**Arthritis & Rheumatology**
Vol. 0, No. 0, Month 2026, pp 1–10
DOI 10.1002/art.70207


American College of Rheumatology
*Empowering Rheumatology Professionals*

# The American College of Rheumatology Updated Recommended Methods for Development of Disease Classification and Response Criteria

Sindhu R. Johnson, MD, PhD,[1,2,3] Tuhina Neogi, MD, PhD,[4] Amy S. Turner,[5] Jaime Guzman, MD, MSc,[6] and Zahi Touma, MD, PhD[1]

Disease classification criteria and response criteria are important tools in the conduct of observational studies and clinical trials for rheumatic diseases. This paper updates methodologic guidance for classification and response criteria funded, developed and/or endorsed by the American College of Rheumatology (ACR), incorporating contemporary methods and bias mitigation strategies. As current and previous committee chairs and staff charged with overseeing ACR criteria development and endorsement, we highlight guiding principles, outline a methodological approach, provide relevant examples, and emphasize considerations for ACR funding and endorsement. We provide an updated checklist to guide investigators and reviewers. The ACR Criteria Subcommittee, ACR Quality of Care Committee and the ACR Board of Directors have approved this paper.

*This criteria set has been approved by the American College of Rheumatology (ACR) Board of Directors. This signifies that the criteria set has been quantitatively validated using patient data, and it has undergone validation based on an independent data set. All ACR-approved criteria sets are expected to undergo intermittent updates*.

*The ACR is an independent, professional, medical and scientific society that does not guarantee, warrant, or endorse any commercial product or service*

## Introduction

Disease classification criteria and response criteria are key tools for observational studies and clinical trials.[1–3] Rheumatologists integrate patient symptoms and signs with laboratory and/or radiographic findings to make a diagnosis. This relies on training in pattern recognition, understanding disease constructs, and knowledge of mimicking conditions in the differential diagnosis. However, disease presentations can vary widely. While rheumatologists are trained to recognize the variations for accurate diagnosis, such heterogeneity complicates research, which requires specificity, accuracy, and homogeneity. Variation in presentation can reduce precision when estimating interventions' effect, or obscure true improvement. To enable comparison across studies, patient populations and outcome measures must be consistent. This is achieved using validated classification and response criteria. Reflecting their importance, the American College of Rheumatology (ACR) has supported the development and/or endorsement of such criteria for many rheumatic diseases.[4–18]

**Classification versus diagnostic criteria.** Classification criteria are developed for research.[1–3,19] They focus on key features rather than all diagnostic possibilities, and typically prioritize specificity to minimize false positives, such as enrolling participants without the disease in trials. The ACR does not support their use for diagnosis, though some components may inform practice.[19] Diagnosis remains the physician judgment, incorporating factors like rare manifestations, partial presentations, and overlapping conditions.[19] Insurers should not use classification criteria to restrict therapy, as they may not capture the full disease spectrum.

**Response criteria.** Response criteria provide a valid way to measure clinically meaningful change when a true change has occurred ("signal"), distinguishing it from random variation ("noise").[3,20] Well-developed response criteria offer measurable

[1]University of Toronto, Toronto, Ontario, Canada; [2]Schroeder Arthritis Institute, Toronto Western Hospital, Toronto, Ontario, Canada; [3]Mount Sinai Hospital, Toronto, Ontario, Canada; [4]Boston University School of Medicine, Boston, Massachusetts; [5]American College of Rheumatology, Atlanta, Georgia; [6]The University of British Columbia, Vancouver, British Columbia, Canada.

Additional supplementary information cited in this article can be found online in the Supporting Information section (https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/art.70207).

Author disclosures are available at https://onlinelibrary.wiley.com/doi/10.1002/art.70207.

Address correspondence via email to Sindhu R. Johnson, MD, PhD, at Sindhu.Johnson@uhn.ca.

Submitted for publication November 17, 2025; accepted in revised form April 13, 2026.

endpoints for interventions in clinical trials (eg, for drug development and approval) and in observational studies. They may reflect either achieving a desirable disease state (eg, rheumatoid arthritis remission[21]) or changes in disease activity (improvement and worsening, eg, ACR 20% improvement[22]). These criteria use validated, standardized definitions of clinical status over time and should prioritize distinguishing effective treatments from placebo.

In summary, classification criteria serve as inclusion criteria for research studies, while response criteria define meaningful change in outcomes.

## Guiding principles

The ACR emphasizes rigorous methodology in developing classification and response criteria and has published guidance documents.[1–3] Methods have evolved, increasing rigor over time. The 2010 rheumatoid arthritis (RA) classification criteria introduced multicriteria decision analysis (also called conjoint analysis or discrete choice experiment) to weight items, and implement bias mitigation strategies (eg, use of labels that do not denote age or sex when describing patient profiles, such as "portico", "cupola", or other architectural labels),[23,24] producing a numeric additive point system. Systemic sclerosis (SSc) classification criteria applied Kirschner and Guyatt's methodologic framework,[25–27] grounded in measurement science,[28] with phases including construct definition, item generation, item reduction, weighting, threshold setting, and validation. Biases mitigation is essential throughout. These approaches improved performance and value of modern criteria.[4–18]

This paper updates guidance based on 4 principles.

1. Collaboration among clinicians, clinical epidemiologists (methodologists) and people affected by the rheumatic condition.
2. Balanced use of expert-based and data-driven methods.
3. Avoidance of circular reasoning.
4. Mindful selection of reference standards.

**Collaboration.** Criteria should be developed jointly by expert clinicians, clinical epidemiologists and patients.[2,3] Clinicians bring disease expertise; clinical epidemiologists ensure sound methods and bias control; patients inform meaningful outcomes.[29,30] Teams typically include a core group and separate panels for derivation and validation.

**Balanced methods.** Expert input ensures relevance and acceptance but may introduce bias, mitigated by diverse backgrounds and structured consensus methods (eg, Delphi methods,[31,32] nominal group technique[33,34] and multicriteria decision analysis[24,35]). Data-driven methods use cohorts or trials. Assembly of a new cohort may be necessary if new items are considered for a criteria system,[36] which typically occurs after new items have been identified and defined. Purely data-driven methods may yield spurious or impractical findings. Therefore, the ACR recommends a balanced use of both expert-based methods and data-driven methods to improve validity.[1–3]

**Avoidance of circular reasoning.** Historically, experts developing the criteria contributed patient data and set the reference standard for criteria testing.[1] These criteria performed well when internally validated, but often did not perform well when externally validated in independent data sets.[1–3] The ACR recommends the experts who develop the criteria be separate from the experts who contribute data or set the reference standard against which the criteria are tested and validated.[1–3]

**Reference standards.** Criteria should be tested against meaningful standards, often expert adjudication or outcomes like disease progression, to demonstrate that the new criteria are meaningfully better than previous criteria. For classification criteria, this is achieved by comparing the sensitivity and specificity of the new and old criteria. For example, an often-used reference standard is the adjudication of each case as a disease or mimicker by an independent panel of three or more experts who have not been involved in the development of the criteria. As an example for response criteria, the RA remission criteria were judged against the subsequent development of damage or functional deficits.[21,37]

## Types of criteria in a system

There may be different types of criteria within classification and response criteria systems (Table 1, Figure 1). However, a criteria system does not need to include all of the following types of criteria.

***Entry criterion*** refers to an item that must be present for the rest of the criteria system to be applied; if absent, the remainder of the criteria should not be applied.

***Exclusion criterion*** refers to an item that if present means the patient should not be classified as having the disease.

***Absolute*** or ***sufficient criterion*** is an item, that if present, means the patient is classified as having the disease (without need for the other criteria).

***Additive criteria*** refer to items that should be scored if present, and whose weights will be added together to produce a total score.

For example, in the gout classification criteria, monosodium urate crystals and in the RA classification criteria erosive disease characteristic of RA were positioned as a sufficient criterion, respectively. The positioning of criteria may be refined further. For example, for calcium pyrophosphate deposition disease classification criteria, crowned dens syndrome was added as 'sufficient criterion' in the criteria refinement phase. However, the

**Table 1.** Types of criteria within a disease classification criteria system*

| Criteria type | Definition | Example |
|---|---|---|
| **Entry criteria** | If present, apply the rest of the criteria system.<br>If absent, do not apply the rest of the criteria system. The patient should not be classified as having the disease. | Antinuclear antibodies (ANA) at a titer of >1:80 on HEp-2 cells or an equivalent positive test (ever).<br>If present, then the rest of the criteria system may be applied for EULAR/ACR SLE classification criteria.[9] |
| **Exclusion criteria** | If present, do not apply the rest of the criteria system. The patient should not be classified as having the disease. | Do not count a specific criterion toward the criteria system if there is a more likely alternative explanation than SLE for it.[9] (e.g., fever due to an infection) |
| **Absolute (or sufficient) criterion** | If present, the patient can be classified as having the disease. | Skin thickening of the fingers of both hands extending proximal to the metacarpophalangeal joints. Weight 9 points. Classification of systemic sclerosis threshold 9.[26] |
| **Additive criteria** | The weights of each criterion present should be added together. | 2010 ACR/EULAR Rheumatoid arthritis classification criteria require a score of ≥6 to be classified as having rheumatoid arthritis.[23] |

* The ACR does not require all types of criteria in a criteria system. ACR, American College of Rheumatology; SLE, systemic lupus erythematosus.

positioning of the different types of criteria must be determined prior to the validation of the final criteria system.

## Methodologic approach

The ACR recommends five phases (Figure 2, Table 2).

1. Construct definition.
2. Item generation.
3. Item reduction.
4. Criteria system derivation and refinement.
5. Criteria system validation.

**Construct definition.** Constructs may evolve with advances in disease understanding. For example, a 20% improvement in the RA core set was considered a good response,[22] but energies are now focused on attaining disease remission.[21,22] Defining the construct underlying the disease can be challenging, as the construct is latent, ie, known to exist but difficult to define or measure. The 2010 RA criteria group reconceptualized RA as an inflammatory arthritis for which most rheumatologists would initiate hydroxychloroquine, as compared to the classical concept of a symmetrical, inflammatory, polyarthritis affecting the small joints of the hands.[23] As an example for response criteria, in 1996, the systemic lupus erythematous (SLE) damage index group conceptualized damage as a non-reversible change, unrelated to inflammation, occurring since the diagnosis of SLE, ascertained by clinical assessment and present for at least six months.[38] Using qualitative methods, the construct underlying SLE damage has been modernized, recognizing that the functional consequences of organ damage may improve through physiologic adaptation or treatment.[39,40] Similarly, the SSc-subset criteria group found that subdividing SSc into limited and diffuse categories risks misclassification and the predictability of end-organ damage does not always hold true.[41] Novel factors underlying the construct of SSc-subsets include rate of skin thickening change and autoantibodies.[41]

The construct should not be defined by a few individuals. Construct definition may require literature review,[39] qualitative[40,41] or survey methods to capture a range of perspectives. Using the

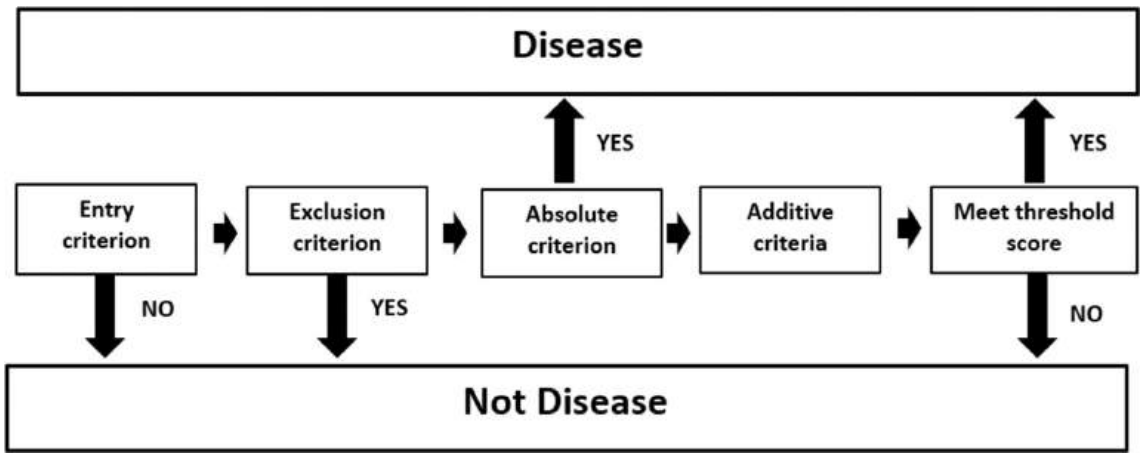


**Figure 1.** Illustration of how types of criteria contribute to classification.

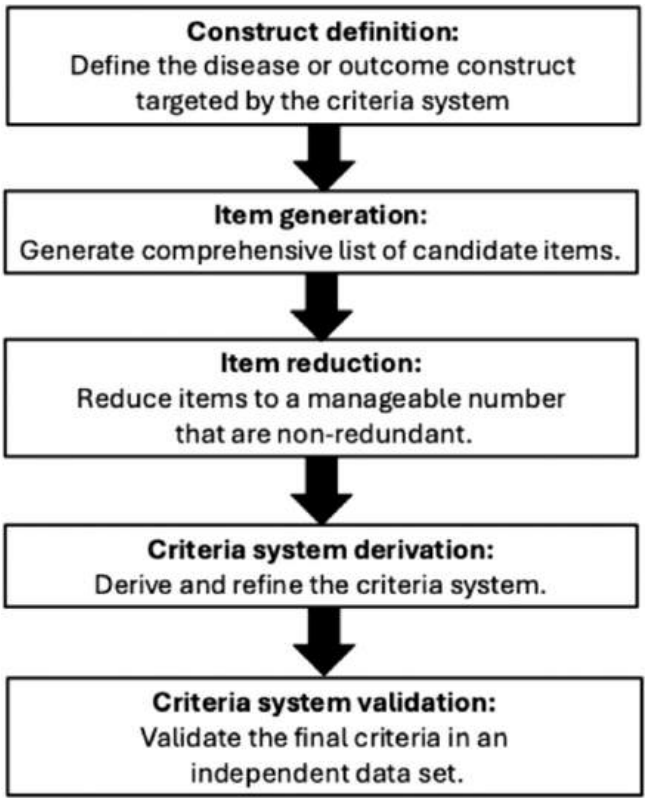


**Figure 2.** Methodologic framework for classification and response criteria development.

Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines for reporting systematic reviews,[42] Scoping Reviews[43] or reporting systematic reviews of outcome measurement instruments[44] enhances transparency and completeness in reporting. Discussions are needed to identify populations for whom the criteria will apply versus populations that are beyond the scope of the criteria. For example, antinuclear antibody negative SLE was determined to be outside the scope of the 2019 SLE classification criteria.[33]

**Item generation.** Multiple methods could be used to identify all items that may be included as criteria. Methods include systematic or scoping reviews,[45,46] or observational data.[47,48] For example, for the 2019 SLE classification criteria, cases and controls with early disease were evaluated for items that discriminate between the two groups.[48] Unexplained fever distinguished early SLE from mimickers with early disease.[48] Another method may include surveying patients and/or clinicians.[49] For example, a cross-sectional survey of SLE patients demonstrated the importance of fever, fatigue, and arthritis for classification of SLE in early disease.[49] This methodologic approach is useful to identify disease features that are important to patients or identifies disease features that clinicians may not have thought of as important parts of the disease.[49] Delphi exercises[31,32] ask a panel of experts to list all items they may consider.[50,51] During the first round of a Delphi exercise participants are asked to nominate additional items that should be considered or added to the list of items identified from previous steps. This may help identify novel items not included in published sources. Delphi exercises are often internet based, allowing for asynchronous completion, and facilitate the inclusion of many investigators from around the world, including those outside of rheumatology.[50,51] Efforts should be undertaken to ensure a response rate of greater than 80% at each Delphi round.[52] Methodologic strategies that have been shown to increase response rate include reminder emails, personalized communication from individuals known to the panelist, and incentives.[52]

**Item reduction.** Several methods could be considered to reduce the number of items to a manageable number. The prevalence, feasibility, sensitivity, specificity, face and content validity, reliability, and responsiveness (in the case of response criteria) of the items could inform item reduction, with poorly performing items being discarded.[36,47,53] This information may be obtained from the literature or cohort data. The added value of including items that rarely occur needs to be balanced by the impact of including too many items on the feasibility of the final criteria system.[47] Reliability is an important consideration for laboratory tests, as there may be substantial variations across laboratories. For example, suboptimal reliability across laboratories led to the exclusion of novel antibodies for the classification of SLE; whereas antinuclear antibody testing by HEp-2 cells or equivalent assay was reliable across laboratories and was retained.[9,46] Follow-up rounds in a Delphi exercise (started for item generation)[31,32] is a common method for item reduction.[50,51] Participants are asked to rate candidate items for appropriateness/importance, for example using a Likert scale of 1 to 9, where one is least appropriate or lowest importance and 9 is most appropriate or highest.[50,51] Items with high appropriateness scores (eg, 7 to 9) may be retained; whereas items with lower scores may be discarded. In rounds 2 and 3 of the Delphi, the overall mean appropriateness scores of each item as well as the participants' own scores from the prior round are presented to participants, to allow them to revise their scores. This is a distinguishing feature of the Delphi method as it allows the group to come to consensus. Another approach to facilitate item reduction is to study the association of candidate items to relevant reference standards (eg, a subsequent diagnosis after adequate follow-up time, or demonstration that no erosive inflammatory arthritis consistent with RA damage develops).[21] The nominal group technique (NGT)[54] is another consensus-based method that may be used for item reduction.[33,34,40,51] Using NGT, a smaller group of experts with a facilitator are asked to discuss each item in turn. The facilitator ensures that each person is allowed to speak once in favor or against discarding the item. The facilitator may allow rebuttals or allow each person to speak for a second time. Items may be removed based on group consensus.

**Table 2.** Methods to consider across the phases of classification and response criteria development*

| Phase | Methods | Bias mitigation strategies and reporting guidelines for systematic reviews |
|---|---|---|
| **Construct definition** | Systematic or scoping reviews of the literature | Use methodological standards for conducting systematic and scoping reviews.[63,64]<br>Use the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines for reporting systematic review[42] or the PRISMA Extension for Scoping Reviews (PRISMA-ScR).[43]<br>Use the PRISMA-COSMIN guidelines for reporting systematic reviews of outcome measurement instruments (OMIs).[44] |
| | Cross-sectional survey<br>Use qualitative methods (e.g., focus groups) to interview expert clinicians, people with the disease and/or additional stakeholders to identify themes | Purposely sample a range of participants with complementary skills and different perspectives. |
| **Item generation** | Systematic or scoping review of the literature for relevant items representing the construct, including review of previous criteria and related measurement instruments. | Use methodological standards for conducting systematic and scoping reviews.[63,64]<br>Use the PRISMA guidelines for systematic review[42] or the PRISMA-ScR.[43] |
| | Focus groups or survey of expert clinicians, people with the disease and/or additional stakeholders | Maintain a response rate of 80% or more.<br>Use of Dillman methods in survey studies[52] to increase the response rate. |
| | Items captured in existing cohorts or trials databases | Use more than 1 cohort.<br>Include data from different racial/ethnical backgrounds, sex, regions, disease spectrum, etc. |
| | Delphi methods[31,32] | Purposely sample a range of participants with complementary skills and different perspectives.<br>Use efforts to achieve response rate of 80% or more per Delphi round.[52] |
| **Item reduction** | Consensus methods:<br>Delphi methods[31,32]<br>Nominal group technique[54] | Purposely sample a range of participants with complementary skills and different perspectives.<br>Use efforts to achieve response rate of 80% or more per Delphi round.[52] |
| | Cohort data:<br>Consider item prevalence, sensitivity, specificity | Use more than 1 cohort.<br>Include data from different geographic regions. |
| **Criteria system derivation & refinement** | Calculation of item weights using methods such as multicriteria decision analysis or modelling methods | Use of real exemplar cases.<br>Use architectural terms to label cases.<br>Avoid alpha-numeric labeling of cases. |
| | Threshold identification | When cohort data is being used for threshold identification, and if physician judgement is the gold standard, have at least 2 of 3 independent investigators adjudicate and agree on cases and controls |
| | Data-based criteria refinement in a derivation cohort | Include a minimum of 100 cases and 100 controls<br>When cohort data is being used for criteria refinement, and if physician judgement is the gold standard, have at least 2 of 3 independent investigators adjudicate and agree on cases and controls. |
| **Criteria system validation** | Independent validation cohort to assess criteria performance and compare to older criteria/alternatives | Include a minimum of 100 cases and 100 controls<br>When cohort data is being used for criteria validation, and if physician judgement is the gold standard, have at least 2 of 3 independent investigators adjudicate and agree on cases and controls.<br>Investigators who contribute data for validation should be different than those who created the criteria system |
| | Randomized trial data (for response criteria) | Use data from trials that were blinded, had allocation concealment, and had small well-described withdrawals in each arm |

* To avoid the bias of circular reasoning, a steering committee may consider using this table to determine where specific investigators/stakeholders may be included. For example, reserve experts with data for the derivation and validation cohorts.

**Criteria system derivation and refinement.** After items are precisely defined so that they will be assessed in a standardized manner (increasing reliability of the criteria system),[53] consideration is given to the need for item weighting. Historically, items within a classification criteria system were considered equally weighted (eg, four of 11 criteria defined 1982 SLE

classification criteria[55]) or divided into minor and major criteria[56] which implies two different weights. Formal methods are now used to assign quantitative weights to individual items. However, a criteria system does not have to have weights. For example, the Boolean criteria for RA remission do not have additive weights. Data-driven or expert-based consensus methods could be used to derive item weights. Multicriteria decision analysis has been frequently used[5,17,18,24,35,57] as it is forward-thinking, allowing experts to derive weights for contemporary and emerging items. Data-driven methods rely on pre-existing data sets that may not contain the items, or have item definitions that are outdated or were not applied uniformly across centers. Data-driven methods using modeling techniques may be considered if the appropriate data exist.

Using multicriteria decision analysis, 20-30 exemplar cases ranging in their probability of having the disease or having responded to treatment are selected. The exemplar cases should be real cases sampled from the clinics of investigators who are not part of the expert panel and not experts contributing data for the derivation and validation cohorts. Inventing or artificial intelligence created cases should be avoided as fabricated cases may result in combinations of disease manifestations that are rarely observed. A standardized data collection form that includes items under consideration should be used to summarize these exemplar cases.[29]

An expert panel (usually a dozen) with relevant expertise and diverse geographic and sex representation is assembled. This panel must have had minimal involvement in the criteria development process (in both the work leading up to and occurring after this phase) due to the considerable influence they have on the final criteria system. For example, an expert who was part of a Delphi exercise of >100 participants could be included, as their contribution to the overall project is small. However, this expert must not contribute data for criteria derivation, refinement, validation nor adjudicate the reference standard, as this risks the bias of circular reasoning.

Independently, panel members rank-order the cases by probability of having the disease or having responded to treatment. The ranking of cases helps the experts clarify their ideas on characteristics that matter most to them (ie, by identifying key positive and key negative items, providing a basis for which items should have the highest or lowest weights). It will also help test the face validity of the new criteria system, by comparing the system to the original expert rankings. The panel attends a consensus meeting to conduct a discrete choice experiment. Using software (eg, 1000Minds) managed by a facilitator, experts assess two cases defined to be identical in every respect except for two items that are contrasted to assess their relative importance. The experts vote which case is most likely to have the disease or responded to treatment (Figure 3). If there is less than 70% agreement discussion ensues. Experts can explain the rationale behind their decision making. After the discussion has completed, the experts vote again until consensus is achieved. The voting preferences are captured using the software. The software calculates the part-weight utilities and are converted to a score. These preliminary weights may be further refined, or it may be determined that differential weighting is not required if items weights are very similar.

*Threshold setting.* The next step is to identify the total score needed to classify a person as having the disease. For example, the same exemplar cases ranked prior to the discrete choice experiment are scored with the preliminary item weights and arranged in order of their total score. The resulting rank order of cases by total score is compared to the rank order provided earlier by the panel members. The total score may be reviewed to assess whether the order makes sense and if revisions to individual item weights are needed.[4,24] The draft scoring system has good face validity if high scoring cases were ranked as high probability and low scoring cases were ranked low probability by the majority of the expert panel. A preliminary threshold for classification can be set where 60%–75% of the panel would classify a case as having the disease.[9,26] An alternative approach is asking

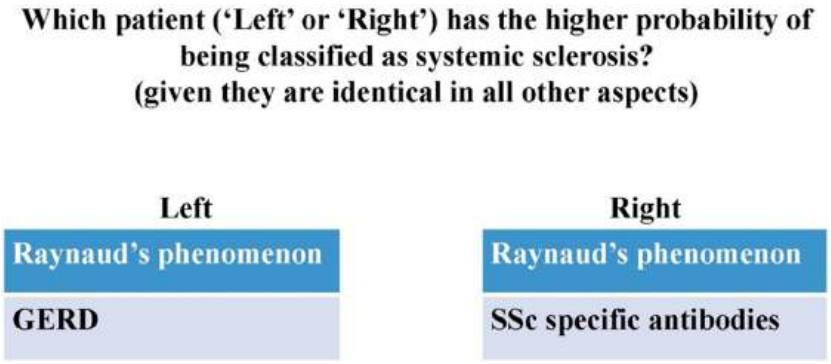


**Figure 3.** An illustrative example of a discrete choice experiment from development the ACR/EULAR classification criteria for systemic sclerosis. Experts assess two cases defined to be identical in every respect, except for two items that are contrasted, to assess their relative importance. The expert panel votes which case is most likely to have the disease (or responded to treatment is the setting of response criteria). ACR, American College of Rheumatology; GERD, gastroesophageal reflux disease; SSc, systemic sclerosis.

the expert panel to indicate whether they would enroll a case into a trial, or whether they would consider the case to have responded to a therapy. The inflection point in the global score at which most of the panel would enroll versus not enroll a case, or consider response versus nonresponse in a trial, is the preliminary threshold.

*Refinement.* A derivation cohort is used to refine the draft criteria system in. A derivation cohort of at least 100 cases (have the disease) and 100 controls (disease mimickers) is required so that confidence intervals for the estimated sensitivity and specificity of the criteria will be reasonably narrow.[2,3] Similarly, for response criteria, a minimum of 100 active treatment and 100 comparator cases is required.[2,3] Increasing the sample size beyond the recommended numbers will increase the precision around sensitivity and specificity so is recommended, when feasible. The data may come from observational data, trials, or new data collected for this purpose. The data may be collected after the item reduction phase so that it includes all the items, using the contemporary item definitions. This derivation cohort should reflect the full spectrum of clinical presentation, disease duration (early and established, as appropriate), and be multiethnic (collecting race/ethnicity data). Investigators who participated in the creation of the draft criteria system should be separate from those who contribute derivation cohort data, to prevent the bias of circular reasoning.[2,3] This will ensure that the items used by one group of clinicians to conceive and evaluate the presence of disease is not simply confirmed by the same group of clinicians providing their patients' data.[2,3] The impact of minor modifications, rounding weights or simplifying weights to whole numbers on the sensitivity and specificity of the criteria system can be evaluated iteratively. However, modifications should not distort the hierarchy of the weights (criteria with larger weights should still have larger weights). The final threshold is set by computing sensitivity and specificity of the draft criteria system (determined by gold standard or independent expert panel) in the derivation cohort, typically with the aim to optimize specificity to minimize false positives. Cases that were misclassified should be evaluated to determine the cause of misclassification and ascertain if this can be remedied without disrupting the hierarchy of the drafted system. Some refinements, such as merging of items, may necessitate re-doing some of the discrete choice experiments to derive updated weights.

**Criteria system validation.** The ACR requires validation of criteria in an independent data set with ≥100 cases and ≥100 controls.[2,3] The cohort should reflect the full spectrum of presentations, including early and established disease, and be multiethnic to allow testing by sex, disease duration and race/ethnicity.[58] The validation cohort must be separate from the derivation cohort, though they may be assembled simultaneously and randomly split. The validation cohort should not be used to revise the criteria; if revisions occur, it becomes a second derivation cohort, requiring another validation cohort. Comparisons with previous criteria are recommended.

For response criteria, validation should use randomized trial data. If unavailable, validation should occur in future trials with appropriate data collection, though provisional validation may be done using observational data.[21,37]

## Methodologic innovation

The ACR values methodologic innovation. The methods and bias mitigation strategies outlined summarize what has been done successfully. The use of innovative methods that improves the validity, reliability or responsiveness of the criteria systems are welcome.

## ACR Funding and Endorsement

In deciding to fund or endorse new classification or response criteria, the ACR requires use of rigorous methods aligned with this guidance. Supplementary Table 1. In addition, the ACR considers other factors as outlined below.

Proposals that fall into one or more of these categories will be favored during reviews for possible ACR funding support:

1. A disease with high prevalence (eg, common rheumatic diseases).
2. A less prevalent disease with a large impact on morbidity or mortality (eg, rare systemic autoimmune rheumatic diseases).
3. A disease for which criteria exist but are suboptimal or for which no criteria exist.
4. Projects that advance the criteria development field by using innovative methods.

Final ACR endorsement of a new criteria system, funded or not funded by the ACR, requires:

1. An improvement in criteria performance relative to older criteria, grounded by robust methods,
2. Incremental value by addressing new knowledge of the disease, including new commonly used laboratory or imaging tests,
3. Or, clearly addressing another unmet need.

For response criteria, "provisional" ACR endorsement may be given if there are no adequate trial data for external validation and criteria have undergone validation using cohort data.[59–62] The ACR encourages such provisional criteria undergo external validation as trial data becomes available and then be considered for full endorsement. The ACR understands that external validation may happen by the original investigator group or by others. The category of provisional endorsement does not apply to classification criteria, as external validation is expected.[59]

## AUTHOR CONTRIBUTIONS

All authors contributed to at least one of the following manuscript preparation roles: conceptualization AND/OR methodology, software, investigation, formal analysis, data curation, visualization, and validation AND drafting or reviewing/editing the final draft. As corresponding author, Dr Johnson confirms that all authors have provided the final approval of the version to be published and takes responsibility for the affirmations regarding article submission (eg, not under consideration by another journal), the integrity of the data presented, and the statements regarding compliance with institutional review board/Declaration of Helsinki requirements.

## REFERENCES


1. Johnson SR, Goek ON, Singh-Grewal D, et al. Classification criteria in rheumatic diseases: a review of methodologic properties. Arthritis Rheum 2007;57(7):1119–1133. doi:https://doi.org/10.1002/art.23018
2. Singh JA, Solomon DH, Dougados M, et al; Classification and Response Criteria Subcommittee of the Committee on Quality Measures, American College of Rheumatology. Development of classification and response criteria for rheumatic diseases. Arthritis Rheum 2006;55(3):348–352. doi:https://doi.org/10.1002/art.22003
3. Felson DT, Anderson JJ. Methodological and statistical approaches to criteria development in rheumatic diseases. Baillieres Clin Rheumatol 1995;9(2):253–266. doi:https://doi.org/10.1016/S0950-3579(05)80189-X
4. Neogi T, Jansen TL, Dalbeth N, et al. 2015 Gout classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis Rheumatol 2015;67(10):2557–2568. doi:https://doi.org/10.1002/art.39254
5. Shiboski CH, Shiboski SC, Seror R, et al; International Sjögren's Syndrome Criteria Working Group. 2016 American College of Rheumatology/European League Against Rheumatism classification criteria for primary Sjögren's syndrome: a consensus and data-driven methodology involving three international patient cohorts. Arthritis Rheumatol 2017;69(1):35–45. doi:https://doi.org/10.1002/art.39859
6. Ravelli A, Minoia F, Davì S, et al; Paediatric Rheumatology International Trials Organisation; Childhood Arthritis and Rheumatology Research Alliance; Pediatric Rheumatology Collaborative Study Group; Histiocyte Society. 2016 classification criteria for macrophage activation syndrome complicating systemic juvenile idiopathic arthritis: a European League Against Rheumatism/American College of Rheumatology/Paediatric Rheumatology International Trials Organisation collaborative initiative. Arthritis Rheumatol 2016;68(3):566–576. doi:https://doi.org/10.1002/art.39332
7. Rider LG, Aggarwal R, Pistorio A, et al; International Myositis Assessment and Clinical Studies Group and the Paediatric Rheumatology International Trials Organisation. 2016 American College of Rheumatology/European League Against Rheumatism criteria for minimal, moderate, and major clinical response in juvenile dermatomyositis: an International Myositis Assessment and Clinical Studies Group/Paediatric Rheumatology International Trials Organisation collaborative initiative. Arthritis Rheumatol 2017;69(5):911–923. doi:10.1002/art.40060
8. Lundberg IE, Tjärnlund A, Bottai M, et al; International Myositis Classification Criteria Project Consortium, the Euromyositis Register, and the Juvenile Dermatomyositis Cohort Biomarker Study and Repository (UK and Ireland). 2017 European League Against Rheumatism/American College of Rheumatology classification criteria for adult and juvenile idiopathic inflammatory myopathies and their major subgroups. Arthritis Rheumatol 2017;69(12):2271–2282. doi:https://doi.org/10.1002/art.40320
9. Aringer M, Costenbader K, Daikh D, et al. 2019 European League Against Rheumatism/American College of Rheumatology classification criteria for systemic lupus erythematosus. Arthritis Rheumatol 2019;71(9):1400–1412. doi:https://doi.org/10.1002/art.40930
10. Wallace ZS, Naden RP, Chari S, et al; American College of Rheumatology/European League Against Rheumatism IgG4-Related Disease Classification Criteria Working Group. The 2019 American College of Rheumatology/European League Against Rheumatism classification criteria for IgG4-related disease. Arthritis Rheumatol 2020;72(1):7–19. doi:https://doi.org/10.1002/art.41120
11. Trincianti C, Van Dijkhuizen EHP, Alongi A, et al; Paediatric Rheumatology International Trials Organisation. Definition and validation of the American College of Rheumatology 2021 Juvenile Arthritis Disease Activity Score cutoffs for disease activity states in juvenile idiopathic arthritis. Arthritis Rheumatol 2021;73(11):1966–1975. doi:https://doi.org/10.1002/art.41879
12. Grayson PC, Ponte C, Suppiah R, et al; DCVAS Study Group. 2022 American College of Rheumatology/European Alliance of Associations for Rheumatology classification criteria for eosinophilic granulomatosis with polyangiitis. Arthritis Rheumatol 2022;74(3):386–392. doi:https://doi.org/10.1002/art.41982
13. Grayson PC, Ponte C, Suppiah R, et al; DCVAS Study Group. 2022 American College of Rheumatology/EULAR classification criteria for Takayasu arteritis. Arthritis Rheumatol 2022;74(12):1872–1880. doi:https://doi.org/10.1002/art.42324
14. Ponte C, Grayson PC, Robson JC, et al; DCVAS Study Group. 2022 American College of Rheumatology/EULAR classification criteria for giant cell arteritis. Arthritis Rheumatol 2022;74(12):1881–1889. doi:https://doi.org/10.1002/art.42325
15. Robson JC, Grayson PC, Ponte C, et al; DCVAS Study Group. 2022 American College of Rheumatology/European Alliance of Associations for Rheumatology classification criteria for granulomatosis with polyangiitis. Arthritis Rheumatol 2022;74(3):393–399. doi:https://doi.org/10.1002/art.41986
16. Suppiah R, Robson JC, Grayson PC, et al; DCVAS Study Group. 2022 American College of Rheumatology/European Alliance of Associations for Rheumatology classification criteria for microscopic polyangiitis. Arthritis Rheumatol 2022;74(3):400–406. doi:https://doi.org/10.1002/art.41983
17. Abhishek A, Tedeschi SK, Pascart T, et al. The 2023 ACR/EULAR classification criteria for calcium pyrophosphate deposition disease. Arthritis Rheumatol 2023;75(10):1703–1713. doi:https://doi.org/10.1002/art.42619
18. Barbhaiya M, Zuily S, Naden R, et al; ACR/EULAR APS Classification Criteria Collaborators. The 2023 ACR/EULAR antiphospholipid syndrome classification criteria. Arthritis Rheumatol 2023;75(10):1687–1702. doi:https://doi.org/10.1002/art.42624
19. Aggarwal R, Ringold S, Khanna D, et al. Distinctions between diagnostic and classification criteria? Arthritis Care Res (Hoboken) 2015;67(7):891–897. doi:https://doi.org/10.1002/acr.22583
20. Johnson SR, Touma Z. Principles of Clinical Outcome Assessment. In: Hochberg MC, Gravallese EM, Smolen JS, van der Heijde D, Weinblatt ME, Weisman MH, eds. Rheumatology. 8th ed. Elsevier; 2023:198–202.
21. Studenic P, Aletaha D, de Wit M, et al. American College of Rheumatology/EULAR remission criteria for rheumatoid arthritis: 2022 revision. Arthritis Rheumatol 2023;75(1):15–22. doi:https://doi.org/10.1002/art.42347
22. Felson DT, Anderson JJ, Boers M, et al. American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum 1995;38(6):727–735. doi:https://doi.org/10.1002/art.1780380602
23. Aletaha D, Neogi T, Silman AJ, et al. 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European

League Against Rheumatism collaborative initiative. Arthritis Rheum 2010;62(9):2569–2581. doi:https://doi.org/10.1002/art.27584.

24. Neogi T, Aletaha D, Silman AJ, et al; American College of Rheumatology; European League Against Rheumatism. The 2010 American College of Rheumatology/European League Against Rheumatism classification criteria for rheumatoid arthritis: phase 2 methodological report. Arthritis Rheum 2010;62(9):2582–2591. doi:https://doi.org/10.1002/art.27580
25. Kirshner B, Guyatt G. A methodological framework for assessing health indices. J Chronic Dis 1985;38(1):27–36. doi:https://doi.org/10.1016/0021-9681(85)90005-0
26. van den Hoogen F, Khanna D, Fransen J, et al. 2013 classification criteria for systemic sclerosis: an American College of Rheumatology/European League against Rheumatism collaborative initiative. Arthritis Rheum 2013;65(11):2737–2747. doi:https://doi.org/10.1002/art.38098
27. Feinstein AR. Clinimetrics. Yale University Press; 1987.
28. Streiner DL, Norman GR. Health Measurement Scales. A Practical Guide to Their Development and Use. Oxford University Press; 2003.
29. Fraenkel L, Miller AS, Clayton K, et al. When patients write the guidelines: patient panel recommendations for the treatment of rheumatoid arthritis. Arthritis Care Res (Hoboken) 2016;68(1):26–35. doi:https://doi.org/10.1002/acr.22758
30. Mirza RD, Bolster MB, Johnson SR, et al. Assessing patient values and preferences to inform the 2023 American College of Rheumatology/American College of Chest Physicians interstitial lung disease guidelines. Arthritis Care Res (Hoboken) 2024;76(8):1083–1089. doi:https://doi.org/10.1002/acr.25346
31. Brown BB. Delphi Process: A Methodology Used for the Elicitation of Opinions of Experts. RAND Corporation; 1968.
32. Beiderbeck D, Frevel N, von der Gracht HA, et al. Preparing, conducting, and analyzing Delphi surveys: cross-disciplinary practices, new directions, and advancements. MethodsX 2021;8:101401. doi:https://doi.org/10.1016/j.mex.2021.101401
33. Johnson SR, Khanna D, Daikh D, et al. Use of consensus methodology to determine candidate items for systemic lupus erythematosus classification criteria. J Rheumatol 2019;46(7):721–726. doi:https://doi.org/10.3899/jrheum.180478
34. Nair R, Aggarwal R, Khanna D. Methods of formal consensus in classification/diagnostic criteria and guideline development. Semin Arthritis Rheum 2011;41(2):95–105. doi:https://doi.org/10.1016/j.semarthrit.2010.12.001
35. Johnson SR, Naden RP, Fransen J, et al. Multicriteria decision analysis methods with 1000Minds for developing systemic sclerosis classification criteria. J Clin Epidemiol 2014;67(6):706–714. doi:https://doi.org/10.1016/j.jclinepi.2013.12.009
36. Touma Z, Cervera R, Brinks R, et al; American College of Rheumatology/European League Against Rheumatism Group. Associations between classification criteria items in systemic lupus erythematosus. Arthritis Care Res (Hoboken) 2020;72(12):1820–1826. doi:https://doi.org/10.1002/acr.24078
37. Felson DT, Smolen JS, Wells G, et al; American College of Rheumatology; European League Against Rheumatism. American College of Rheumatology/European League Against Rheumatism provisional definition of remission in rheumatoid arthritis for clinical trials. Arthritis Rheum 2011;63(3):573–586. doi:https://doi.org/10.1002/art.30129
38. Gladman D, Ginzler E, Goldsmith C, et al. The development and initial validation of the systemic lupus international collaborating clinics/American college of rheumatology damage index for systemic lupus erythematosus. Arthritis Rheum 1996;39(3):363–369.
39. Barber MRW, Johnson SR, Gladman DD, et al. Evolving concepts in systemic lupus erythematosus damage assessment. Nat Rev Rheumatol 2021;17(6):307–308. doi:https://doi.org/10.1038/s41584-021-00611-4
40. Johnson SR, Gladman DD, Brunner HI, et al. Evaluating the construct of damage in systemic lupus erythematosus. Arthritis Care Res (Hoboken) 2023;75(5):998–1006. doi:https://doi.org/10.1002/acr.24849
41. Johnson SR, Soowamber ML, Fransen J, et al. There is a need for new systemic sclerosis subset criteria. A content analytic approach. Scand J Rheumatol 2018;47(1):62–70. doi:https://doi.org/10.1080/03009742.2017.1299793
42. Moher D, Liberati A, Tetzlaff J, et al; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg 2010;8(5):336–341. doi:https://doi.org/10.1016/j.ijsu.2010.02.007
43. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and Explanation. Ann Intern Med 2018;169(7):467–473. doi:https://doi.org/10.7326/M18-0850
44. Elsman EBM, Mokkink LB, Terwee CB, et al. Guideline for reporting systematic reviews of outcome measurement instruments (OMIs): PRISMA-COSMIN for OMIs 2024. J Clin Epidemiol 2024;173:111422. doi:https://doi.org/10.1016/j.jclinepi.2024.111422
45. Nevskaya T, Pope JE, Turk MA, et al. Systematic analysis of the literature in search of defining systemic sclerosis subsets. J Rheumatol 2021;48(11):1698–1717. doi:https://doi.org/10.3899/jrheum.201594
46. Leuchten N, Hoyer A, Brinks R, et al; Systemic Lupus Erythematosus Classification Criteria Steering Committee. Performance of antinuclear antibodies for classifying systemic lupus erythematosus: a systematic literature review and meta-regression of diagnostic data. Arthritis Care Res (Hoboken) 2018;70(3):428–438. doi:https://doi.org/10.1002/acr.23292
47. Johnson SR, Fransen J, Khanna D, et al. Validation of potential classification criteria for systemic sclerosis. Arthritis Care Res (Hoboken) 2012;64(3):358–367. doi:https://doi.org/10.1002/acr.20684
48. Mosca M, Costenbader KH, Johnson SR, et al. Brief Report: how do patients with newly diagnosed systemic lupus erythematosus present? A multicenter cohort of early systemic lupus erythematosus to inform the development of new classification criteria. Arthritis Rheumatol 2019;71(1):91–98. doi:https://doi.org/10.1002/art.40674
49. Leuchten N, Milke B, Winkler-Rohlfing B, et al; on behalf of the SLE Classification Criteria Steering Committee. Early symptoms of systemic lupus erythematosus (SLE) recalled by 339 SLE patients. Lupus 2018;27(9):1431–1436. doi:https://doi.org/10.1177/0961203318776093
50. Schmajuk G, Hoyer BF, Aringer M, et al; SLE classification criteria steering committee and the international SLE expert panel of the initiative. Multicenter Delphi exercise to identify important key items for classifying systemic lupus erythematosus. Arthritis Care Res (Hoboken) 2018;70(10):1488–1494. doi:https://doi.org/10.1002/acr.23503
51. Fransen J, Johnson SR, van den Hoogen F, et al. Items for developing revised classification criteria in systemic sclerosis: results of a consensus exercise. Arthritis Care Res (Hoboken) 2012;64(3):351–357. doi:https://doi.org/10.1002/acr.20679
52. Dillman DS, Smyth JD, Christian LM. Internet, Mail, and Mixed-Mode Surveys. The Tailored Design Method. John Wiley & Sons, Inc; 2009.
53. Tedeschi SK, Johnson SR, Boumpas D, et al. Developing and refining new candidate criteria for systemic lupus erythematosus classification: an international collaboration. Arthritis Care Res (Hoboken) 2018;70(4):571–581. doi:https://doi.org/10.1002/acr.23317
54. Delbecq AL, Van de Ven AH. A group process model from problem identification and program planning. J Appl Behav Sci 1971;7(4):466–492. doi:https://doi.org/10.1177/002188637100700404
55. Tan EM, Cohen AS, Fries JF, et al. The 1982 revised criteria for the classification of systemic lupus erythematosus. Arthritis Rheum 1982;25(11):1271–1277. doi:https://doi.org/10.1002/art.1780251101
56. Masi AT; Subcommittee for Scleroderma Criteria of the American Rheumatism Association Diagnostic and Therapeutic Criteria

Committee. Preliminary criteria for the classification of systemic sclerosis (scleroderma). Arthritis Rheum 1980;23(5):581–590. doi: https://doi.org/10.1002/art.1780230510

57. Tedeschi SK, Johnson SR, Boumpas DT, et al. Multicriteria decision analysis process to develop new classification criteria for systemic lupus erythematosus. Ann Rheum Dis 2019;78(5):634–640. doi: https://doi.org/10.1136/annrheumdis-2018-214685
58. Johnson SR, Brinks R, Costenbader KH, et al. Performance of the 2019 EULAR/ACR classification criteria for systemic lupus erythematosus in early disease, across sexes and ethnicities. Ann Rheum Dis 2020;79(10): 1333–1339. doi:https://doi.org/10.1136/annrheumdis-2020-217162
59. Johnson SR, Grayson PC. Use of "provisional" designation for response criteria. Arthritis Care Res (Hoboken) 2018;70(6):811–812. doi:https://doi.org/10.1002/acr.23555
60. Khanna D, Berrocal VJ, Giannini EH, et al. The American College of Rheumatology provisional composite response index for clinical trials in early diffuse cutaneous systemic sclerosis. Arthritis Rheumatol 2016;68(2):299–311. doi:https://doi.org/10.1002/art.39501
61. Brunner HI, Holland M, Beresford MW, et al; PRCSG and PRINTO Investigators. American College of Rheumatology provisional criteria for global flares in childhood-onset systemic lupus erythematosus. Arthritis Care Res (Hoboken) 2018;70(6):813–822. doi:https://doi.org/10.1002/acr.23557
62. Brunner HI, Holland MJ, Beresford MW, et al; Paediatric Rheumatology International Trial Organisation and Pediatric Rheumatology Collaborative Study Group. American College of Rheumatology provisional criteria for clinically relevant improvement in children and adolescents with childhood-onset systemic lupus erythematosus. Arthritis Care Res (Hoboken) 2019;71(5):579–590. doi:https://doi.org/10.1002/acr.23834
63. Pollock D, Evans C, Menghao Jia R, et al. "How-to": scoping review? J Clin Epidemiol 2024;176:111572. doi:https://doi.org/10.1016/j.jclinepi.2024.111572
64. Higgins JPT, Thomas J, Chandler J, et al, eds. Cochrane Handbook for Systematic Reviews of Interventions. Version 6.5. Cochrane; 2024.