APPENDIX 1: SUPPLEMENTARY TABLES

Table S1. Reference ranges for laboratory tests according to center

Center	ESR (mm/hr)	CRP (mg/L)	RF (IU/ml)	ACPA (U/ml)
1	< 25	< 5	<40	< 20
2	2-37 in 1st hr	< 5	<30	0-10
4	<10	< 5	<40	<20.
6	< 8 in 1st hr, <18 in 2nd hr	< 5	< 14	< 10
8	0-9	< 5	0-13	0-6
8	0-9	< 8	0-14	0-10
13	FEMALE 2-20 MALE 2-15	<10	<16	< 7
14	Male: Yrs divided by 2 Female: yrs+10 divided by 2	< 5	1-20	<25l
18	1-20	<10	< 25	< 25
20	< 15	< 5	<50	<50
21	< 28	< 5	<20	<26
22	< 30	< 5	<20	<10
23	<37	<10	16	< 8
24	<37	<10	16	< 8
25	1-15	<10	<40	<7
30	<37	<10	16	< 8
31	0-29	< 8	<10	<15.6
32	Male: 0-22 Female: 0-29	≤8	<10	<15.6
33	0 - 29	< 8	<14	<20
34	<20	<8	<14	<16
40	<20 women < 14 men	< 5	1-20	<25l
41	women – (age+10)/2 men – age/2	≤ 8	≤ 35	< 7 – negative > 10 – positive

Table S2. Diagnoses of the 169 Comparison Subjects

New-onset rheumatoid arthritis (RA)	49 (29%)
New-onset other sero-negative arthritis *(e.g. spondyloarthritis, psoriatic arthritis)*	20 (12%)
New-onset connective tissue diseases and vasculitis *(e.g. ANCA-associated vasculitis, systemic lupus erythematosus, inflammatory myopathy)*	9 (5%)
Shoulder conditions *(e.g. bilateral rotator cuff syndrome and/or adhesive capsulitis, rotator cuff tear, glenohumeral osteoarthritis)*	52 (31%)
Chronic pain	26 (15%)
Endocrinopathy	2 (1%)
Neurological disorder	2 (1%)
Previously undiagnosed malignancy	4 (2%)
Infection	5 (3%)

Table S3. Factor Analysis: Factor Loadings for Individual Criteria

Criterion	Factor 1	Factor 2	Factor 3	Factor 4
Duration of symptoms >=2 weeks				-0.207
Shoulder tenderness			0.162	**0.578**
Pain or limited hip range of motion	**0.825**		0.142	
Bilateral pelvic girdle (hip) aching	**0.778**	0.173		
Hip tenderness	**0.641**		0.170	0.349
Neck aching		0.127		0.265
Morning stiffness > 45 minutes duration		0.112	**0.916**	
Recent weight loss of >2 kg	0.104	0.243	0.221	-0.134
Carpal Tunnel		0.294	0.149	
Peripheral synovitis		**0.677**		
Other joint pain		**0.757**		0.132
Abnormal RF and/or ACPA	-0.265		0.141	
MHAQ	0.244	0.409	0.351	

Abbreviations: RF = rheumatoid factor; ACPA= anti-citrullinated protein antibody; MHAQ= modified health assessment questionnaire

*factor loadings >0.5 are shown in bold. Factor loadings between -0.1 and 0.1 were not included in the table.

Table S4. Receiver Operating Characteristic (ROC) analysis for both scores

Clinical score w/o ultrasound			Score with ultrasound		
Score	Sensitivity	Specificity	Score	Sensitivity	Specificity
0	92.8	41.4	0	92.5	42.2
1	92.8	43.8	1	92.5	42.9
2	92.8	46.8	2	92.5	44.8
3	87.2	61.5	3	91.7	51.3
4	68.0	77.5	4	82.5	70.1
5	47.2	88.8	5	65.8	80.5
6	19.2	98.2	6	36.7	92.2
			7	20.0	98.0
			8	8.3	100

Table S5. Factor Analysis: Factor Loadings for Ultrasound Criteria

Criterion	Factor 1	Factor 2	Factor 3	Factor 4
at least 1 shoulder with subdeltoid bursitis, biceps tenosynovitis and/or glenohumeral synovitis(either posterior or axillary)	0.121	0.236	**0.959**	
Both shoulders with subdeltoid bursitis, biceps tenosynovitis and/or glenohumeral synovitis(either posterior or axillary)		**0.803**	0.351	0.141
at least 1 shoulder with subdeltoid bursitis, biceps tenosynovitis and/or glenohumeral synovitis(either posterior or axillary)	0.112	0.373	**0.775**	
Both shoulders with subdeltoid bursitis, tenosynovitis and/or glenohumeral synovitis(either posterior or axillary)		**0.960**	0.245	
at least 1 hip with synovitis and/or trochanteric bursitis	**0.906**			0.286
Both hips with synovitis and/or trochanteric bursitis	0.492		0.104	**0.720**
At least 1 shoulder with subdeltoid bursitis and/or biceps tenosynovitis and/or glenohumeral synovitis(either posterior or axillary) AND at least 1 hip with synovitis and/or trochanteric bursitis	**0.933**		0.167	0.301
US findings as detailed above in BOTH shoulder and BOTH hips	0.292	0.209		**0.930**

*factor loadings >0.5 are shown in bold

APPENDIX 2: Blinded re-evaluation of selected PMR patients and comparison subjects

While the prospective study was ongoing, we performed a blinded re-evaluation of a subset of PMR and comparison subjects included in the prospective study. Clearly, the development of classification criteria is inherently exposed to the risk of circularity of reasoning, since the same rheumatologists who judge whether the patient has the disease or not also develop the list of candidate criteria items. Outcome prediction at follow-up can obviate, at least to some extent, the problem of circularity. Therefore, in order to ensure the quality of the data and create a standard for assessment of patients included in the study, we performed a blinded re-evaluation of a subset of patients. This evaluation was intended to ensure uniformity of decision to classify PMR, to constitute a reproducibility assessment of the newly developed criteria and to estimate the reliability of classification based on assessment at the first visit. Every investigator was sent a set of data on 30 patients (10 PMR, 20 comparison subjects). For each disease feature the investigator was asked (on a five point scale) to circle a weight to indicate how much this disease feature influenced their clinical judgment or their decision whether the patient has or does not have PMR (1=strongly influences diagnosis of PMR to 5=strongly influences the diagnosis was not PMR). Each reviewer was then asked to re-evaluate the diagnosis (PMR or not PMR) and indicate on a 5-point scale their degree of confidence in the diagnosis. The reviewer was then asked whether they would treat such a case with corticosteroid, and whether they would enter such a case in a clinical trial for PMR.

To assess the diagnostic accuracy of each candidate criteria, the mean rating across all raters was taken. This composite score was then used to determine the areas under the receiver operating characteristic (ROC) curve (AUC), denoted as the c-statistic. Subjects were categorised into 3 groups based on raters' misclassification rates. Group 1: greater than 50% misclassified; Group 2: 20–50% misclassified, Group 3: less than 20% misclassified.

The assessment of multi-rater discrimination of PMR (10 patients) from comparison subjects (20 subjects) was undertaken by 23 investigators. The results are shown in Table below. Misclassification proportion was high in 10 patients. In Group 1 (n=3, 1 PMR, 2 comparison subjects), the factors that contributed to the misclassification were normal CRP and/or ESR, poor or ill- sustained corticosteroid response, and RF positivity without peripheral synovitis. In Group 2 (n=7; 4 PMR, 3 comparison subjects), misclassification was related to persistent synovitis, lack of complete/sustained corticosteroid response, RF or CCP positivity and low baseline CRP and/or ESR.

The c-statistic suggested that gender, duration of symptoms, systemic symptoms such as weight loss, neck pain, limitation of movement and serum electrophoresis were unhelpful to

the blinded rater, in discriminating PMR from comparison subjects (c-statistic < 0.8 in all). Bilateral hip pain, morning stiffness, CRP and ESR levels (pre- and especially post-corticosteroid), and corticosteroid response were good discriminators of PMR from comparison subjects (c-statistic > 0.8 in all, Table).

Table Appendix 2: Blinded Multi-rater Evaluation of Diagnosis and Candidate Criteria

Candidate Criteria	PMR subjects Mean (SD)	Comparison subjects Mean (SD)	P value	C statistic (95% CI)
Bilateral pelvic girdle aching	1.8 (1.1)	3.3 (1.1)	0.01	0.80 (0.54, 0.92)
Morning stiffness >45min	1.7 (1.0)	3.0 (1.4)	<0.01	0.87 (0.63, 0.96)
Abnormal CRP at baseline	1.8 (0.9)	3.2 (1.4)	<0.01	0.81 (0.58, 0.92)
Abnormal ESR at 26 weeks	2.3 (1.0)	3.2 (0.8)	<0.01	0.89 (0.64, 0.97)
Abnormal CRP at 26 weeks	2.2 (0.9)	3.1 (1.1)	<0.01	0.85 (0.62, 0.94)
Rapid steroid response	2.6 (1.5)	4.6 (0.6)	<0.01	0.99 (0.90, 1.00)
Complete steroid response	2.2 (1.4)	4.5 (0.7)	<0.01	0.98 (0.84, 1.00)
Sustained steroid response	2.6 (1.4)	4.3 (0.7)	<0.01	0.99 (0.90, 1.00)

PMR = polymyalgia rheumatica; ESR = erythrocyte sedimentation rate; CRP = C-reactive protein; CI = confidence interval

Furthermore, the clinical scoring algorithm was evaluated in these 30 subjects. The algorithm correctly classified 8 of the 10 PMR subjects and 16 of the 20 comparison subjects. The 6 subjects who were incorrectly classified by our algorithm were also frequently misclassified by the expert raters. The median percentage of raters who incorrectly classified these patients was 33% compared to 4% for the other 24 subjects; p=0.014). This indicated our criteria perform well in distinguishing PMR from comparison subjects among patients that are agreed upon by experts. In addition, the subjects misclassified by our algorithm are more likely to be those which experts could not agree upon.

The re-evaluation exercise showed that most candidate criteria items performed well in discriminating PMR from comparison subjects. However, a significant proportion (a third of the sample) of PMR/comparison subjects was difficult to classify, as noted by the high percentage of raters who misclassified these subjects. The high c-statistic levels associated with the corticosteroid response and post- treatment CRP and ESR suggested that the uncertainty originated from the pivotal role of the corticosteroid in the investigator assessment, in deciding whether a patient does or does not have PMR. Questions such as whether PMR may not always adequately respond to corticosteroid and whether polymyalgic RF positive disease without peripheral synovitis can occur clearly require further investigation.