Editorial

Clinical Trials, Outcome Measures, and Response Criteria

After a patient has completed a clinical trial, the physician is anxious to know: "Did my patient get better?" Usually, the next question is: "Is the new treatment any good?" The emphasis of the first question tends to be whether the patient showed meaningful clinical improvement by the end of the study. For the second question, the emphasis is more toward whether the response in the group receiving the treatment of interest was superior to that in the control or placebo group. A positive answer for one question does not necessarily imply a "yes" answer for the other. For example, the patient's physician may think: "I regret that the trial ended up with a negative result, but my patient got better." The clinical trialist may think: "I regret that your patient didn't get better, but the treatment that we tested was clearly superior to the placebo." Here, we will look into how these questions have been approached, with an emphasis on clinical trials for rheumatoid arthritis (RA). We will also address how an alternative analytic approach may be used in RA as well as in conditions for which, in contrast to RA, widely agreed-upon response criteria may not yet be available.

Over an extended period of time, investigators in many countries have sought the optimal outcome measures for clinical trials in various rheumatic diseases. For RA, the efforts included the examination of a number of individual measures to find the best one[1]. These efforts have included exercises on how to count and score joints, and which joints to evaluate[2,3]. When considerable data became available after decades of work, an international effort was launched to determine which measures should be included in a core set to best describe the disease activity. After the Outcome Measures in Rheumatoid Arthritis Clinical Trials (OMERACT) meeting, the American College of Rheumatology (ACR) published a core set of criteria that were recommended for inclusion in clinical trials of RA[4-6]. This core set was subsequently ratified by the World Health Organization (WHO) and the International League of Associations for Rheumatology (ILAR) as the WHO/ILAR core set[7]. However, the question of how to optimally combine these outcome measures remained[8,9]. Paulus had developed rules-based criteria for response in RA trials based on data from the Cooperative Systemic Studies of Rheumatic Diseases (CSSRD) program (Table 1)[10]. Further activities by the ACR followed a heuristic or rules-based approach. The ACR analyzed data from CSSRD trials and applied rules that were essentially different ways of weighting and summing up the data across the core set of outcome measures[11]. An optimal rule would give the greatest difference between the active treatment and the placebo group, while giving a low placebo response rate. The goal was to achieve the greatest power to detect a treatment with the smallest number of patients. Different response criteria proposed by the European League Against Rheumatism (EULAR) and others were tested in analyses of the same CSSRD trials[11]. These analyses resulted in the selection of a rule that could be applied by clinicians to determine whether a patient was or was not a responder for a given therapy. The EULAR criteria use a more complex approach than the ACR, where each outcome measure is divided into categories based on percentage change from baseline[12,13]. As shown in Table 1, the EULAR criteria are based on both an improvement and the achievement of a low disease activity state, as measured by the Disease Activity Score (DAS). In the DAS, the weights for summing across different measures have been determined using a multivariate approach. This results in a more complicated equation that may have more precision for the set of clinical trial data on which it was originally based than, say, the ACR response criteria.

There are problems with the heuristic approach. The first of these problems is inherent in extrapolating data from a single set of clinical trials to a later period. We may think of the conduct of clinical trials as the process of following cohorts of subjects who have been exposed or not exposed to the treatment of interest over a limited period of time.

See Defining therapeutic success in rheumatoid arthritis clinical trials:
from statistical significance to clinical significance, *page 411*

Table 1. Response criteria for rheumatoid arthritis clinical trials.

Response Criterion	Requirements
ACR	20% improvement in joint tenderness and joint swelling counts and 3 of 5 other measures: physician's global assessment, patient's global assessment, acute-phase reactant (ESR), patient self-assessed disability (e.g., Health Assessment Questionnaire), and patient's pain assessment[11]
Paulus	20% improvement in 4 of 6 measures: joint tenderness scores, joint swelling scores, physician's global assessment, patient's global assessment, ESR, and morning stiffness[10]
EULAR	Response criteria are based on change of Disease Activity Score (DAS) from baseline and the DAS score attained at followup. DAS is calculated as follows: DAS = 0.54 (sq rt [Ritchie Articular Index]) + 0.065 (swollen joint count) + 0.33 (ln ESR) + 0.0072 (General Health Status)[13] Change of DAS from baseline is categorized as good improvement > 1.2; moderate > 0.6 but ≤ 1.2; no improvement ≤ 0.6. The DAS score is divided into 3 categories: low disease activity ≤ 2.4; moderate disease activity > 2.4 and ≤ 3.7; high disease activity > 3.7 Responses have been defined as follows: good > 1.2 improvement in the DAS from baseline and followup DAS ≤ 2.4; nonresponse ≤ 0.6 or improvement of > 0.6 but ≤ 1.2 and DAS at followup of > 3.7; otherwise the patients would be classified as moderate responders[7]

Cohort, period, and age effects may occur and produce an effect on the magnitude of response[14]. Response may be affected by the age of individuals in a trial regardless of date of birth of the subjects (the age effect). For example, assume that the average age of individuals participating in trials is higher in 2001 than in 1991. Further, let us assume that older individuals tend to be less responsive to the treatment. The result would be that the response rate would tend to be lower in the later trials. Another problem is the potential cohort effect, where the response rate may depend on the year of birth regardless of age. For example, a cohort born during a war may suffer longterm effects of malnutrition and trauma that could influence response rates. A period effect could result in the response rate varying by the calendar time regardless of the age or birth cohort. An example of a circumstance resulting in a period effect would be the change in available medications between 1991 and 2001. Earlier trials for the treatment of RA were performed with a background of nonsteroidal antiiinflammatory agents and analgesics. Currently, the background therapy may include methotrexate, which would raise the comparison group's response rate and decrease the detectable difference given the same magnitude of treatment effect. The extent of improvement that should be required in testing of newer, more powerful agents has been revisited[15].

Another problem is that the rules for determining whether or not an individual participating in a trial has responded may be applied to a substantially different setting or population from the original. For example, the CSSRD trials used to develop the criteria were mainly Phase III multicenter trials[10]. However, currently a number of Phase II trials have utilized the ACR response criteria for evaluation of new agents for the treatment of RA. These patients may have earlier or more aggressive disease than the CSSRD patients that participated in Phase III trials. Populations in different geographic regions of the world may differ in the genetics and response to treatment[18,19]. With all approaches it is also unclear what the effects of using modifications of the original outcomes may have, e.g., using a different health status measure, using a different joint scoring system from the original.

If we do not use rules based strictly on data from previously conducted trials, is there any other option? A statistical method that can be applied is the global statistical approach[20-22]. The method is described as global, because global estimates of the effect of the treatment are obtained across several outcome measures. It should be emphasized that the global statistical approaches should be distinguished from the physician's and patient's global assessment of disease activity in the ACR core outcome measures for RA.

Some global statistical approaches require classifying each patient as a success or failure on each outcome, thus requiring assumptions similar to the ACR criteria. Thus these types of global statistics have the same problems of extrapolation as previously described for the ACR criteria[20]. However, another simple global statistical approach developed by O'Brien may obviate some of the problems discussed above for the rules-based approach[21,23]. O'Brien's global statistical approach does not require using past experience to categorize patients as successes or failures, but instead uses the full range of values that each measure can take on. The global approach of O'Brien also allows easy inclusion of new measures to assess treatment benefit without a lengthy debate about what constitutes success on this new measure.

In both global statistical approaches, adjustments are made to decrease the contribution of positively-correlated outcomes to avoid overestimating benefit. (It is a well known statistical property that when a variance is estimated assuming independence when the variables depend on one another, the variance under independence is an underestimate. Computing the variance without adjustment could lead to an error in rejecting the null hypothesis.) For example, swollen joints might often be expected to be tender as well. Thus, the information obtained in swollen and tender joint scores may overlap. It is not desirable to count the overlapping data twice, since this inflates the contributions of the joint counts. In the O'Brien approach, the comparisons of the active treatment versus the placebo or control group are made by ranking patients across treatment groups separately for each outcome, then summing the ranks for each patient across outcomes to obtain a score for each patient. These scores can be easily compared between treatment groups using standard statistical tests such as t tests or rank sum tests.

Both global approaches require treatment to have a similar effect on all outcome measures (common-dose assumption). Because of the power and efficiency of the method, potentially fewer subjects are needed to detect efficacy when the common-dose assumption is met[24]. In our analysis of the Minocycline in Rheumatoid Arthritis (MIRA) trial and 2 CSSRD trials, applying the global statistical approach, we found that the method was more powerful than the ACR response criteria[24], even when the common-dose assumption was violated. No assumptions are required regarding how individual outcome measures should be weighted and combined. Thus, the global test can assess overall treatment benefit incorporating new outcome measures for RA trials that may be developed in the future, but are not currently part of the ACR data set. Also, the method can be readily applied to any number of disorders other than RA for which multiple outcome measures are required. For example, a global statistical test can easily be applied to systemic lupus erythematosus, scleroderma, or to Sjögren's syndrome. Clinical trials of treatments for Sjögren's syndrome have usually included subjective and objective oral and ocular measures as well as laboratory measures[25-27]. However, neither a widely accepted core set of measures nor response criteria have been developed[28,29]. The global statistical test method could be particularly useful for disorders such as Sjögren's syndrome for which no standard response criteria exist.

A difficulty is that the global statistical test method gives an overall comparison of benefit between the active treatment and placebo groups[24]. To understand the contribution of individual measures, in the presence of a significant global test, requires a second analysis where the effect of treatment on each outcome is estimated separately. To assist the reader of a clinical trial it would also be helpful to present the ACR and/or EULAR results for the trial as a descriptive secondary outcome measure. The global statistical test also does not allow classification of the individual patients as treatment successes or failures, although it would be possible to use the patient's total score on O'Brien's global statistical test to determine how they ranked among the other patients in the trial. After the trial, clinicians could still use rules-based criteria as a guide to assess whether individual patients in their practices are responding to the new treatment.

Clinical trials in the rheumatic diseases are a challenge because of the variability of the clinical course and fluctuations in the outcomes measured. This is compounded by the relatively modest size of the improvement required and the slow action of many currently used agents. Because of these considerations, any factors that might mitigate sample size required are of greater importance than in fields where diseases are acute and have outcomes with little variability and treatments have a rapid and large effect. We believe that multivariate analytic approaches, such as the global statistical test, have the potential for reducing sample size required in clinical trials of the rheumatic and other diseases, and may be useful in comparing the overall benefits of treatments in diseases for which agreed-upon response criteria do not exist. However, in clinical practice, the rules-based approach, when available, will remain easier to apply, and could still be used to assess responses in individual patients.

STANLEY R. PILLEMER, MD,
Gene Therapy and Therapeutics Branch,
National Institute of Dental and Craniofacial Research,
National Institutes of Health,
10 Center Drive MSC 1190, Room 1N113,
Bethesda, Maryland 20892-1190;
BARBARA TILLEY, PhD,
Chair, Department of Biometry and Epidemiology,
Medical University of South Carolina,
Charleston, South Carolina 29425, USA

Address reprint requests to Dr. Pillemer. E-mail: Pillemer@nih.gov

REFERENCES

1. Anderson JJ, Chernoff MC. Sensitivity to change of rheumatoid arthritis clinical trial outcome measures. J Rheumatol 1993;20:535-7.
2. Egger MJ, Huth DA, Ward JR, Reading JC, Williams HJ. Reduced joint count indices in the evaluation of rheumatoid arthritis. Arthritis Rheum 1985;28:613-9.
3. Fuchs HA, Pincus T. Reduced joint counts in controlled clinical trials in rheumatoid arthritis. Arthritis Rheum 1994;37:470-5.
4. OMERACT, Conference on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Proceedings. Maastricht, The Netherlands, April 29-May 3, 1992. J Rheumatol 1993;20:527-91.
5. Tugwell P, Boers M. OMERACT Conference on Outcome Measures in Rheumatoid Arthritis Clinical Trials: introduction. J Rheumatol 1993;20:528-30.
6. 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.

7. 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;21 Suppl 41:86-9.
8. Roberts RS. Pooled outcome measures in arthritis: the pros and cons. J Rheumatol 1993; 20(3):566-567.
9. Boers M, Tugwell P. The validity of pooled outcome measures (indices) in rheumatoid arthritis clinical trials. J Rheumatol 1993;20:568-74.
10. Paulus HE, Egger MJ, Ward JR, Williams HJ. Analysis of improvement in individual rheumatoid arthritis patients treated with disease-modifying antirheumatic drugs, based on the findings in patients treated with placebo. The Cooperative Systematic Studies of Rheumatic Diseases Group. Arthritis Rheum 1990;33:477-84.
11. 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.
12. 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.
13. van Gestel AM, Anderson JJ, van Riel PL, et al. ACR and EULAR improvement criteria have comparable validity in rheumatoid arthritis trials. J Rheumatol 1999;26:705-11.
14. Holford TR. Understanding the effects of age, period, and cohort on incidence and mortality rates. Annu Rev Pub Health 1991;12:425-57.
15. 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.
16. Moreland LW, Schiff MH, Baumgartner SW, et al. Etanercept therapy in rheumatoid arthritis. A randomized, controlled trial. Ann Intern Med 1999;130:478-86.
17. Moreland LW, Morgan EE, Adamson TC III, et al. T cell receptor peptide vaccination in rheumatoid arthritis: a placebo- controlled trial using a combination of V beta 3, V beta 14, and V beta 17 peptides. Arthritis Rheum 1998;41:1919-29.
18. Dawkins RL, Kay PH, Christiansen FT. Immunogenetics of rheumatoid arthritis. Ann Acad Med Singapore 1983;12:155-63.
19. Reveille JD, Alarcon GS, Fowler SE, et al. HLA-DRB1 genes and disease severity in rheumatoid arthritis. The MIRA Trial Group. Minocycline in Rheumatoid Arthritis. Arthritis Rheum 1996;39:1802-7.
20. Lefkopoulou M, Ryan L. Global tests for multiple binary outcomes. Biometrics 1993;49:975-88.
21. O'Brien PC. Procedures for comparing samples with multiple endpoints. Biometrics 1984;40:1079-87.
22. Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. Biometrics 1987;43:487-98.
23. Tilley BC, Marler J, Geller NL, et al. Use of a global test for multiple outcomes in stroke trials with application to the National Institute of Neurological Disorders and Stroke t-PA Stroke Trial. Stroke 1996;27:2136-42.
24. Tilley BC, Pillemer SR, Heyse SP, Li S, Clegg DO, Alarcon GS. Global statistical tests for comparing multiple outcomes in rheumatoid arthritis trials. MIRA Trial Group. Arthritis Rheum 1999;42:1879-88.
25. Fox RI, Dixon R, Guarrasi V, Krubel S. Treatment of primary Sjogren's syndrome with hydroxychloroquine: a retrospective, open-label study. Lupus 1996;5 Suppl 1:S31-S36.
26. Ship JA, Fox PC, Michalek JE, Cummins MJ, Richards AB. Treatment of primary Sjogren's syndrome with low-dose natural human interferon-alpha administered by the oral mucosal route: a phase II clinical trial. IFN Protocol Study Group. J Interferon Cytokine Res 1999;19:943-51.
27. Vivino FB. The treatment of Sjogren's syndrome patients with pilocarpine-tablets. Scand J Rheumatol 2001;115 Suppl:1-9.
28. Bowman SJ, Pillemer S, Jonsson R, et al. Revisiting Sjogren's syndrome in the new millennium: perspectives on assessment and outcome measures. Rheumatology (Oxford) 2001;40:1180-8.
29. Asmussen KH, Bowman SJ. Outcome measures in Sjogren's syndrome. Rheumatology (Oxford) 2001;40:1085-8.