QUALITY OF CARE COMMITTEE

Quality of Care Committee Reviewer (1):
Vote: Not certain that this should be approved with such low performance characteristics
Comments: The cardinal issue is whether these criteria have adequate performance characteristics to be considered "valid" criteria.

As they state in their abstract, the main criteria have only low-moderate sensitivity and specificity for PMR; it may be that no criteria can do better for PMR, but this seems to be an important issue.

"A score ≥4 had 69% sensitivity and 77% specificity for discriminating all comparison subjects from PMR."

Quality of Care Committee Reviewer (2):
Vote: Approve

Quality of Care Committee Reviewer (3):
Vote: Approve

Quality of Care Committee Reviewer (4):
Vote: Provisional Approval (full approval pending validation in an independent sample).

Quality of Care Committee Reviewer (5):
Vote: Approve
Comments: Very thorough work in a difficult condition. Authors are to be commended.

A few concerns:

Abstract conclusion is not entirely accurate...classification of PMR does not occur "in the absence of periph synovitis or positive RA serologies"--the algorithms simply make the classification of PMR less or more likely when considering individual criterion. This statement is repeated in the end of the discussion and should be correct there too.

If the purported problem that these criteria aim to address is "The lack of standardized classification criteria has been a major factor hampering development of rational therapeutic approaches in management of PMR[12, 16, 17]. These deficiencies have contributed to difficulties in evaluating patients in clinical studies." then the goal should be to define a cohort that is as uniform as possible, with little variation or question that the subjects do, indeed, have PMR (even if it is only a restricted

"sub-type" of PMR). Once trials demonstrate effective treatment/clearer understanding of the clinical characteristics for that homogenous group, further studies can be performed to assess whether clinical findings and treatment responses extend to other "sub-types" of PMR, or whether those conditions are distinct entities. But I think it is critical to avoid a heterogeneous population in this circumstance. Thus, sensitivity should be less emphasized compared to specificity, as the authors mention is their response to the Criterion Committee review--does it make sense to set the cut off higher for each algorithm and sacrifice sensitivity? Inclusion of the full ROC data is VERY helpful.

No formal analysis comparing ultrasound-based algorithm to the non-U/S algorithm was performed. While the specificity in the U/S algorithm is numerically higher, I am not certain whether the incremental benefit of the U/S algorithms is statistically significant. The following statement in the discussion (1st paragraph) is not correct: Ultrasound findings of bilateral shoulder abnormalities...significantly improve both the sensitivity and specificity of the clinical criteria.
The sensitivity of the ultrasound algorithm at the recommended cut off of 5 points (65.8) is numerically lower than the non-U/S algorithm at the recommended cut off of 4 points (68). More accurate to say that U/S-based algorithm with a 5 point cut off **may** improve specificity.

Recognizing that a single reference laboratory was not used, were there even vague rules about what constitutes an abnormal CRP/ESR value? Audience will be lost without some guidance. It might be useful to at least include a statement about what reference laboratories used as cut-offs--at least for CRP and ESR, whether age was considered while interpreting ESR, etc. Assuming all labs were certified, the absolute CRP concentration and ESR score should have been roughly comparable across labs. I'm simply trying to determine if most labs used CRP cut offs of 3mg/L, 5mg/L or 10mg/L. If no guidelines were given, then simply saying so would be helpful.

The C in Table 2 and the Table Appendix 2 (should this be Appendix 2 Table?), should be defined in the text under the table or labeled "C statistic", I think

I do think the authors have done about as good a job as they can. Kudos to all involved.

Quality of Care Committee Reviewer (6):
Vote: Approve
Comments: A fairly rigorous study and certainly an addition to the literature.
1. In table 3, the authors should insert values for C-statistic for the different models shown.

2. In the methods, it would be very helpful to provide the clinical audience some interpretation of the c-statistic. There are published conventions suggesting that a C-stat >= 0.8 is probably of some value where values lower than this are probably only marginally helpful in clinical decision making.

Quality of Care Committee Reviewer (7):
Vote: Approval **based on minor revisions** (see comments below)
Comment: For ESR and/or CRP, the authors state that institution specific values were used to define an elevated level. It would be helpful to know the range of these thresholds and the methods used to measure ESR and CRP. Also, it seems like the authors characterized a person as positive when either ESR or CRP was elevated (if both were obtained). If true, perhaps they should say ESR or CRP rather than 'and/or'.

The paper also needs to be edited by a native English speaker.

Quality of Care Committee Reviewer (8):
Vote: Approve **as provisional criteria**
Comment: I vote for provisional acceptance since the criteria have not been validated in an external independent dataset.

I don't feel that the authors adequately addressed every comment, but overall I think the clarity of the manuscript was improved.

CRITERIA SUBCOMMITTEE

Criteria Subcommittee Reviewer (1):
Vote: Approval **based on minor revisions** (see comments)
Comment: For ESR and/or CRP, the authors state that institution specific values were used to define an elevated level. It would be helpful to know the range of these thresholds and the methods used to measure ESR and CRP. Also, it seems like the authors characterized a person as positive when either ESR or CRP was elevated (if both were obtained). If true, perhaps they should say ESR or CRP rather than 'and/or'.

The paper also needs to be edited by a native English speaker.

Criteria Subcommittee Reviewer (2):
Vote: Approve **as provisional criteria**
Comment: I think they have done a very good job of responding to our points and that these criteria are definitely better than anything that is out there thus far. I vote that we endorse them as the provisional ACR criteria for classification of PMR, as they will need independent validation.

Criteria Subcommittee Reviewer (3):
Vote: Approve **as provisional criteria**

Comment: The group has answered the comments appropriately.

Criteria Subcommittee Reviewer (4):
Vote: Approve **as provisional criteria**
Comment: I was underwhelmed by their accuracy which may be due to lack of sufficient item generation or overzealous item reduction.

Nonetheless, this seems to be the best there is at the present time and there is a clear need to do the upcoming RCTs at least with some kind of standard diagnostic criteria.

Also, I do not see that there is a prospective validation (beyond the development phase) provided as far as I can see. Rather the prospective dataset is used to develop the scoring system. Although the development dataset size is acceptable, if a separate validation dataset was now considered a prerequisite for [full] endorsement, then [full] endorsement should not be provided.

Criteria Subcommittee Reviewer (5):
Vote: Approve

Criteria Subcommittee Reviewer (6):
Vote: Approve

Criteria Subcommittee Reviewer (7):
Vote: Approve **as provisional criteria**
Comment: For criteria that will be used to select subjects for clinical trials of potential therapies, one would prefer highly specific criteria that exclude patients who do not have the condition being studied. This may require accepting lower sensitivity, i.e., excluding more patients who will have PMR from the clinical trial, but will strengthen the trial design to require fewer subjects to demonstrate benefit for the subjects who actually have PMR. Perhaps the authors could include a small table showing the effects of requiring scores of 4, 5 and 6 (without ultrasound) or 5, 6, 7 and 8 (with ultrasound) on the specificity and sensitivity of the proposed criteria. Then investigators and statisticians could use this information when designing clinical trials, to balance recruitment issues (sensitivity) versus analysis issues (specificity).

Criteria Subcommittee Reviewer (9):
Vote: **Disapproval**
Comment: Overall I think the paper is remarkably improved and the data is elegantly presented. The authors did a great job of responding to the comments. In the end, I'm still very concerned about the selection of only patients with bilateral shoulder involvement. Moreover, the resulting criteria still have high misclassification rates. I understand that the aim is for specificity given the proposed use in clinical trials but the low sensitivity is concerning. Are these criteria really selecting for a particular subpopulation of PMR, albeit a large subpopulation? I realize there is a trade of but wonder if we could do better.

Criteria Subcommittee Reviewer (10):

Vote: Approve **as provisional criteria**
Comment: It seems to me the authors addressed reviewers' comments in a reasonable manner. It also seems to me from reading the critiques that there did seem to be some concern from the group with the approach even though the authors defend it. Given that they are looking for endorsement, perhaps a provisional approval would be warranted. It seems like replication in another cohort would be appropriate.

Criteria Subcommittee Reviewer (11):
Vote: Approve **as provisional criteria**