

Development of Classification and Response Criteria for Rheumatic Diseases

CLASSIFICATION AND RESPONSE CRITERIA SUBCOMMITTEE OF THE AMERICAN COLLEGE OF RHEUMATOLOGY COMMITTEE ON QUALITY MEASURES

Relevance to the clinician. Clinicians already know that not all patients who are diagnosed with rheumatic diseases really have them. Moreover, determining which patients have improved and by how much is also difficult. Classification criteria allow clinical researchers to recruit patients with similar diseases (e.g., rheumatoid arthritis or systemic lupus erythematosus) into studies. Response criteria help to determine whether treatments really work, i.e., whether they actually produce clinically important improvement. As the science of clinical research advances, we must update our standards for considering classification and response criteria. In this editorial, members of the American College of Rheumatology (ACR) Subcommittee on Classification and Response Criteria describe the purpose of criteria sets, their development and validation, and the role of the ACR in adopting them.

Many rheumatic diseases are characterized by overlapping organ system involvement, lack of a pathognomonic diagnostic test, and widespread manifestations. These characteristics hinder easy diagnosis or recognition of changes in disease status, and also present a tremendous challenge when conducting and evaluating the results from clinical research.

Criteria that classify a disease (classification criteria), define disease activity, or specify a measurement of change in disease activity in response to an intervention (response criteria) are critical for conducting most types of clinical studies with the aim of improving patient care. These studies range from epidemiologic investigations to clinical trials. The classification criteria ensure recruitment of patients with a similar clinical entity, and the response criteria allow for standardized definitions of response or outcomes. In this article, members of the American College of Rheumatology (ACR) Subcommittee on Classification and Response Criteria discuss the different types of criteria

sets, the methods for developing and validating such criteria sets, and the role of the ACR with respect to such activities. In addition, “common pitfalls” encountered when developing and/or validating such criteria sets are identified.

The Role of Criteria Sets in Rheumatic Disease

Role of classification criteria. Classification criteria help to distinguish patients with the disease in question from those without the disease. The criteria are used to standardize disease definitions across geographically diverse centers and across studies to ensure that the same disease entity is consistently studied. Examples are the ACR (formerly the American Rheumatism Association [ARA]) 1987 revised classification criteria for rheumatoid arthritis (RA) (1) and the 1982 revised classification criteria for systemic lupus erythematosus (2). Criteria sets that define disease entities for clinical studies are generally not described as diagnostic criteria but rather as classification criteria. Classification criteria are often useful teaching aides for trainees and will almost always mirror the list of criteria that one uses for diagnosis, but they are not synonymous with diagnostic criteria. In a typical clinical setting, meeting prespecified criteria is not required for diagnosis. Rather, a diagnosis results from a clinical evaluation for features that suggest the presence of disease. Some experts advocate the use of diagnostic criteria as a screening tool for primary care, especially when trying to identify patients in the early stages of rheumatic conditions.

Role of response criteria. Response criteria help standardize the measurement of clinical status over time. In the context of a clinical trial, response criteria may facilitate the determination of the efficacy of a particular drug or a treatment strategy and can help standardize this assessment across trials. There are 2 broad types of response

The work of the Classification and Response Criteria Subcommittee is supported by the American College of Rheumatology.

Members of the Classification and Response Criteria Subcommittee of the American College of Rheumatology Committee on Quality Measures: Jasvinder A. Singh, MD, MPH (Co-Chair), Daniel H. Solomon, MD, MPH (Co-Chair), Maxime Dougados, MD, David Felson, MD, MPH, Gillian Hawker, MD, MPH, Patricia Katz, PhD, Hal Paulus, MD, Carol Wallace, MD.

The American College of Rheumatology is an independent, professional, medical and scientific society which does not guarantee, warrant, or endorse any commercial product or service.

Address correspondence to Daniel H. Solomon, MD, MPH, Brigham and Women's Hospital, 1620 Tremont Street, Suite 3030, Boston, MA 02120. E-mail: dhsolomon@partners.org

Submitted for publication March 3, 2006; accepted in revised form March 9, 2006.

Table 1. Recommendations for development and validation of criteria sets

<p>Development</p> <ol style="list-style-type: none"> 1. The list of possible inclusion and exclusion criteria should be developed using appropriate consensus methodology. 2. Each of the potential criteria should be reliable (reproducible), precise in its measurement, easy and feasible to measure, and clinically sensible. 3. Potential criteria redundancy should be assessed and minimized. <p>Validation</p> <p>Classification criteria</p> <ol style="list-style-type: none"> 1. Selection of cases (patients considered to have disease): <ol style="list-style-type: none"> A. Cases should be chosen to include the spectrum of disease severity. B. If the criteria are to be used for epidemiologic studies, both clinical and community cases should be included. 2. Selection of controls (patients considered not to have the disease): <ol style="list-style-type: none"> A. Controls should be chosen with a view to the intended purpose of the criteria, i.e., to distinguish individuals with disease from those without disease versus to distinguish individuals with a particular rheumatic disease from individuals with other rheumatic diseases. Ideally, multiple control groups should be used. 3. An adequate number of cases and controls should be chosen. (Adequate must be defined in the context of the sensitivity and specificity of a given criteria set.) 4. For each individual criterion, and for combinations of criteria, the sensitivity and specificity for detecting and ruling out disease should be calculated. These results, together with clinical opinion, should be used to reduce the number of criteria for inclusion. 5. Criteria to be included should be those with the greatest demonstrated validity, sensitivity, and specificity. 6. Acceptable statistical approaches should be used to create the classification criteria from the reduced number of criteria. 7. The final classification criteria should be validated in samples of cases and controls distinct from the patients used to develop the criteria. <p>Response criteria</p> <ol style="list-style-type: none"> 1. An appropriate statistical method should be used to identify the number of factors or constructs represented by the various criteria. 2. An appropriate statistical approach should be used to determine which factors/elements differentiate patients across the spectrum of disease severity. 3. The final disease response criteria should be validated in an independent prospectively collected clinical trial sample.

criteria: disease state and change in disease state. Although these are linked, they may be measured using different criteria. For example, the ACR20 is a well-accepted measure of change (improvement) in RA disease state (3). Although measures of disease state and change may include many of the same elements from the ACR RA core set (4), the elements are defined differently. Therefore, to measure remission in individuals with RA (a disease state), one can use the ARA RA remission criteria (5) but not the ACR20 criteria (3). The 2 measures both include many of the same elements from the ACR RA core set, but they are defined differently. An important consideration regarding response criteria is whether they should be measured using a continuous (or ordered) scale versus a dichotomous measure. The ACR20 is a dichotomous measure (whether the patient achieved a 20% improvement or did not achieve a 20% improvement). Measures such as the Disease Activity Score can be used as either a continuous or ordered scale (6). Dichotomous outcomes may be easier to explain to patients and colleagues, whereas continuous (or ordered) measures have greater statistical power to detect small differences across a larger range of outcomes. Common pitfalls include not considering continuous (or ordinal) measures when developing response criteria sets, and not specifying how response criteria should be presented (i.e., at the individual or group level) and when they should be measured (i.e., any time during treatment or at the end of a treatment trial).

Development and Validation of Criteria Sets

Specific recommendations for developing and validating classification and response criteria are outlined in Table 1, and a checklist for developing criteria sets is shown in Appendix A. In the past, classification and response criteria were often developed by a limited panel of experts. More recently, the use of consensus methodology (such as Delphi questionnaires followed by meetings using the Nominal Group Technique) has resulted in a more inclusive process that can involve a large number of participants (7,8). This approach allows for international, collaborative, and broadly based efforts that have a high likelihood of resulting in criteria sets that will be widely accepted and utilized.

Classification criteria. Specific guidelines from the Subcommittee on Classification and Response Criteria are modified from those recommended by Felson and Anderson (9). The first step in developing classification criteria is to create a list of potential criteria (both inclusion and exclusion criteria) followed by identification of a large sample of patients with the disease and a comparable number of controls without the disease. These criteria are then applied to determine which criteria (or combination of criteria) best differentiate disease state from nondisease state (10). Once a final set of criteria has been developed, the recommendation is to test its validity in a large sample of subjects distinct from the original sample used for criteria set development.

Table 2. Validity and other methodologic characteristics relevant to criteria sets (adapted from reference 13)

Characteristic	Definition
Face validity	Credibility: Are the criteria sensible?
Content validity	Comprehensiveness: Do the criteria sample all of the domains of the disease?
Criterion validity	Do the criteria predict or correlate or agree with a “gold standard”?
Sensitivity and specificity	Do the criteria identify those with the clinical construct, i.e., disease? Do the criteria identify those without the disease in question?
Responsiveness	Are the criteria sensitive to change? Do the criteria detect the smallest clinically important change?

Validity issues are central to criteria development (9) and are described in Table 2. Face and content validity are routinely assessed during the development of classification criteria. However, for classification criteria, sensitivity, specificity, and criterion validity are most important to consider. Sensitivity refers to how well the criteria set identifies persons with a specific clinical entity, and specificity refers to whether the criteria are able to exclude persons with other clinical entities. For example, some patients with psoriatic arthritis may be included in studies of RA because the ARA RA classification criteria lack sufficient specificity, highlighting the importance of careful selection of nondisease controls when validating classification criteria. One wants to choose control subjects who have diseases similar to the one under study to better test specificity. Finally, criterion validity describes the level of concordance or agreement between the developed criteria set and a gold standard, often a cohort of patients that expert clinicians agree have a given disease. Such patient cohorts should represent the spectrum of a given disease, ideally are actual patients, and should be distinct from any patient cohort used to develop the criteria set.

Response criteria. Development and validation of response criteria present issues similar to those for classification criteria, with the addition that sensitivity to change, or responsiveness is important in response criteria. Defining responsiveness is conceptualized as a process of data-driven consensus using 2 sources of information: data derived from reanalysis of trials, and consensus based on clinician choice of instrument and on clinician’s identification of a clinically important response. The process is most likely to produce a durable and widely accepted response definition if both clinical and analytic inputs are used. In the first process, using actual patient data from a clinical trial, experts rate the response of given patients. Then the proposed criteria set is tested against the experts’ rating to determine whether the criteria have adequate ability to discriminate between patients with important clinical improvement and those without. In the second process, the discriminant validity of the candidate definitions of response is tested. This is best measured using data from methodologically sound clinical trials, in which the tested treatment demonstrated efficacy.

- Although criteria sets can be validated using data collected outside of a treatment trial, such criteria would be considered provisional by the ACR Subcommittee until tested in the context of a trial (see below). Once validated using prospectively collected trial data, an ACR Provisional Criteria Set would be considered a fully approved ACR Criteria Set (see below). One common pitfall would be considering criteria sets with face and content validity as fully validated; quantitative validity testing is critical. A second common pitfall is using circular reasoning, where the same group of experts who develop criteria also validate the criteria; appropriate validation should be ensured by using different sets of experts and actual patient scenarios. A third common pitfall is using clinical consensus to arrive at a threshold to define response using a single-response variable (e.g., at least a 20% improvement in pain). Outcome measurement is likely to be more sensitive to change if this measure is analyzed as a continuous or ordered variable, not simply a dichotomous threshold. An exception to this is when an index of individual measures (e.g., ACR20) is dichotomized rather than existing as just one measure.
- It is important to recognize the important role that response criteria play in clinical trials. Validated response criteria allow investigators, clinicians, regulators, and patients to determine the efficacy (or lack thereof) of a given intervention and to communicate about response using the same metric. However, response criteria can be used inappropriately, i.e., measured at incorrect time points or applied to the wrong populations. One common pitfall is not clarifying how response criteria should be applied; how and when response criteria should be assessed in the course of a treatment trial need to be specified. Another common pitfall is that eligibility criteria for trials may not parallel response criteria; how response criteria relate to trial eligibility should be recommended (e.g., should morning stiffness be part of eligibility criteria for a trial if it is not part of the response measure?).

Role of the American College of Rheumatology

One of the subcommittees of the newly constituted ACR Quality Measures Committee is the Subcommittee on Clas-

sification and Response Criteria. This subcommittee is responsible for encouraging development and validation of new and improved classification and response criteria sets for various rheumatic conditions. Although many prior criteria sets were approved by the ACR without the broad-based development and rigorous testing described above, measurement science has advanced and rheumatology must follow by adopting a more inclusive and robust approach. To this end, the ACR has developed a policy that criteria sets will require quantitative validation prior to receiving ACR approval (11). Criteria sets without quantitative validation will no longer be approved by the ACR as being "preliminary." However, the ACR strongly encourages publication of such proposed criteria sets so that a variety of researchers can test and hopefully validate them. Response criteria sets that have undergone quantitative validation using data from previously collected cohorts outside of a trial setting can be approved by the ACR and will be termed ACR Provisional Criteria Sets to recognize that they require further validation using data from trials. Criteria sets that have undergone prospective validation and testing in an external data set can be considered ACR Criteria Sets.

The ACR Subcommittee is still grappling with how to balance the clinician's and patient's perspectives in response criteria. Clinicians may value certain aspects of a patient's disease status (i.e., inflammation and structure), whereas patients place more importance on pain and daily function. These various perspectives may underlie subtle differences between determining the efficacy of a given treatment in a trial and how treatment decisions are made in typical practice.

There is great enthusiasm for having criteria sets adopted simultaneously by the ACR and the European League Against Rheumatism so that international research can be harmonized in a meaningful way. This is demonstrated by publication in this issue of *Arthritis Care & Research* of a criteria set for juvenile systemic lupus erythematosus that was developed by a European-based research group, the Pediatric Rheumatology International Trials Organization, and approved by the ACR Board of Directors (12). The ACR is well aware of other organizations' efforts regarding classification and response criteria and will attempt to support such efforts whenever possible. Groups working on classification and response criteria for rheumatic disease should look to the ACR Subcommittee on Classification and Response Criteria as a resource. This subcommittee wants to partner with such groups to ensure that rigorous methods, such as those outlined above, are used when developing and validating criteria sets.

This is an exciting era for rheumatologists. The pace of drug discovery for rheumatic conditions is quickening, and the need for carefully constructed classification and response criteria sets is growing. The ACR hopes to facilitate this process so that additional safe and effective treatments can be efficiently brought to our patients' bedsides.

REFERENCES

1. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum* 1988;31:315-24.
2. Tan EM, Cohen AS, Fries JF, Masi AT, McShane DJ, Rothfield NF, et al. The 1982 revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum* 1982;25:1271-7.
3. Felson DT, Anderson JJ, Boers M, Bombardier C, Furst D, Goldsmith C, et al. American College of Rheumatology preliminary definition of improvement in rheumatoid arthritis. *Arthritis Rheum* 1995;38:727-35.
4. Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. *Arthritis Rheum* 1993;36:729-40.
5. Pinals RS, Masi AT, Larsen RA, and the Subcommittee for Criteria of Remission in Rheumatoid Arthritis of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee. Preliminary criteria for clinical remission in rheumatoid arthritis. *Arthritis Rheum* 1981;24:1308-15.
6. 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.
7. Taylor WJ. Preliminary identification of core domains for outcome studies in psoriatic arthritis using Delphi methods. *Ann Rheum Dis* 2005;64 Suppl 2:ii110-2.
8. Jamieson M, Griffiths R, Jayasuriya R. Developing outcomes for community nursing: the Nominal Group Technique. *Aust J Adv Nurs* 1998;16:14-9.
9. Felson DT, Anderson JJ. Methodological and statistical approaches to criteria development in rheumatic diseases. *Baillieres Clin Rheumatol* 1995;9:253-66.
10. Bloch DA, Moses LE, Michel BA. Statistical approaches to classification: methods for developing classification and other criteria rules. *Arthritis Rheum* 1990;33:1137-44.
11. American College of Rheumatology policies and procedures for quality measures. URL: <http://www.rheumatology.org/practice/qmc/PolicyProcedure.asp>.
12. Ruperto N, Ravelli A, Oliveira S, Alessio M, Mihaylova D, Pasic S, et al. The Pediatric Rheumatology International Trials Organization/American College of Rheumatology Provisional criteria for the evaluation of response to therapy in juvenile systemic lupus erythematosus: prospective validation of the definition of improvement. *Arthritis Rheum* 2006;55:355-63.
13. Katz PP. Introduction to special patient outcomes in rheumatology issue of *Arthritis Care & Research*. *Arthritis Rheum* 2003;49 Suppl 5:S1-4.

APPENDIX A: CHECKLIST FOR DEVELOPING CRITERIA SETS

Diagnostic/classification criteria development checklist:

1. Will a comprehensive list of possible criteria be considered (content validity)?
2. Will each of the potential criteria be reliable (reproducible), precise in its measurement, easy to measure, and clinically sensible?
3. Are the potential criteria redundant (i.e., highly correlated)? Will this be assessed?
4. Selection of cases (patients considered to have the condition of interest):
 - A. Will cases be chosen across the spectrum of disease severity?
 - B. If the criteria are to be used for epidemiologic studies, will both clinical and community cases be included?
5. Selection of controls (patients considered not to have the condition of interest):
 - A. Will the controls be chosen with a view to the intended purpose of the criteria, i.e., to distinguish individuals with disease from those without disease versus to distinguish individuals with a particular rheumatic disease from individuals with other diseases? Ideally, multiple control groups will be used.
6. Will at least 100 cases and 100 controls be chosen?
7. For each individual criteria, and for combinations of criteria, will the sensitivity and specificity for detecting and ruling out the disease of interest be calculated (construct validity, convergent and divergent validity)? Will these results, together with clinical opinion, be used to reduce the number of criteria for inclusion?
8. Are the criteria to be included those with the greatest content and construct validity? How will this be demonstrated?
9. Will acceptable statistical approaches (see reference 9) be used to create the diagnostic/classification criteria from the reduced number of criteria?
10. Will the final diagnostic/classification criteria be validated in different samples of cases and controls? How will those other samples be chosen?

Response criteria/disease severity/damage criteria development checklist:

1. Will a comprehensive list of criteria for potential inclusion be developed (content validity)? How?
 2. Will each of the chosen elements be reliable (reproducible), precise, easy to measure, and clinically sensible?
 3. Selection of cases (patients with the given disease):
 - A. Will cases be chosen across the spectrum of disease severity or damage?
 - B. If the criteria are to be used for epidemiologic studies, will both clinical and community cases be included?
 4. Will at least 200 cases be chosen?
 5. Will the chosen criteria be redundant (i.e., highly correlated)? How will this be examined?
 6. Will an appropriate statistical method (see reference 9) be used to identify the number of factors or constructs represented by the various criteria?
 7. Will an appropriate statistical approach be used to determine which factors/elements differentiate patients across the spectrum of disease severity?
 8. Will the final disease severity/damage criteria be validated in an independent sample?
-