Rationale and Strategies for Reevaluating the ACR20

DAVID T. FELSON, DANIEL E. FURST, and MAARTEN BOERS

ABSTRACT. ***Objective.*** To assess whether the American College of Rheumatology response criteria ACR20 should be replaced by another definition of response with enhanced discriminant validity.

Methods. We worked with statisticians to define over 100 different ways of defining response, including dichotomous definitions (e.g., ACR20; ACR50; ACR70; low disease activity), ordinal definitions (EULAR response; ACR20, ACR50, ACR70), disease activity indexes [Disease Activity Score (DAS); Disease Activity Index, SDAI], continuous definitions (mean percentage improvement in all core set measures; nACR, ACRn), and hybrid definitions (ACR20, ACR50, ACR70 defined for a patient as 0, 1, 2, 3 scale with continuous measures between intervals) along with variations on each of these approaches (e.g., percentage vs absolute change in DAS; e.g., measures requiring vs not requiring joint count improvement). To test clinical validity, we administered a survey using patients from a trial who had various levels of improvement and asked rheumatologists whether and by how much these patients improved. For Sn-to-Chge, we are collecting data from large disease modifying antirheumatic drug multicenter trials in rheumatoid arthritis and ranking candidate definitions of response on their average p values in distinguishing active treatment from placebo or combination compared to single comparator.

Results. We surveyed 52 rheumatologists about which trial patients had improved and by how much. Trial data were obtained and tested for sensitivity to change.

Conclusion. A rigorous data-driven consensus process was used to reassess the ACR20. (J Rheumatol 2007;34:1184–7)

Key Indexing Terms:
CLINICAL TRIALS
OUTCOME ASSESSMENT
RHEUMATOID ARTHRITIS
TREATMENT OUTCOME

Prior to the development of the American College of Rheumatology (ACR) preliminary definition of improvement (called the ACR20) and similar efforts in Europe, a multitude of outcomes was used to evaluate rheumatoid arthritis (RA) treatments. Trial reports often contained 10 or 15 primary outcome measures, and each trial evaluated different ones, so that there was little standardization across trials. This led to chance significant findings and great difficulty in comparing different treatments using the same metrics. Also, multiple interpretations were possible if, for example, there were only one or 2 significant results out of 10 or 15 outcomes tested. As well, authors could actually test even more primary outcomes and report only those that were positive, creating positive reporting bias.

In the early 1990s, using a data-driven consensus approach, representatives from the ACR, EULAR, and the WHO developed a core set of outcome measures for RA trials. Measures were selected based on their sensitivity to change, their lack of redundancy, their content validity (whether they sampled from multiple domains of RA activity), and whether they predicted important outcomes in RA, including disability, radiographic damage and death. The core set measure ratified internationally at OMERACT 1 included 7 measures: tender joint count, swollen joint count, physician global assessment, patient global assessment, patient pain assessment, patient self-reported disability, and an acute-phase reactant (either erythrocyte sedimentation rate or C-reactive protein). For studies lasting a year or longer, the group also recommended radiographs[1,2].

With the core set measures, there was a uniformity of the measures that were to be used in RA trials, but there remained 7 measures, and a single measure is preferred. Further, the core set measures, like previous outcome assessments in RA trials, focused on comparing the mean improvement of treated groups in a trial, rather than on individual patients with RA. Evaluating how many patients improved would be a clinically more relevant measure than mean improvement of treated groups.

With those 2 limitations in mind, and following a process initiated at the first Outcomes In Rheumatology (OMERACT) conference[3], an ACR committee joined by international representatives proceeded to develop a definition of improvement in RA, using 2 parallel strategies: a survey of paper patients drawn from randomized trials to determine which patients had at least minimal clinical improvement according to rheuma-

From the Boston University School of Medicine, Boston, Massachusetts, USA; Department of Clinical Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands; and University of California, Los Angeles, Los Angeles, California, USA.

Supported by NIH AR47785 and by a grant from the American College of Rheumatology.

D.T. Felson, MD, MPH, Professor of Medicine and Epidemiology, Boston University School of Medicine; M. Boers, MSc, MD, PhD, Professor of Clinical Epidemiology, Department of Clinical Epidemiology and Biostatistics, VU University Medical Center; D.E. Furst, MD, Professor of Medicine, UCLA, Los Angeles, CA, USA.

Address reprint requests to Dr. D.T. Felson, A203, Boston University School of Medicine, 80 East Concord Street, Boston, MA 02118. E-mail: dfelson@bu.edu

tologists, and an analysis of trial data in RA to determine which potential definition of improvement would best separate active treatment from placebo and strong treatment from weaker treatment.

Using definitions of response that had been proposed in the literature and also developing variations on those measures and coming up with our own measures, the committee ultimately tested 40 candidate measures of improvement, evaluating their clinical validity (their agreement with rheumatologists' impression of improvement) and their discriminant validity (how well they distinguished active treatment from placebo and stronger from weaker treatments). The candidate measure that met both needs, in that it corresponded well to clinicians' impressions of patient improvement and also discriminated well those on active treatment from those on placebo (or stronger vs weaker treatment) was selected as the ACR preliminary definition of improvement (later called the ACR20)[4].

Promulgation of the ACR20 yielded a number of salutary effects in clinical trial measurement and reporting and in terms of standardization. It was immediately adopted by most ongoing RA trials and almost immediately appeared as an outcome measure in published trial reports. Although it was not the primary intention when developed, the ACR20 was, by virtue of being an index (i.e., a combination of different measures), more powerful than any of the single core set measures in discriminating between drugs that differed in efficacy. About two-thirds of RA trial reports currently contain the ACR20, either as a primary or as a secondary measure of efficacy. Additionally, the multiple outcome measures that previously filled "Results" sections of RA trials vanished, with the focus instead on the ACR20 and/or variations of the ACR20 (see below). Non-core set measure outcomes also disappeared, except in trials targeted toward preventing structural damage. Trial standardization was greatly advanced, and a focus of trials became the response to treatment of individual patients, rather than the mean response of a group of treated patients. These overall effects made it easier to discern the relative efficacy of treatments in RA. For example, ACR20 rates were high for a new class of treatment, the tumor necrosis factor inhibitors, and for novel combinations of traditional disease modifying antirheumatic drugs, allowing the rheumatology community to discern their potent efficacy.

In addition, the need to define response of individual patients in a standardized way across trials became apparent to those studying other rheumatic diseases, and that, in turn, spurred efforts by experts to develop response criteria in juvenile RA[5], osteoarthritis[6], ankylosing spondylitis[7], and even in low back pain[8].

While there were clearly great benefits of the introduction of the ACR20, and of the core set measures in general, problems arose almost immediately after the promulgation and wide acceptance of the ACR20. First, with the introduction of new therapies that were perhaps more efficacious than the ones previously available, there was threshold creep. It seemed that requiring only 20% of improvement in core set measures was not enough, and that requiring a higher percentage of improvement might be more meaningful, especially for new possibly more potent therapies. Appearing in trial reports were higher thresholds for the ACR20, including ACR50 and ACR70. Each of these was defined the same way as ACR20, but with a 50% or 70% threshold, respectively. Also, ACR20 was not used consistently across trials, with some trial analyses focusing on ACR20 response rates during the trial and others at the end of the trial. This created discrepancies between response rates of the same drug in different trials[9].

In addition, new definitions of response began to appear and were touted as showing more sensitivity to change than the ACR20. Examples included the ACRn[10] and even patient-specific measures that were evaluated as continuous rather than dichotomous measures[11]. Indeed, Anderson and colleagues[12], using simulation studies of RA trial data, showed that the ACR20 had far less power than an optimal statistic such as the O'Brien test, a test that defined response on a continuous basis and that searched through response likelihood in 2 treatment groups in a trial to find the point of maximal discrimination between an active treatment and control. Continuous or ordinal definitions of response appeared to have better sensitivity to change than the ACR20, and *a priori* definitions of improvement like the ACR20 often did not perform as well as ways of evaluating response to therapy that were data-driven, varying from trial to trial.

Even so, ACR20 became the standard vocabulary for describing treatment response. In review talks on RA treatments, including those at the ACR meeting, most contained comparisons the ACR20 response rates of different treatments. (This was true even though these comparisons often were not valid because of other differences between trials.) ACR20 rates reported in trials and presented by reviewers in educational talks provided a metric that was easy to grasp and that allowed rheumatologists and others to gauge the comparative efficacy of treatments and the likely response of patients to treatment. There are a number of other strengths to having a standard measure including the ability to do metaanalyses of trials using the same outcome measure and the continued ability to make comparisons, albeit imperfect ones, between treatments. Further, the wide acceptance of a single measure of response has discouraged the practice of reporting and testing multiple different primary outcomes, a significant problem before the promulgation of the ACR20.

With the recognition that ACR20 did not have as much discriminant validity as other potential definitions of improvement, an ACR committee was formed in 2003 to reevaluate improvement criteria in RA. The goals of this committee were to define response so that the discriminant validity of response definition could be maximized and trials in RA could be carried out with fewer patients. The second goal of the commit-

tee was to define response in a way that preserved clinical validity and was readily understandable to rheumatologists, to their students, and even to patients. The third committee goal was to define response so that its evaluation could be standardized, optimizing communication, minimizing a multiplicity of outcomes, and optimizing trial comparisons.

Agenda for the Committee to Reevaluate Improvement Criteria in RA

The committee started its work with one important assumption, that the basic elements of the core set are sound. Different analysts had evaluated different ways of measuring core set elements (e.g., restricted joint counts vs extended joint counts; visual analog vs Likert pain or global scales; C-reactive protein vs erythrocyte sedimentation rate) and had not found that variations of measurement had important effects on the overall discriminant validity of the core set or of measures derived from the core set that defined response.

The general approach to redefining improvement criteria was the same 2-track parallel approach that was used to define the original preliminary definition of improvement, the ACR20.

Truth

One part of the approach was to survey rheumatologists using paper patients from real trials and ask them to evaluate not only whether the depicted patients had improved during the trial, but how much they had improved. The goals of this survey were to evaluate how well candidate measures of improvement corresponded to clinician's impressions of the degree of improvement. Given analyses that had already shown that ordinal and continuous measures of response had better discriminant validity than the dichotomous ACR20, there was a high likelihood that a new response definition would define response on a continuum, not just dichotomously as improved/not improved. Thus, the survey asked clinicians to grade the amount of response being experienced by patients so that survey analysis could test the correlation of this improvement with the amount of improvement defined by the candidate measures.

Discrimination

The second track pursued to redefine response was to analyze trial data and to test candidate measures to see which measures had the best discriminant validity.

Prior to developing the survey or analyzing trial data, the committee identified a comprehensive list of potential candidate measures of response. Over 100 potential measures of response were examined and tested both in the survey and in the analyses of trial data. These included widely used dichotomous measures of improvement, including the ACR20, ACR50, and ACR70, the defined measure of low disease activity[13], indices such as the DAS and the SDAI[14], ordinal outcomes in which individual patients could be characterized based on the amount of improvement they experienced (e.g., EULAR definition of response)[13]. One example of a new ordinal approach was to create an ordinal measure out of the ACR20, ACR50, and ACR70 so that the patient could have 4 levels of improvement: < ACR20, ACR20 but not 50, ACR50 but not 70, and ≥ ACR70. This creates a 0, 1, 2, 3 variable for each patient. EULAR also had developed an ordinal measure that was trichotomous. Lastly, we tested a variety of continuous measures of response, such as the ACRn[7], the nACR — a count of the number of core set measures improved by at least 20% in a patient, the mean percentage improvement in core set items, and the median percentage improvement in core set items.

These candidate definitions dealt with tender and swollen joint counts differently. For example, the nACR measures do not treat joint counts differently than any other core set measures. However, ACR20 requires at least 20% improvement in tender and swollen joint count for a patient to be labeled as improved. In addition to evaluating the performance of response measures, the committee evaluated how important it was for joint count improvement to be part of a patient's improvement picture. The survey was designed to include examples of patients from the RA trials who had improvements in many variables, but not in both of their joint counts. This allowed for an examination of whether rheumatologists felt these patients were improved.

To evaluate the discriminant validity of these candidate measures, the committee, with the help of industry sponsors, assembled a list of 11 large randomized trials published since the dissemination of the core set and the ACR20. Trial data were provided and core set measures were available in all of the trials, so that it was possible to test the discriminant validity of these candidate measures in over 3600 patients who had participated in placebo control or comparative randomized trials in RA.

Ultimately, the goal, like the process that generated the ACR20, was to reevaluate the ACR20, determine if it should be retained, and if not, select a new definition of response that contained many of the positive qualities of the ACR20. These included its understandability and its use to standardize trials. On the other hand, a goal was to improve on discriminant validity of the response definition so that the efficacy of treatments could be detected with fewer subjects.

Feasibility

The feasibility of final measure would be determined based on testing in trials and response of trialists to its introduction.

Research Agenda

At the special interest group, we reviewed data presented to the ACR committee including comparisons of the sensitivity to change of currently used measures of outcome versus new candidate measures. Respondents were asked to express opinions about unresolved concerns in defining response and were

solicited for suggestions about additional issues to be addressed before finalizing the ACR choices, as follows.

1. Define whether swollen joint count had to be included in the revised response criteria. There was considerable discussion of defining response using candidate measures that treated all core set measures equally without requiring improvement in joint counts. Two extreme points of view were presented. On one hand, some members of the group felt that if swollen joint count improvement was not required, analgesic therapies might achieve claims of effecting RA improvement. If pain related measures improved with analgesia, there would be no necessity for therapies to lead to improvement in biological measure improvement in order to gain credibility as an RA treatment. According to these members, this violated the construct validity of improvement in RA, which should require measureable improvement in biological parameters.

On the other hand, others suggested strongly that measures of response that performed optimally ought to be selected whether they required joint count improvement or not. It was noted that swollen joint count improvement was often an insensitive measure to change and compromised the discriminant validity of candidate measures. Between these 2 viewpoints, there was strong disagreement whether swollen joint count ought to be required (preserving clinical construct validity) or should not be required (enhancing sensitivity to change).

2. Examine the usefulness and need for the inclusion of worsening in the definition of response. Other questions were posed during the discussions that bear importantly on the selection and validation of any new measure of response. These included whether a measure should incorporate worsening of patients into the definition of response (most respondents felt that it should).

3. Ascertain the usefulness and need for specific time to response as part of the revised response definition.

4. Ascertain whether and to what extent a new measure needs to be intuitively and easily understandable. Some participants noted that neither the ACR20 nor the DAS was especially easy to understand. Even so, newly proposed measures, especially those that are continuously defined, may not be readily understandable, and there was felt to be a tradeoff between improving sensitivity to change and preserving understandability. This tradeoff was likely to be specific for each candidate measure tested and needs to be such that the new measure remains both understandable and sensitive to change. Many felt that understandability needed to be maintained at the potential cost of some decrease in discrimination. A related issue is that the new measure needed to be standardizable so that response rates could be reported and compared across trials, as has been done with the ACR20. A new measure building on ACR20 so that ACR20 could be extracted from the new measure would be preferable.

Conclusion

The special interest group discussion generated many questions that participants felt were unresolved in the committee's approach to developing new response criteria for rheumatoid arthritis, and additional work was thought to be needed prior to the promulgation of new response criteria.

REFERENCES

1. Felson DT, Anderson JJ, Boers M, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum 1993;36:729-40.
2. Boers M, Tugwell P, Felson DT, 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 1994;22 Suppl 41:86-9.
3. Goldsmith CH, Boers M, Bombardier C, Tugwell P. Criteria for clinically important changes in outcomes: Development, scoring and evaluation of rheumatoid arthritis patient and trial profiles. J Rheumatol 1993;20:561-5.
4. 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.
5. 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.
6. Pham T, van der Heijde D, Altman RD, et al. OMERACT-OARSI initiative: Osteoarthritis Research Society International set of responder criteria for osteoarthritis clinical trials revisited. Osteoarthritis Cartilage 2004;12:389-99.
7. van der Heijde D, Dougados M, Davis J, et al. ASsessment in Ankylosing Spondylitis International Working Group/Spondylitis Association of America recommendations for conducting clinical trials in ankylosing spondylitis. Arthritis Rheum 2005;52:386-94.
8. Deyo RA, Battie M, Beurskens AJ, et al. Outcome measures for low back pain research. A proposal for standardized use. Spine 1998;23:2003-13.
9. Felson DT. Whither the ACR20? J Rheumatol 2004;31:835-7.
10. Bathon JM, Martin RW, Fleischmann RM, et al. A comparison of etanercept and methotrexate in patients with early rheumatoid arthritis. New Engl J Med 2000;343:1586-93.
11. 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.
12. Anderson JJ, Bolognese JA, Felson DT. Comparison of rheumatoid arthritis clinical trial outcome measures: a simulation study. Arthritis Rheum 2003;48:3031-8.
13. 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.
14. Smolen JS, Breedveld FC, Schiff MH, et al. A simplified disease activity index for rheumatoid arthritis for use in clinical practice. Rheumatology Oxford 2003;42:244-57.