Editorial

Defining Therapeutic Success in Rheumatoid Arthritis Clinical Trials: From Statistical Significance to Clinical Significance

New drugs developed recently for the treatment of rheumatoid arthritis (RA) have renewed our hopes for significantly changing the course of the disease. Long-lasting disease remission, however, remains an elusive goal for very few patients. For practical purposes the effectiveness of a drug thus continues to be best described by the probability of achieving the desired outcome given the specific circumstances of the patient at hand[1,2]. For a good part of the last half century the subject of much of the rheumatologic literature has been how to define such a desired outcome.

Tender and swollen joint counts and their changes were traditionally used as primary outcomes in RA clinical trials until the mid-1990s[3]. These outcomes continue to serve as the most important somatic correlate of the underlying disease process. While joint counts are suitable to capture and measure effectiveness as a property of the intervention, they are of little help in ascertaining the probability of achieving a desired outcome or success, however defined, in a given patient. The past decade has seen immense progress towards a commonly accepted definition of treatment success or, more modestly, treatment response. Taking a closer look at most of these efforts, it is apparent that clinical meaningfulness served as a major constraining criterion, to develop criteria that may better serve the needs of the practicing physician.

In addition to being constrained by clinical meaningfulness, an acceptable definition of treatment response is also constrained by the ethical mandate to subject only as many or few patients as necessary to the potentially inferior control regimen in clinical trials. To determine minimum sample size we conventionally obey the rules of statistical hypothesis testing when applied to the results of clinical experiments[4]. These taken for granted, it can be shown that the 2 identified constraints, clinical meaningfulness and minimum sample size, seem to compete against each other: A definition considered clinically meaningful by some may require a sample size that is higher than one achievable with a different definition, considered clinically meaningful by others.

That rheumatology entered the treatment response fray was hardly avoidable due to the inadequacy of tender and swollen joint counts to satisfy the constraints of clinical meaningfulness and minimum sample size within the analysis of clinical trials. To illustrate, the analysis of a typical RA trial prior to the use of response criteria was such that the difference in tender or swollen joint counts at the end of the study period was calculated and weighted by the extent of variation of joint counts, but most importantly also by the total number of patients in the study. The difference thus weighted becomes a number that is bigger, i.e., the smaller the variation of the data, the larger the number of patients in the study. This test statistic, derived from the weighted difference in joint counts, is very sensitive to the number of patients in the study. A small, clinically irrelevant difference in joint counts may result in a large test statistic if only the minimum required number of patients are recruited into the study. The criterion of "statistical significance" alone, therefore, fails to put a cap on the potential number of patients to be recruited into the study. This is particularly a weakness when analyzing trials of therapies with marginal benefit. To safeguard against falsely determining a therapy to be efficacious when it is not, it is necessary to determine a difference in joint counts that can be considered clinically meaningful. This turns out to be an impossible task because it largely depends on the number of tender or swollen joints assessed at baseline. Not only is it possible to admit to trial patients with different degrees of disease activity, but also the number of joints subjected to assessment can vary from 28 to 68, depending on the investigators' preferences. Therefore, there is no possibility to clearly determine the difference in the absolute number of tender or swollen joints that can be considered clinically meaningful.

Fortunately, there is a strong tradition of defining relevant somatic endpoints for therapeutic studies in RA. These endpoints are all characterized by: (1) the inclusion of several clinical features, such as joint counts, pain, morning stiffness, function, etc., and (2) a mix between relative and absolute improvements that determine what is clinically meaningful. In probably the first appearance of explicit criteria in the RA literature, a "satisfactory clinical response" was considered present if a patient met 5 out of the following 7 criteria (4 out

See Clinical trials, outcome measures, and response criteria, *page 407*

of 6 if the patient was not taking steroids), and all criteria had to be met to classify the patient as "remission"[5]: (1) reduction of early morning stiffness to less than 30 minutes; (2) disappearance of nocturnal rheumatic pain; (3) elimination of synovitis to a maximum of one large joint or 3 small joints; (4) reduction of corticosteroid dosage by 50%; (5) reduction of erythrocyte sedimentation rate (ESR) by at least 50% from pretreatment; (6) patient answering "yes" to: "Do you believe that the treatment has helped you and is worthwhile to continue?"; and (7) consensus of 4 physicians evaluating the treatment as a success.

Most response criteria, however, were not developed to satisfy the constraint of "minimum sample size" until the OMERACT 2 (Outcome Measures in Rheumatology Clinical Trials) Conference, where a group of international rheumatology researchers proposed a set of criteria ultimately called the "American College of Rheumatology (ACR) Criteria for Improvement"[6]. A response, according to these criteria, requires a predefined percentage change from baseline in both swollen and tender joint counts, and improvement in at least 3 of the following 5 measures: patient global assessment of disease, physician global assessment of disease, a functional status measure, patient global assessment of pain intensity, and a laboratory measure of inflammation [ESR or C-reactive protein (CRP)]. The predefined percentage change may vary between 20% (ACR 20), 50% (ACR 50), or 70% (ACR 70) improvement, with 20% being the most commonly used definition.

The ACR 20 cutoff was chosen for reasons that have to do directly with the assumptions employed when performing statistical tests to compare proportions. Findings from the landmark trial of etanercept 25 and 10 mg subcutaneously twice weekly compared to placebo may serve to illustrate the effect of these assumptions. In this trial, 9 and 4 out of 80 patients randomized to placebo qualified for an ACR 20 and ACR 50 response, respectively, at 6 months, while 46 and 31 out of 78 did so in the group of patients randomized to etanercept 25 mg. An investigator hypothesizing the observed ACR 20 difference before starting a new trial would need to recruit 18 patients, given the rules (80% power, $p < 0.05$), while hypothesizing the observed ACR 50 difference would require a sample size of 27 per group. Any lower hypothesized proportion of ACR 50 responders in the active treatment group would require a larger sample size. The authors of this study, incidentally, did not specify sample size calculations in their publication and may have subjected more patients than necessary to placebo. To a practical observer, the ACR 50 difference may appear more clinically convincing than the ACR 20 difference. In practice, however, few rheumatologists seem to discount a 20% response; otherwise it would be hard to explain that the percentage of patients continuing methotrexate in observational studies is by far larger than the percentage of patients known to respond by ACR 20 criteria in clinical trials[7].

The ACR 20 response criteria turn out to be a solution that quite elegantly satisfies the constraints of minimum sample size and clinical meaningfulness[8]. They minimize those instances where a treatment shows promise but would be considered inefficacious by the chosen cutoff; moreover, they safeguard against falsely considering a treatment to be efficacious, when it is not. Although the latter seems less a danger given the powerful drugs currently in development, it seems safer to be alert against such instances.

Currently, researchers and practicing rheumatologists are working on improved definitions to remedy some of the shortcomings of the ACR 20 criterion, in particular its relative nature and inability to capture an absolute state of disease activity. In this issue of *The Journal* Pillemer and Tilley take a position on issues related to these and other response criteria[9].

Proposals for improved criteria or alternative definitions of treatment response or success will need to document how they can maintain the practical efficiency of the ACR 20 criteria. Joint counts and other quasi-continuous measures are unlikely to be as practical or efficient. Whichever definition is proposed, it will need to help physicians teach their patients that, while treatment success is not guaranteed, the desired outcome is likely. Much progress was made by harnessing the demon of statistical significance and subordinating it to a practical, but biologically plausible understanding of clinical significance.

ANDREAS MAETZEL, MD, MSc, PhD,
Division of Clinical Decision Making,
Department of Health Policy, Management and Evaluation,
University Health Network and University of Toronto,
200 Elizabeth Street EN6-232A,
Toronto, Canada M5G 2C4.

Address reprint requests to Dr. Maetzel. E-mail: maetzel@uhnresearch.ca

REFERENCES

1. Miettinen OS. The modern scientific physician: 5. The useful property of an intervention. CMAJ 2001;165:1059-60.
2. Miettinen OS. The modern scientific physician: 1. Can practice be science? [comment]. CMAJ 2001;165:441-2.
3. Suarez-Almazor ME, Belseck E, Shea B, Wells G, Tugwell P. Methotrexate for treating rheumatoid arthritis (Cochrane Review). In: The Cochrane Library, Issue 4. Chichester, UK: John Wiley & Sons, Ltd.; 2003.
4. Goodman SN. p values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate [see comments]. Am J Epidemiol 1993;137:485-96; discussion 497-501.
5. Rothermich NO, Philips VK, Bergen W, Thomas MH. Chrysotherapy. A prospective study. Arthritis Rheum 1976; 19:1321-7.
6. 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.
7. Maetzel A, Wong A, Strand V, Tugwell P, Wells G, Bombardier C. Meta-analysis of treatment termination rates among rheumatoid arthritis patients receiving disease-modifying anti-rheumatic drugs. Rheumatology Oxford 2000;39:975-81.
8. 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.
9. Pillemer SR, Tilley BC. Clinical trials, outcome measures and response criteria [editorial]. J Rheumatol 2004;31:407-10.