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 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 α = 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-
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.

2. Consider the problems of the ACR20 measure, including sensitivity to change and variability in how it was used.

3. Propose new measures that address these problems, such as ACR-N or nACR, and compare them with the ACR20.

4. Evaluate the new measures in clinical trials, paying attention to both efficacy and ease of use.

5. Choose the most effective and practical measure for future use in RA trials.

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>
<thead>
<tr>
<th>Variable name</th>
<th>Variable definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>O’Brien test</td>
<td>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)</td>
</tr>
<tr>
<td>O’Brien.abs</td>
<td>Sum of ranks of absolute changes (ac) in all 7 core measures, where ( ac = final - baseline ) for each measure</td>
</tr>
<tr>
<td>O’Brien%</td>
<td>Sum of ranks of percentage changes (pc) in all 7 core measures, where for each measure ( pc = \frac{final - baseline}{baseline} )</td>
</tr>
<tr>
<td>O’Brien.symm%</td>
<td>Sum of ranks of symmetric percentage changes (spc) in all 7 core measures, where for each measure ( spc = \frac{final - baseline}{final + baseline} )</td>
</tr>
</tbody>
</table>

Continuous definitions of improvement*

Percent change measures (pc)

<table>
<thead>
<tr>
<th>MeanACR</th>
<th>Mean % improvement in the American College of Rheumatology (ACR) 7 core set measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACR-N</td>
<td>Minimum (% tender joint count [TJC], % swollen joint count [SJC], median of other 5)</td>
</tr>
<tr>
<td>Mean3ACR</td>
<td>Mean (% TJC, % SJC, mean of % for other 5)</td>
</tr>
<tr>
<td>MedianACR</td>
<td>Median of % changes in 7 core set measures</td>
</tr>
<tr>
<td>FifthACR</td>
<td>Fifth highest % change</td>
</tr>
<tr>
<td>ThirdACR</td>
<td>Third highest % change</td>
</tr>
<tr>
<td>MeanACR2</td>
<td>Mean of change is all core set measures/mean of all core set measures at baseline</td>
</tr>
<tr>
<td>MeanACR3</td>
<td>Mean of the change in each core set measure/square root of measure score at baseline</td>
</tr>
<tr>
<td>MedianACR2</td>
<td>Median of change is all core set measures/median of all core set measures at baseline</td>
</tr>
</tbody>
</table>

Symmetric percentage change measures

<table>
<thead>
<tr>
<th>Meanspc</th>
<th>Mean symmetric % improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wmeanspc</td>
<td>Mean of 7 with symmetric % TJC and symmetric % SJC double weighted</td>
</tr>
<tr>
<td>Medianspc</td>
<td>Median of 7 symmetric % changes</td>
</tr>
</tbody>
</table>

Disease activity indexes

<table>
<thead>
<tr>
<th>%DAS</th>
<th>Percentage change on DAS (Disease Activity Score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDAS</td>
<td>Absolute change on DAS</td>
</tr>
<tr>
<td>SpecDAS</td>
<td>Symmetric percentage change of DAS</td>
</tr>
<tr>
<td>%SDAI</td>
<td>Percentage change on SDAI (Simplified Disease Activity Index, Smolen)</td>
</tr>
<tr>
<td>DSDAI</td>
<td>Absolute change on SDAI</td>
</tr>
<tr>
<td>ODAI</td>
<td>For each core set measure, create ordinal scale 1,2,3,4 and average them</td>
</tr>
<tr>
<td>%ODAI</td>
<td>Percentage change in ODAI</td>
</tr>
<tr>
<td>DODAI</td>
<td>Absolute change in ODAI</td>
</tr>
<tr>
<td>EULAR</td>
<td>European League Against Rheumatism (EULAR) response criteria, where 0 = no response, 1 = moderate response, 2 = good response</td>
</tr>
</tbody>
</table>

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: \( \frac{mean of core set measures score at end - mean of measures at start}{mean of measures at start} \)
Acmean: \( \frac{mean of core set measures score at end - mean of measures at start}{mean of measures at start + mean of measures at end} \)
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

<table>
<thead>
<tr>
<th>nACR</th>
<th>nACR (number of core set measures improved by ( \geq 20% ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>nACRw</td>
<td>Weighted by TJC and SJC, if TJC and SJC ( \geq 20% ), ( nACRw = nACR ), else ( nACRw = 0 )</td>
</tr>
<tr>
<td>nACRdw</td>
<td>Double weight of TJC and SJC</td>
</tr>
<tr>
<td>nACR50, w50, dw50</td>
<td>Similar as nACR-type measurements, but threshold for improvement increased to 50%</td>
</tr>
<tr>
<td>nACR70, w70, dw70</td>
<td>Similar as above, but threshold for improvement increased to 70%</td>
</tr>
</tbody>
</table>

(continued)
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.

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.

**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.

---

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

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Variable definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>n2ACR</td>
<td>nACR + nACR50</td>
</tr>
<tr>
<td>n2ACRW</td>
<td>nACRW + nACRW50</td>
</tr>
<tr>
<td>n2ACRDw</td>
<td>nACRDw + nACRDw50</td>
</tr>
<tr>
<td>n3ACR, w, dw</td>
<td>Similar as n2ACR-type measurements, but is a sum of nACR + nACR50 + nACR70 (or its variants)</td>
</tr>
</tbody>
</table>

**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 (Continued)
- 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).
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.

#### 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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Spearman’s correlation</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>%SDAI</td>
<td>0.96569</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>%CDAI</td>
<td>0.93778</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>nACR50</td>
<td>0.92975</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>nACR70</td>
<td>0.92975</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>nACR70w</td>
<td>0.89547</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>nACR90</td>
<td>0.87737</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>nACR120</td>
<td>0.86971</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>nACR180</td>
<td>0.86144</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>nACR200</td>
<td>0.85044</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ACR-N</td>
<td>0.83657</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ACRhybrid</td>
<td>0.83150</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>bdACRh3hybrid</td>
<td>0.82231</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>%ODAI</td>
<td>0.80802</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ACRstep</td>
<td>0.79135</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>nACR50</td>
<td>0.78574</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>nACR70</td>
<td>0.77289</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>nACR70w</td>
<td>0.77121</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ACRunw50</td>
<td>0.76574</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>FifthACR50</td>
<td>0.76574</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ACRunw50</td>
<td>0.76574</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>nACR50</td>
<td>0.76574</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>nACR70</td>
<td>0.76574</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>nACR70w</td>
<td>0.76574</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ACR50</td>
<td>0.73140</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ACR70</td>
<td>0.71090</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>LDAS1</td>
<td>0.68817</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>EULAR</td>
<td>0.66271</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>nACR50</td>
<td>0.65902</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>MeanwACR50</td>
<td>0.64574</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ARRunw30</td>
<td>0.62228</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>FifthACR30</td>
<td>0.62228</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ACR50</td>
<td>0.60604</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ThirdACR50</td>
<td>0.58918</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ACRunw20</td>
<td>0.58586</td>
<td>0.0001</td>
</tr>
<tr>
<td>FifthACR20</td>
<td>0.58586</td>
<td>0.0001</td>
</tr>
<tr>
<td>ACR70</td>
<td>0.57165</td>
<td>0.0002</td>
</tr>
<tr>
<td>MeanACR</td>
<td>0.55360</td>
<td>0.0003</td>
</tr>
<tr>
<td>MeanACR_bd</td>
<td>0.55338</td>
<td>0.0003</td>
</tr>
<tr>
<td>MeanACR30</td>
<td>0.50842</td>
<td>0.0011</td>
</tr>
<tr>
<td>LDAS2</td>
<td>0.48516</td>
<td>0.0020</td>
</tr>
<tr>
<td>MeanACR20</td>
<td>0.46735</td>
<td>0.0031</td>
</tr>
<tr>
<td>MeanACR30</td>
<td>0.46149</td>
<td>0.0035</td>
</tr>
<tr>
<td>MeanACR70</td>
<td>0.45426</td>
<td>0.0042</td>
</tr>
<tr>
<td>MedianACR20</td>
<td>0.43411</td>
<td>0.0095</td>
</tr>
<tr>
<td>DASimp</td>
<td>0.35198</td>
<td>0.0302</td>
</tr>
</tbody>
</table>

* For definitions, see Table 1.
Hybrid ACR Response Measure

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-
response 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>
<thead>
<tr>
<th>Candidate definition</th>
<th>Rank of definition among all tested using weighted average of trial data</th>
<th>Lowest (best) rank for candidate among all trials</th>
<th>Highest (worst) rank for candidate among all trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>O’Brien%†</td>
<td>1</td>
<td>1</td>
<td>27</td>
</tr>
<tr>
<td>n3ACR†</td>
<td>3</td>
<td>1</td>
<td>52</td>
</tr>
<tr>
<td>n3ACRdw</td>
<td>4</td>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td>bdACRhybrid†</td>
<td>7</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>MeanACR_bd†</td>
<td>8</td>
<td>1</td>
<td>54</td>
</tr>
<tr>
<td>n2ACR</td>
<td>9</td>
<td>4</td>
<td>61</td>
</tr>
<tr>
<td>n2ACRdw</td>
<td>10</td>
<td>5</td>
<td>58</td>
</tr>
<tr>
<td>MeanACR2</td>
<td>15</td>
<td>6</td>
<td>60</td>
</tr>
<tr>
<td>ACRhybrid†</td>
<td>18</td>
<td>2</td>
<td>49</td>
</tr>
<tr>
<td>%DAS†</td>
<td>20</td>
<td>1.5</td>
<td>89.5</td>
</tr>
<tr>
<td>%SDAH</td>
<td>27</td>
<td>1</td>
<td>122</td>
</tr>
<tr>
<td>MeanACR</td>
<td>28</td>
<td>5</td>
<td>85</td>
</tr>
<tr>
<td>MedianACR</td>
<td>31</td>
<td>5</td>
<td>89</td>
</tr>
<tr>
<td>nACR†</td>
<td>32</td>
<td>3</td>
<td>80</td>
</tr>
<tr>
<td>nACRdw</td>
<td>33</td>
<td>1</td>
<td>78</td>
</tr>
<tr>
<td>nACR50</td>
<td>34</td>
<td>4</td>
<td>66</td>
</tr>
<tr>
<td>nACRdw50</td>
<td>35</td>
<td>6</td>
<td>73</td>
</tr>
<tr>
<td>%CDAI</td>
<td>45</td>
<td>21</td>
<td>86</td>
</tr>
<tr>
<td>%ODAI</td>
<td>46</td>
<td>20</td>
<td>70</td>
</tr>
<tr>
<td>MedianACR2</td>
<td>49</td>
<td>20</td>
<td>79</td>
</tr>
<tr>
<td>n3ACRw</td>
<td>50</td>
<td>8</td>
<td>105</td>
</tr>
<tr>
<td>ACR-N†</td>
<td>53</td>
<td>14</td>
<td>93</td>
</tr>
<tr>
<td>EULAR†</td>
<td>54</td>
<td>7</td>
<td>111</td>
</tr>
<tr>
<td>n2ACRw</td>
<td>55</td>
<td>12</td>
<td>102</td>
</tr>
<tr>
<td>nACRw</td>
<td>62</td>
<td>7</td>
<td>98</td>
</tr>
<tr>
<td>ACRstep</td>
<td>64</td>
<td>12</td>
<td>97</td>
</tr>
<tr>
<td>(ACR20,50,70)</td>
<td>69</td>
<td>1</td>
<td>131</td>
</tr>
<tr>
<td>ThirdACR30</td>
<td>70</td>
<td>17</td>
<td>130</td>
</tr>
<tr>
<td>MedianACR30</td>
<td>72</td>
<td>51</td>
<td>103</td>
</tr>
<tr>
<td>ACRstep3</td>
<td>74</td>
<td>5</td>
<td>126</td>
</tr>
<tr>
<td>DASimp†</td>
<td>77</td>
<td>36</td>
<td>113</td>
</tr>
<tr>
<td>ThirdACR50</td>
<td>78</td>
<td>13.5</td>
<td>111.5</td>
</tr>
<tr>
<td>ACRuw30</td>
<td>79</td>
<td>13.5</td>
<td>111.5</td>
</tr>
<tr>
<td>FifthACR30</td>
<td>83</td>
<td>5</td>
<td>116</td>
</tr>
<tr>
<td>MeanwACR30</td>
<td>87</td>
<td>7</td>
<td>131</td>
</tr>
<tr>
<td>ACRuw20</td>
<td>88</td>
<td>7</td>
<td>131</td>
</tr>
<tr>
<td>FifthACR20</td>
<td>90</td>
<td>22</td>
<td>122</td>
</tr>
<tr>
<td>MeanACR20</td>
<td>91</td>
<td>11</td>
<td>129</td>
</tr>
<tr>
<td>MeanACR30</td>
<td>92</td>
<td>39.5</td>
<td>125.5</td>
</tr>
<tr>
<td>MedianACR20</td>
<td>96</td>
<td>31</td>
<td>129</td>
</tr>
<tr>
<td>nACRw50</td>
<td>97</td>
<td>41</td>
<td>129</td>
</tr>
<tr>
<td>MeanwACR50</td>
<td>101</td>
<td>40</td>
<td>122</td>
</tr>
<tr>
<td>ACR20†</td>
<td>107</td>
<td>18</td>
<td>131</td>
</tr>
<tr>
<td>LDAS1</td>
<td>110</td>
<td>76</td>
<td>118</td>
</tr>
<tr>
<td>ACR50</td>
<td>113</td>
<td>11</td>
<td>130</td>
</tr>
<tr>
<td>ThirdACR20</td>
<td>116</td>
<td>84</td>
<td>128</td>
</tr>
<tr>
<td>ACR40</td>
<td>118</td>
<td>61</td>
<td>118</td>
</tr>
<tr>
<td>ACRuw50</td>
<td>119</td>
<td>61</td>
<td>118</td>
</tr>
<tr>
<td>FifthACR50</td>
<td>123</td>
<td>68</td>
<td>133</td>
</tr>
<tr>
<td>ACR50†</td>
<td>130</td>
<td>18</td>
<td>133</td>
</tr>
<tr>
<td>ACR70†</td>
<td>135</td>
<td>82</td>
<td>133</td>
</tr>
<tr>
<td>LDAS2</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
</tbody>
</table>

* Most widely clinically used measures and other representative results are presented. For definitions, see Table 1.
† Candidates of special interest.
shall propose an approach to timing of the hybrid ACR based on timing of measurement. In another report, we ways in which the hybrid ACR measure could be used which the ACR20 has been defined. There are a variety of we provide additional tables in which we compute the sure score (see Table 4). To make it clear how this works, old score. Combined with the mean percent improvement in swollen and tender joint counts to reach a certain thresh- response measure, which requires improvement in the problem. We have proposed one solution, the hybrid ACR compromise the sensitivity to change of outcome measure- jointly and especially the swollen joint count can produce 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, 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 remission. 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.

The table below shows the scoring method for the Hybrid American College of Rheumatology (ACR) response measure. We propose a scoring system where the ACR20, ACR50, and ACR70 are combined to create a new hybrid outcome measure. This method involves calculating the mean percent improvement in core set measures, and then determining whether the patient has achieved ACR20, ACR50, or ACR70.

<table>
<thead>
<tr>
<th>ACR status</th>
<th>Mean % change in core set measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not ACR20</td>
<td>Mean % change 19.99</td>
</tr>
<tr>
<td>ACR20 but not ACR50</td>
<td>20 Mean % change 49.99</td>
</tr>
<tr>
<td>ACR50 but not ACR70</td>
<td>50 Mean % change 69.99</td>
</tr>
<tr>
<td>ACR70</td>
<td>70 Mean % change 100%</td>
</tr>
</tbody>
</table>

* 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 more than 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 percent improvement in core set items; the Hybrid ACR score is where they intersect in the table.

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.


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

Background indices. Pincus.

References


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.

<table>
<thead>
<tr>
<th>Core set item</th>
<th>Range</th>
<th>Pretreatment</th>
<th>Final value</th>
<th>% change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tender joint count</td>
<td>0–68</td>
<td>43</td>
<td>14</td>
<td>67</td>
</tr>
<tr>
<td>Swollen joint count</td>
<td>0–66</td>
<td>38</td>
<td>44</td>
<td>−16*</td>
</tr>
<tr>
<td>Pain</td>
<td>0–10</td>
<td>4.6</td>
<td>2.0</td>
<td>60</td>
</tr>
<tr>
<td>Patient global</td>
<td>0–10</td>
<td>8.5</td>
<td>3.1</td>
<td>63</td>
</tr>
<tr>
<td>Physician global</td>
<td>0–10</td>
<td>7.5</td>
<td>2.7</td>
<td>64</td>
</tr>
<tr>
<td>Health Assessment Questionnaire</td>
<td>0–3</td>
<td>2.8</td>
<td>2.0</td>
<td>28</td>
</tr>
<tr>
<td>C-reactive protein level (mg/dl)</td>
<td>&gt;0</td>
<td>11.6</td>
<td>7.2</td>
<td>38</td>
</tr>
</tbody>
</table>

* 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.

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

**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.

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