Editorial

Whither the ACR20?

Wisdom lies neither in fixity nor in change,
but in the dialectic between the two. — Octavio Paz, 1989

Clinical trials in rheumatoid arthritis (RA) provide the main evidence for the efficacy of treatments. Through the early 1990s, trials generally included at least 10 outcome measures and often many more; as well, there was disconcerting heterogeneity in the measures assessed. Also, outcomes that were insensitive to change were widely used, biasing against the detection of a treatment's efficacy. The reliance on multiple outcome measures, the use of different sets of measures in different trials without a common standard, and the inclusion of outcomes insensitive to change in some, but not all, trials made it impossible to accurately assess or compare the efficacy of therapies.

In the early 1990s, a committee of the American College of Rheumatology (ACR) subjected outcome measures in RA trials to a critical evidence-based evaluation and selected a "core set" of outcome measures that met basic criteria of validity and non-redundancy[1]. Measures were chosen that were sensitive to change (a measure with high sensitivity to change or discriminant validity detects differences in a trial between an effective treatment and placebo with high levels of statistical significance, whereas a measure with poor sensitivity to change might not show the same treatment as significantly more efficacious than placebo) and sampled broadly from different domains of disease activity including pain, acute phase reaction, and number of tender or swollen joints.

Using this core set, the committee proceeded to define a single measure of response, the ACR preliminary definition of improvement[2], later called the ACR20 (Table 1). The ACR20 met 2 important criteria: first, it was at least as sensitive to change as other candidate measures of response that were tested; and second, it corresponded with rheumatologists' characterizations of patients as having improved (clinical face validity). The ACR20 has been widely adopted as the primary outcome measure in RA clinical trials, has been recommended by the US Food and Drug Administration in evaluating drugs for regulatory approval, and has been used as a standard for comparison of therapies. ACR20 has made it possible to recognize the potency of new tumor necrosis factor-α inhibitors compared to conventional therapies, and has successfully created a standard on which different treatments can be compared (even though comparison of ACR20 rates across trials is methodologically problematic). The data-driven consensus method exemplified by the development of the ACR20 has been replicated in proposed definitions of response for ankylosing spondylitis[3], juvenile rheumatoid arthritis[4], low back pain, and other rheumatic and musculoskeletal conditions.

Table 1. American College of Rheumatology preliminary definition of improvement (aka ACR20).

Required for a patient to improve
≥ 20% improvement in tender joint count, AND
≥ 20% improvement in swollen joint count, AND
≥ 20% improvement in 3/5 other core set items
Core set items
MD global assessment
Patient global assessment
Patient pain
Disability (self-reported using validated instrument)
Erythrocyte sedimentation rate/C-reactive protein

Given the dramatic success of the ACR20 in altering the approach to outcome evaluation in RA trials, why should there be any consideration of changing it? First, there has been "threshold creep," with the emergence of the ACR50 and ACR70, in which patients must improve at least 50% or 70%,

See Criteria for improvement in rheumatoid arthritis: alternatives to the ACR 20, *page 856*

respectively, in individual core set measures to be characterized as improved. This threshold creep has arisen because an ACR20 response in a patient was felt not to reflect a major clinical response to treatment. Further, the increase in threshold reflects the availability of more efficacious therapies than when ACR20 was promulgated. While ACR20 has been widely adopted, its use differs from trial to trial in terms of the timing of response, and, as a consequence, response rates to the same drug vary across trials[5].

Lastly, other ways of defining response using core set measures have been proposed. These include: the number of ACR core set measures improved by at least 20% (nACR) and an average of 3 variables, the percentage improvement in tender joint count, the percentage improvement in swollen joint count, and the median percentage improvement in the other 5 core set measures (ACRn). ACRn and nACR have been tested in individual trials and have been shown to be more sensitive to change than the ACR20. Anderson, *et al*[6] recently performed a simulation study, i.e., a type of study with broader generalizability to RA trials than analyses of single trials; they found the same results: the ACR20 was less sensitive to change than other response measures using the core set.

The ACR20 defines response dichotomously — when treated, an individual patient is characterized as either having or not having responded. Response is defined in advance and can be easily interpreted by clinicians and compared across trials. However, transforming a continuous measure of response into a dichotomous one dictates a loss of information. Patients who fail to reach an ACR20 response, for example, have been shown to benefit radiographically from treatment[7], suggesting that absence of an ACR20 response leaves out valuable response information for a patient. A continuous or ordinal approach when measuring range of response would be more informative and provide more statistical power to distinguish between treatments than a dichotomous approach. In addition, averaging response over time eliminates day-to-day variation in a patient's RA activity and minimizes measurement variability, resulting in a more reproducible estimate of response. This improves statistical power.

It is in the context of these questions that the valuable article by Albert and colleagues in this issue of *The Journal*[8] can best be understood. Albert and colleagues returned to the Cooperating Clinics (CSSRD) trial database, a database of large US National Institutes of Health-funded multicenter RA trials used to develop the ACR20, to evaluate whether continuous or ordinally-measured response to treatment (they focus on the number of core set elements that improve by 20%, 50%, or 70% at different time points) would have better discriminant validity than ACR20, ACR50, or ACR70 taken by themselves. Not surprisingly, their results suggest that the ACR20 or other dichotomous measures of response are not as sensitive to change as an ordinal one. They also report that an area under the curve approach, which averages response over time, performs even better.

While the findings of Albert, *et al* corroborate those of others and begin to point out alternatives to the ACR20 that might work well, there are flaws in the work that may detract from the validity of their findings. First, while the CSSRD trials were used to develop ACR20 at a time when newer therapies were not yet available, these trials do not include modern methods of core set measurement: validated self-reported measures of physical function were not part of these trials, and the global assessments used broad interval scales, and finer intervals are more sensitive to change[9]. Current measurement instruments, as recommended in the ACR core set, might have yielded different results.

The approach suggested by Albert and colleagues counts the number of core set measures improved by 20% (or 50% or 70%). In doing so, it treats tender and swollen joint counts equally to other core set measures, without requiring that they must improve for a patient to be characterized as improved. Such an approach generates a definition of response that is more sensitive to change than measures of response that require joint count improvement[10]. In fact, of the core set measures, patient-measured responses are, in general, the most sensitive to change and swollen joint counts are among the least sensitive. Nonetheless, in surveys done during development of the ACR20, rheumatologists consistently characterized swollen joint count as the most important measure of disease activity and relied heavily on it to decide whether a patient was improved. So, if the approach recommended by Albert, *et al* were to be used, a therapy could be characterized as efficacious if it produced no improvement in swollen and tender joint count, a scenario that most rheumatologists would reject as not having clinical validity.

Lastly, the report by Albert and colleagues suggests that placebo-controlled trials may require different outcome measures than trials comparing 2 active agents. We find no conceptual or empirical evidence to support this contention. The ACR20 was developed in a placebo trial data set and validated in a comparative trial data set[2], and it had high discriminant validity in both. Subsequent validation studies of the ACR20, which have confirmed its sensitivity to change, have been carried out in placebo and comparative trials[11]. If there is one set of response measures for comparative trials and another for placebo-controlled trials, that will create different sets of standards that could make it hard to compare treatments across trials.

This raises one other limitation to the Albert, *et al* study and a caution to ACR20 reevaluation in general, the problem of multiplicity[5]. The ACR20 constituted a major advance in outcome measurements, in large part because it put forward one single outcome measure at a time when 10 to 15 outcome measures were being used in trials. With the wide use of ACR50 and 70, this single outcome measure is now often supplanted by 3 or even more primary outcome measures[5]. If the recommendation of Albert and colleagues were to take hold, then at least 4 outcome measures (and perhaps more)

would be reported in each trial, each of which would be tested for statistical significance. This multiplicity of outcome measures would permit authors to report positive results when any one measure reached significance. This harkens back to the pre-ACR20 days when interpretation of a trial's result could differ depending on how many and which outcome measures showed statistically significant efficacy.

The ACR has constituted a committee to reevaluate the ACR20 and outcome measurement in RA trials. As highlighted by the study by Albert and colleagues and the comments above, outcome measurement in RA trials currently poses a number of extremely challenging methodologic conundrums. First, how do we create a measure more responsive then ACR20, 50, or 70, yet preserve the single outcome measure standardization that has been so clinically valuable? Second, how do we reevaluate the definition of improvement and create a measure even more sensitive to change, yet still place emphasis on joint count improvement when, if joint count improvement is required, the measure's sensitivity to change falls? Third, how can we standardize an approach that defines response by averaging measures over time, when different treatments have different onsets and durations of action? Fourth, should we continue to define response by looking at percentage of improvement in core set measures, or should other fundamentally different approaches be adopted?

Because the ACR20 is popular, yet does not perform optimally, any solution to these problems will not likely be met with immediate and widespread acceptance. Thus, the ideal of a measure that is optimally sensitive to change, clinically valid and understandable, and serves as the single outcome measure for all RA trials, permitting cross-treatment comparisons, will continue to be elusive.

DAVID T. FELSON, MD, MPH,
Clinical Epidemiology Research and Training Unit,
Boston University School of Medicine,
715 Albany Street, A207,
Boston, Massachusetts 02118, USA

Supported by NIH AR47785. Address reprint requests to Dr. Felson.

REFERENCES

1. Felson DT. Choosing a core set of disease activity measures for rheumatoid arthritis clinical trials. J Rheumatol 1993;20:531-4.
2. Felson DT, Anderson JJ, Boers M, et al. American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum 1995;38:727-35.
3. Anderson JJ, Baron G, van der Heijde D, Felson DT, Dougados M. Ankylosing spondylitis assessment group preliminary definition of short-term improvement in ankylosing spondylitis. Arthritis Rheum 2001;44:1876-86.
4. Giannini EH, Ruperto N, Ravelli A, Lovell DJ, Felson DT, Martini A. Preliminary definition of improvement in juvenile arthritis. Arthritis Rheum 1997;40:1202-9.
5. Felson DT. Assessing the efficacy and safety of rheumatic disease treatments: obstacles and proposed solutions. Arthritis Rheum 2003;48:1781-7.
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. Boers M, COBRA Study Group. Combinatietherapie Bij Reumatoide Artritis. Demonstration of response in rheumatoid arthritis patients who are nonresponders according to the American College of Rheumatology 20% criteria: the paradox of beneficial treatment effects in nonresponders in the ATTRACT trial. Anti-Tumor Necrosis Factor Trial in Rheumatoid Arthritis with Concomitant Therapy. Arthritis Rheum 2001;44:2703-4.
8. Albert DA, Huang G, Dubrow G, Brensinger C, Berlin J, Williams HJ. Criteria for improvement in rheumatoid arthritis: alternatives to the ACR 20. J Rheumatol 2004;31:856-66.
9. Anderson JJ, Felson DT, Meenan RF, Williams HJ. Which traditional measures should be used in rheumatoid arthritis clinical trials? Arthritis Rheum 1989;32:1093-9.
10. Pincus T, Strand V, Koch G, et al. 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.
11. Felson DT, Anderson JJ, Lange ML, Wells G, LaValley MP. Should improvement in rheumatoid arthritis clinical trials be defined as fifty percent or seventy percent improvement in core set measures, rather than twenty percent? Arthritis Rheum 1998;41:1564-70.