The Journal of Rheumatology 2021;48:xxxx
doi:10.3899/jrheum.200248
First Release December 15 2020

The Journal of Rheumatology

Outcomes Measured in Polymyalgia Rheumatica and Measurement Properties of Instruments Considered for the OMERACT Core Outcome Set: A Systematic Review

Helen Twohig[1], Claire Owen[2], Sara Muller[1], Christian D. Mallen[3], Caroline Mitchell[4], Samantha Hider[3], Catherine Hill[5], Beverley Shea[6], and Sarah L. Mackie[7]

ABSTRACT. ***Objective.*** To systematically identify the outcome measures and instruments used in clinical studies of polymyalgia rheumatica (PMR) and to evaluate evidence about their measurement properties.

Methods. Searches based on the MeSH term "polymyalgia rheumatica" were carried out in 5 databases. Two researchers were involved in screening, data extraction, and risk of bias assessment. Once outcomes and instruments used were identified and categorized, key instruments were selected for further review through a consensus process. Studies on measurement properties of these instruments were appraised against the COSMIN-OMERACT (COnsensus-based Standards for the selection of health Measurement Instruments–Outcome Measures in Rheumatology) checklist to determine the extent of evidence supporting their use in PMR.

Results. Forty-six studies were included. In decreasing order of frequency, the most common outcomes (and instruments) used were markers of systemic inflammation [erythrocyte sedimentation rate (ESR), C-reactive protein (CRP)], pain [visual analog scale (VAS)], stiffness (duration in minutes), and physical function (elevation of upper limbs). Instruments selected for further evaluation were ESR, CRP, pain VAS, morning stiffness duration, and the Health Assessment Questionnaire. Five studies evaluated measurement properties of these instruments, but none met all of the COSMIN-OMERACT checklist criteria.

Conclusion. Measurement of outcomes in studies of PMR lacks consistency. The critical patient-centered domain of physical function is poorly assessed. None of the candidate instruments considered for inclusion in the core outcome set had high-quality evidence, derived from populations with PMR, on their full range of measurement properties. Further studies are needed to determine whether these instruments are suitable for inclusion in a core outcome measurement set for PMR.

Key Indexing Terms: OMERACT, outcome measures, polymyalgia rheumatica, systematic review

Polymyalgia rheumatica (PMR) is the most common inflammatory rheumatic condition of older people[1] and is characterized by proximal pain and stiffness, raised inflammatory markers, and a therapeutic response to glucocorticoids[2]. A recent UK study using the Clinical Practice Research Datalink found an annual incidence of 96 per 100,000 people aged over 40 years, with incidence rising markedly with increasing age[3].

Although it is common, PMR remains underresearched, and

This work was supported by a Wellcome Trust PhD Programme for Primary Care Clinicians [203921/Z/16/Z], which supports Helen Twohig. CDM is funded by the National Institute for Health Research (NIHR) Applied Research Collaboration West Midlands, the NIHR School for Primary Care Research and a NIHR Research Professorship in General Practice (NIHR-RP-2014-04-026). The views expressed in this paper are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.

[1]H. Twohig, MRCP, MRCGP, S. Muller, PhD, Primary Care Centre Versus Arthritis, School of Primary, Community and Social Care, Keele University, Staffordshire, UK; [2]C. Owen, MBBS (Hons), FRACP, Department of Rheumatology, Austin Health, and Department of Medicine, University of Melbourne, Melbourne, Australia; [3]C.D. Mallen, FRCGP, PhD, S. Hider, FRCP, PhD, Primary Care Centre Versus Arthritis, School of Primary, Community and Social Care, Keele University, and Midlands Partnership Foundation Trust, Staffordshire, UK; [4]C. Mitchell, FRCGP, MD, Academic Department of Primary Medical Care, University of Sheffield, Sheffield, UK; [5]C. Hill, FRACP, MD, Rheumatology Unit, The Queen Elizabeth and Royal Adelaide Hospitals, and Discipline of Medicine, The University of Adelaide, Adelaide, Australia; [6]B. Shea, PhD, Ottawa Hospital Research Institute, University of Ottawa, Ottawa, Ontario, Canada; [7]S.L. Mackie, MRCP, PhD, Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds, and NIHR Leeds Biomedical Research Centre, Leeds Teaching Hospitals NHS Trust, Woodhouse, Leeds, UK.

S.L. Mackie declares consultancy to Roche, Chugai, and Sanofi on behalf of University of Leeds (no money paid to her directly in last 3 years); Patron of PMRGCAuk; current or recent site investigator on clinical trials for GSK and Sanofi; and EULAR2019 attendance supported by Roche. S. Muller is a trustee of PMRGCAuk.

Address correspondence to Dr. H. Twohig, Primary Care Centre Versus Arthritis, School of Primary, Community and Social Care, Keele University, Staffordshire, UK. Email: h.j.twohig@keele.ac.uk.

Full Release Article. For details see Reprints and Permissions at jrheum.org

Accepted for publication July 13, 2020.

there are many unanswered questions about its management[4]. A core outcome measurement set of standardized instruments for use in clinical studies of PMR would make it easier to synthesize future research evidence.

In 2016, a core domain set ("what" to measure) was endorsed by the Outcome Measures in Rheumatology (OMERACT) group. This comprises pain, stiffness, physical function, and systemic inflammation[5]. We now need to establish "how" to best measure these domains. A previous systematic review[6] found a wide range of instruments had been used but was limited in its search strategy and inclusion criteria, and did not assess the quality of the evidence found. Further, no review of the evidence for measurement properties of instruments in PMR has been carried out.

We therefore set out to systematically (1) identify all of the outcome measures and instruments previously used in clinical studies of PMR, and (2) evaluate the literature on the measurement properties of selected instruments to determine whether they sufficiently met the OMERACT Filter 2.1 requirements for discriminative ability[7].

MATERIALS AND METHODS

Protocol and registration. The review protocol was registered in Prospero (www.crd.york.ac.uk/prospero; registration number CRD42017080058).

Eligibility criteria. Studies were eligible if they included patients with PMR and reported original quantitative data on outcomes of PMR. A range of study types, including randomized controlled trials (RCT), other interventional trials, prospective cohort studies, case control studies, and cross-sectional studies, were eligible for inclusion. Editorials, commentaries, review articles, case reports, and letters were excluded.

Studies evaluating measurement properties of an instrument in patients with PMR were included and tagged to identify them for the second part of the review process.

Studies that considered patients with PMR and giant cell arteritis (GCA) as a single group (i.e., PMR-specific data not available), diagnostic studies, and studies that solely reported outcomes not pertaining directly to PMR (e.g., cardiovascular events in patients with PMR) were excluded.

Information sources. Five databases [MEDLINE (OVID), CINAHL (EBSCO), Embase (HDAS), Web of Science, and the Cochrane Library] were searched from inception until September 30, 2017.

Clinical trial registries (ClinicalTrials.gov, ISCTRN, and the EU Clinical Trials Register) were reviewed to track any unpublished studies. Experts in the field were contacted to see if they were aware of any ongoing studies of relevance.

Searches. The search strategy (Table 1) was developed by the lead author (HT) with advice from a specialist health librarian. It was based on the MeSH term "polymyalgia rheumatica" and adapted for each database.

Table 1. Search strategy for OVID Medline.

1	polymyalgia rheumatica.mp
2	Polymyalgia Rheumatica/
3	rheumatic polymyalgia.mp
4	polymyalgia arteritica.mp
5	forestier certonciny syndrome.mp
6	rheumatic myalgia.mp
7	rhizomelic pseudopolyarthritis.mp
8	polymyalgi*.mp
9	senile gout.mp
10	1–9 combined with OR

Study selection. Identified studies were imported into Endnote X8 (endnote.com) and duplicates removed. HT screened these titles and uploaded eligible studies to Covidence (www.covidence.org). HT screened all abstracts and full texts against the inclusion and exclusion criteria, and each was independently screened by 1 other review author (CO, SM, CDM, CM, or CH). Disagreements were resolved by discussion and, if needed, by consensus with a third reviewer (SH).

Data collection. Data from all included studies were extracted by HT. A second review author (CO, SM, CDM, CM, CH, or SH) checked the extracted data for each. Extracted data comprised lead author, journal, and year of publication; study design; setting; criteria used to define PMR; sample size; participant age and sex distribution; type of intervention; duration of follow-up; outcomes measured; instruments used; and key findings.

Data extraction for the review of measurement properties was carried out independently by HT and CO. The additional information extracted for studies of measurement properties comprised measurement properties evaluated, methods used, and findings in relation to the measurement properties.

Risk of bias. To inform judgment of overall study quality, risk of bias was assessed using criteria from 3 domains of the Quality In Prognosis Studies (QUIPS) tool[8]: domains 1 (study participation), 2 (study attrition), and 4 (outcome measurement). The other 3 domains of the QUIPS tool were not applied, as they were not relevant to all study types in the review. Additional relevant criteria from the Cochrane Risk of Bias tool[9] were applied to included RCT (adequacy of the randomization and blinding process, and whether the groups were treated equally throughout).

Risk of bias assessment was carried out at the same time as data extraction. Studies were categorized as high, moderate, or low risk for each domain. HT carried out this process with review by a second team member (CO, SM, CDM, CM, CH, or SH). Any disagreements were discussed, and consensus was reached.

The assessment of risk of bias for each study was used in critical judgment of the weight given to the study when deciding which outcome measures to take forward for evaluation of their measurement properties.

Strengths and limitations of studies of measurement properties were evaluated independently by HT and CO. Studies were assessed against the COSMIN-OMERACT Good Methods checklist (Table 2) and given a rating to signify whether they should be used as evidence for each measurement property evaluated (red = no, do not use this as evidence; amber = some cautions but this will be used as evidence; green = yes, likely low risk of bias). Results of this assessment were discussed with the wider review team and used to inform overall judgment on whether there was sufficient evidence to support the use of the instrument in PMR.

Planned methods of analysis. Outcomes and instruments were categorized according to the core domain set agreed upon in 2016 by the OMERACT PMR Working Group[5]. Instruments measuring domains that were not in the core set were also collated to establish other constructs assessed in studies of PMR to inform the future research agenda. A narrative review of the results was carried out.

The findings and quality assessment of all studies on individual measurement properties of each selected instrument were tabulated. This information was synthesized into an overall rating of the body of evidence for each measurement property of each instrument in PMR.

RESULTS

Study selection. Forty-six studies were selected for inclusion in the review (Figure 1). No additional studies meeting the eligibility criteria were identified from reference lists or through contacting experts in PMR. Eight ongoing or unpublished studies were identified from clinical trial registries.

Study characteristics. The 46 included studies were carried out

Table 2. Quality criteria for each measurement property, taken from the COSMIN-OMERACT Good Methods checklist[27].

Measurement Property	Quality Criteria
Construct validity (hypothesis testing)	Clear description given of the construct measured by the comparator instrument Measurement properties of the comparator instrument described and adequate Design and statistical methods adequate for the hypothesis to be tested Otherwise free of any important flaws
Test-retest reliability	Patients stable in the interim period Time interval appropriate Test conditions similar for the measurements Correct statistic used (ICC for continuous data, κ for dichotomous/ordinal/nominal scores) Otherwise free of important flaws
Responsiveness (longitudinal construct validity)	Criteria for change considered an adequate gold standard or the construct for change is clear, either as a situation of change or an actual indicator of change Measurement properties of the comparator standard described and adequate Statistical methods appropriate for the testing situations: · For comparison to gold standard: ROC, AUC, predicative values, sensitivity and specificity, correlation of change with external anchor · For constructs: effect size, SRM, correlation Otherwise free of important flaws
Clinical trial discrimination	Time interval between testing stated and appropriate A proportion of people were expected to change in 1 or both groups *A priori* hypotheses stated regarding the anticipated mean differences in change scores between subgroups (positive, negative, or no change expected) Statistical methods adequate for the hypotheses tested (relative efficiencies, pooled treatment effect sizes, standardized mean differences) Otherwise free of important flaws
Thresholds of meaning	Patient group similar to target population Criterion (external anchor, benchmarks, comparable population) selected in a credible manner Analysis done separately for improvement and deterioration or only in direction anticipated in the target application Multiple criteria used and results triangulated Analysis includes either a Youden index threshold from ROC or another cutoff on a ROC approach. If a threshold approach was used, was it tested for diagnostic utility (sensitivity and specificity)? Otherwise free of any flaws

AUC: area under the curve; COSMIN-OMERACT: COnsensus-based Standards for the selection of health Measurement Instruments–Outcome Measures in Rheumatology; ICC: intraclass correlation coefficient; ROC: receiver-operating characteristic curve; SRM: standardized response mean.

between 1995 and 2017. Forty were carried out in Europe, 5 in North America, and 1 in Japan. Only 1 study recruited exclusively from primary care[10].

Study types. The most frequent study type was prospective cohort study (n = 23), followed by RCT (n = 10). There were 5 pilot efficacy/safety studies, 3 nonrandomized, noncontrolled intervention studies, 3 case series, and 2 case-control studies.

Numbers of participants and follow-up. The sample size of individual studies ranged from 4[11] to 652[10]. Aside from the study by Cawley, *et al*[10], all studies had < 150 participants. In longitudinal studies, follow-up duration ranged from 4 weeks to 4 years.

Age and sex of participants. Mean age ranged from 62 to 78 years, and most studies (n = 42) had more female than male participants.

Criteria used for diagnosis. A range of classification criteria were used to identify participants with PMR. The most commonly used were the Healey[12] and Chuang[13] criteria (9 and 8 studies, respectively). Five studies used the 2012 American College of Rheumatology/European League Against Rheumatism criteria[14], 6 used Bird criteria[15], and 6 used Jones and Hazleman criteria[16]. Twelve studies used clinician diagnosis or a specified combination of clinical features.

Risk of bias within studies. Thirteen of 46 studies were judged to have low risk of bias using the study participation domain as a marker of overall risk of bias. Twenty-five were judged to have a moderate risk of bias, and 8 were judged to have a high risk of bias. The most common reasons for high risk of bias rating were inadequate information about the recruitment process/response rate and small sample size for the study design.

Those judged to be at a low risk of bias did not measure noticeably different outcomes to studies where risk of bias was higher, and therefore the rating did not significantly influence the decision on which outcome measures to evaluate further.

Outcomes measured. A summary of outcomes measured by domain is given in Table 3. Eighteen of 46 studies measured an outcome representing each of the core OMERACT domains, of which only 2 were RCT[17,18].

Laboratory markers of inflammation. Laboratory markers of

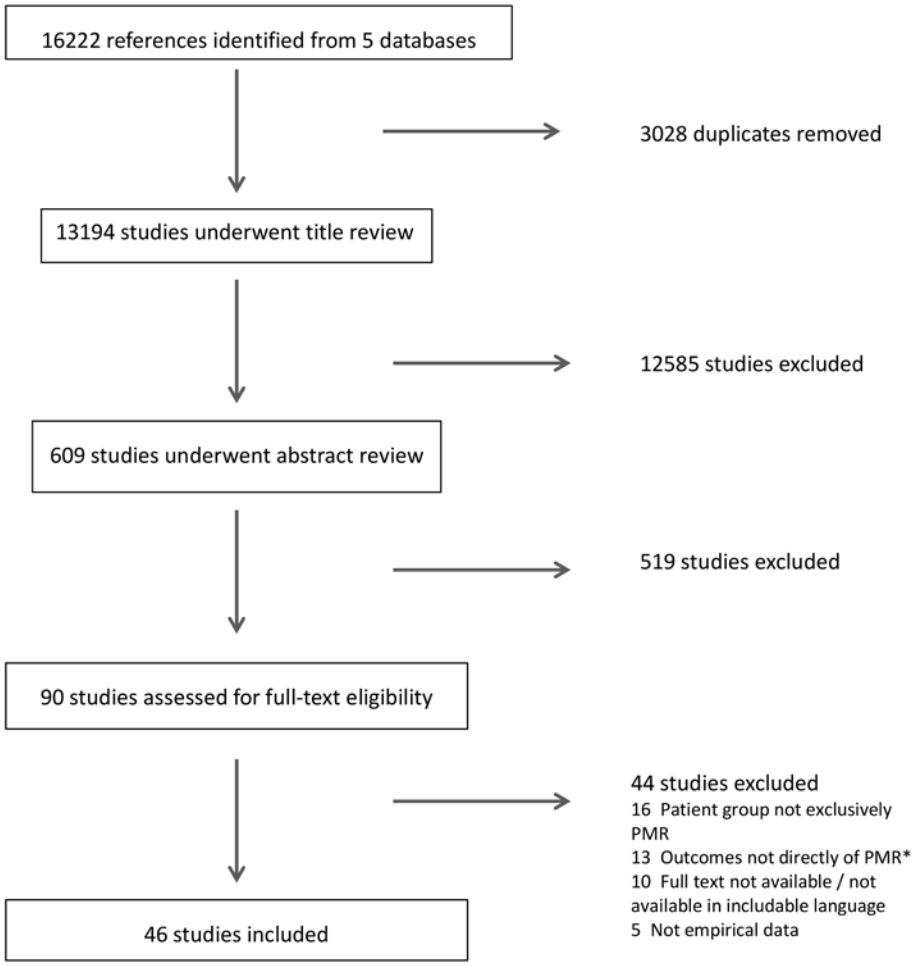

Figure 1. PRISMA diagram of study selection process. PMR: polymyalgia rheumatica. *Studies in this group included those that examined outcomes such as rates of cardiovascular disease or fractures in PMR or that analyzed biochemical markers involved in the pathogenesis of the disease.

inflammation were reported in 43 of 46 studies. Most studies measured both erythrocyte sedimentation rate (ESR) and C-reactive protein (CRP; n = 32). The 5 measuring only ESR were from before the year 2000, whereas the 5 measuring only CRP were published after.

Pain. Thirty-two of 46 studies assessed pain. The most common instrument used (n = 29) was a pain severity visual analog scale (VAS), but the anchor question was rarely stated.

Stiffness. Twenty-eight of 46 studies included an assessment of stiffness. In 26 studies, duration of morning stiffness in minutes was recorded. Four studies additionally assessed stiffness severity using either a VAS or numeric rating scale (NRS).

Physical function. Twenty-two of 46 studies assessed physical function, with 8 of these using > 1 measure of function. In 13 studies, the functional assessment was "elevation of the upper limbs" on a 0–3 scale, measured as part of the composite Polymyalgia Rheumatica Activity Score (PMR-AS)[19], which is defined as follows:

$$\text{CRP} + \text{MST} \times 0.1 + \text{VASpain} + \text{VASphysician} + \text{EUL0–3}$$

where CRP is measured in mg/dL, MST is morning stiffness duration in minutes, VAS has a possible range: 0-10), and EUL is elevation of the upper limbs (possible range 0–3).

Twelve studies used the Health Assessment Questionnaire (HAQ)[20] in some form, either the Health Assessment Questionnaire–Disability Index (HAQ-DI; n = 9) or the modified HAQ (mHAQ; n = 3).

Disease activity/global assessment. Thirteen of 46 studies recorded PMR-AS[19]. Six studies that did not use the PMR-AS included a physician global assessment VAS. Nine studies included some form of patient global assessment. The wording of the questions and the scales for the global VAS varied between studies.

Imaging. Nine of 46 studies included a form of imaging in their outcome set. In 5 of these, assessment of the utility of the imaging technique in PMR was part of the study's aims.

Ongoing or unpublished studies. Five of the ongoing or

Table 3. Summary of outcomes measured by domain.

Domain	No. Studies Assessing This Domain, n (%)	Most Frequent Instrument Used (no. studies)	Other Instruments Used (no. studies)
Laboratory markers of inflammation	43/46 (93)	ESR/CRP (42)	· IL-6 (10) · Fibrinogen (6) · TNF-α (1)
Pain	32/46 (70)	VAS (29)	· NRS (2) · Physician assessment of pain (1) · Pain site manikins (2)
Stiffness	28/46 (63)	Morning stiffness duration in minutes (26)	· Stiffness severity VAS/NRS (4) · Physician assessment of stiffness (1) · Stiffness site manikins (2)
Physical function	22/46 (48)	Elevation of upper limbs on 0–3 scale (13)	· HAQ (12) · SF-36 physical component[36] (3) · American Rheumatism Association functional class assessment[37] (1)
Global assessment/disease activity	21/46 (46)	PMR-AS (13)	· PGA (6) · PtGA (9)
Imaging	9/46 (2)	Ultrasound (6)	· MRI (3) · FDG PET-CT (2)
Other			
Physical examination, presence of synovitis, fever, or weight loss	10		
No. relapses, duration of treatment, or cumulative steroid dose	7		
Other blood variables (e.g., FBC, HbA1c, ACTH/cortisol)	17		
Fatigue	6	VAS (4)	· NRS (1) · Time to onset of fatigue for daily chores (1)
Health status	5	Unspecified questionnaire/VAS (4)	Back-to-normal question (1)
Mood/anxiety	1	GAD-7[38] (1) PHQ-8[39] (1)	

OMERACT core set domains in bold. CRP: C-reactive protein; ESR: erythrocyte sedimentation rate; FBC: full blood count; FDG-PET CT: fluorodeoxyglucose–positron emission tomography/computed tomography; GAD-7: Generalized Anxiety Disorder-7; HAQ: Health Assessment Questionnaire; IL: interleukin; MRI: magnetic resonance imaging; NRS: numeric rating scale; PGA: physician global assessment; PHQ-8: Patient Health Questionnaire-8; PMR-AS: Polymyalgia Rheumatica Activity Score; PtGA: patient global assessment; SF-36: 36-item Short Form Health Survey; TNF: tumor necrosis factor; VAS: visual analog scale.

unpublished studies specified their outcomes. While there were no new outcomes used among these, 3/5 measured fatigue, and 2/5 measured stiffness severity as well as duration of morning stiffness, possibly suggesting a trend toward these factors being attributed greater importance.

Evaluation of measurement properties. The OMERACT PMR Special Interest Group, comprising clinicians, researchers, and patient partners, met in 2018 to determine whether instruments mapping to the core domains had satisfied tests for domain match and feasibility, and if they should continue through the remaining steps of the OMERACT 2.1 Filter. This process has been described in detail in a previous publication[21]. Results from the first part of the review informed this discussion, and the following instruments were selected for further evaluation: laboratory markers of inflammation (CRP and ESR), pain (VAS and NRS), stiffness (VAS, NRS, and duration of morning stiffness), and function (mHAQ and HAQ-DI).

Through the search strategy described, 5 studies were identified that evaluated measurement properties of these instruments. Results of the appraisal of these studies are summarized in Table 4. Table 5 presents an overview of the quality of evidence that exists for each instrument.

The standardized OMERACT Summary of Measurement Properties tables were also completed for each instrument, and the example for pain VAS is available as Supplementary Material (available from the authors on request).

Pain VAS. No studies explicitly aimed to assess construct validity, but the reporting of the change in pain VAS in response to treatment, and the correlation between pain VAS and other instruments demonstrated in Leeb[22] and Matteson[23], can be taken as some evidence supporting the validity of this measure in assessing PMR-related pain. However, neither study set out hypotheses about the expected relationship with other outcomes, and the comparator measures used were either not themselves validated

Table 4. Critical appraisal of the studies of measurement properties of instruments considered for inclusion in the core outcome set.

Instrument	Measurement Property	First Author, Yr	Quality Assessment	Findings	Rating
Pain VAS	Construct validity	Leeb 2003[22]	Comparison made to pretreatment levels and correlation between VAS pain and other instruments was assessed. No *a priori* hypotheses about magnitude of change or strength of correlation with other instruments stated. The comparator instruments were not measuring the same construct and/or were not themselves validated in PMR.	Highly significant improvement at W24 compared to baseline. VAS pain was highly correlated with other measures including ESR/CRP and duration of morning stiffness. Multiple regression analysis with VAS pain as the dependent variable showed that it correlated with self-reported myalgia and elevation of the upper limbs.	Red
		Matteson 2012[23]	Comparison made to pretreatment levels No *a priori* hypotheses about magnitude of change or correlation with other instruments stated.	Statistically significant improvement between baseline and W1 and W1 and W4 but not between W4 and W26.	Red
	Responsiveness	McCarthy 2014[25]**	Situation of change clear (newly diagnosed, started on treatment). PMR-AS used as gold standard for assessment of remission (accepted as a validated measure). Statistical methods were appropriate but no hypotheses about magnitude of change were made.	SRM = 0.89 ESS = 0.96	Red
		Kalke 2000[24]	Small sample size, n = 18 Situation of change clear (newly diagnosed, started on treatment). Statistical methods are appropriate but no hypotheses about magnitude of change were made.	SRM = 1.7	Red
	Test-retest reliability	Matteson 2012[23]	Small sample size, n = 14 Patients were stable in the interim time period; the time period was appropriate and test conditions were stable. Statistical methods were appropriate (ICC).	Global pain ICC = 0.82	Amber
	Thresholds of meaning	Matteson 2012[23]	Patient group is sufficiently similar to target population. Not enough information on methods given. No attempt to calculate minimally important difference to patients.	SDC and %MDC = 28.9	Red
Duration of morning stiffness	Construct validity	Leeb 2003[22]	Comparison made to pretreatment levels. No *a priori* hypotheses about magnitude of change or strength of correlation with other instruments stated.	Highly significant improvement at W24 compared to baseline.	Red
		Matteson 2012[23]	Comparison made to pretreatment levels. No *a priori* hypotheses about magnitude of change or correlation with other instruments stated.	Statistically significant improvement between baseline and W1 and W1 and W4 but not between W4 and W26.	Red
	Responsiveness	McCarthy 2014[25]	Situation of change clear in active group (newly diagnosed, started on treatment). PMR-AS used as gold standard for assessment of remission (accepted as a validated measure). Statistical methods were appropriate but no hypotheses about magnitude of change were made.	SRM = 0.89 ESS = 0.96	Red
		Kalke 2000[24]	Small study, n = 18 Situation of change clear (newly diagnosed, started on treatment). Statistical methods are appropriate but no hypotheses about magnitude of change were made.	SRM = 1.7	Red
	Test-retest reliability	Matteson 2012[23]	Small sample size, n = 14 Patients were stable in the interim time period; the time period was appropriate and test conditions were stable. Statistical methods were appropriate (ICC).	ICC 0.11	Red

Table 4. Continued.

Instrument	Measurement Property	First Author, Yr	Quality Assessment	Findings	Rating
	Thresholds of meaning	Matteson 2012[23]	Patient group is sufficiently similar to target population. Not enough information on methods given. No attempt to calculate minimally important difference to patients.	SDC = 231 %MDC = 16.1	Red
HAQ-DI	Construct validity	Kalke 2000[24]	Small sample size, n = 18 No clear description of the construct measured by the comparator instrument (not measures of function). No *a priori* hypotheses about magnitude of change or correlation with other instruments stated.	Significant improvement in HAQ score between pre- and posttreatment measurements Linear regression coefficient with duration of morning stiffness, pain VAS, and CRP were 0.66, 0.72, and 0.63, respectively	Red
	Responsiveness	Kalke 2000[24]	Small sample size, n = 18 Situation of change clear (newly diagnosed, started on treatment). Statistical methods are appropriate but no hypotheses about direction of change or strength of correlation between instruments were made.	SRM = 3	Red
mHAQ	Construct validity	Matteson 2012[23]	Each instrument was compared to its pretreatment levels. No *a priori* hypotheses about magnitude of change or correlation with other instruments stated.	Statistically significant improvement at all measurement timepoints	Red
		McCarthy 2014[25]	Each instrument was compared to its pretreatment levels. Comparator measures were not evaluating the same construct. No *a priori* hypotheses about magnitude of change or correlation with other instruments stated.	Statistically significant improvement between W1 and W6 in the active group. Correlation coefficients between mHAQ and PMR-AS, ESR and CRP were 0.68, 0.45 and 0.39 respectively	Red
	Responsiveness	McCarthy 2014[25]	Situation of change clear in active group (newly diagnosed, started on treatment). PMR-AS used as gold standard for assessment of remission; accepted as a validated measure. Statistical methods were appropriate but no hypotheses about magnitude of change were made.	SRM = 1.36 ESS = 1.65	Red
	Test-retest reliability	Matteson 2012[23]	Small sample size, n = 14 Patients were stable in the interim time period; the time period was appropriate and test conditions were stable. Statistical methods were appropriate (ICC).	ICC = 0.72	Amber
	Thresholds of meaning	Matteson 2012[23]	Patient group is sufficiently similar to target population. Not enough information on methods given. No attempt to calculate minimally important difference to patients.	SDC = 0.78 %MDC = 25.9	Red
ESR/CRP	Construct validity	Leeb 2003[22]	Each instrument was compared to its pretreatment levels and correlation between VAS pain and ESR/CRP was assessed. No *a priori* hypotheses about magnitude of change or strength of correlation with other instruments stated.	Highly significant improvement at W24 compared to baseline.	Red
		Matteson 2012[23]	Each instrument was compared to its pretreatment levels. No *a priori* hypotheses about magnitude of change or correlation with other instruments stated.	Statistically significant improvement between baseline and W1 and W1 and W4 but not between W4 and W26.	Red

Table 4. Continued.

Instrument	Measurement Property	First Author, Yr	Quality Assessment	Findings	Rating
		McCarthy 2014[25]	Each instrument was compared to its pretreatment levels. Comparator instrument for correlation was the mHAQ which measures a different construct. No explicit *a priori* hypotheses about magnitude of change or correlation with other instruments stated.	Statistically significant improvement from W1 to W6 in the active group Correlation coefficient between mHAQ and ESR/CRP = 0.45/0.39	Red
	Responsiveness	McCarthy 2014[25]	Situation of change clear in active group (newly diagnosed, started on treatment). PMR-AS used as gold standard for assessment of remission – accepted as a validated measure. Statistical methods were appropriate but no hypotheses about magnitude of change were made.	ESR SRM/ESS = 1.2/1.15 CRP SRM/ESS = 1.05/1.14	Red
		Kalke 2000[24]	Small study, n = 18 Situation of change clear (newly diagnosed, started on treatment). Statistical methods are appropriate but no hypotheses about magnitude of change were made.	CRP SRM 1.6	Red
	Thresholds of meaning	McCarthy 2013[28]**	Appropriate patient group. Criteria for assessment of disease activity and definition of remission satisfactory. Thresholds for ESR and CRP cutoffs justified from the literature. Statistical methods satisfactory but did not use multiple methods to triangulate findings.	Ability of ESR > 40 mm/h or CRP > 6 mg/L to detect active disease: · ESR: sensitivity 92%, specificity 66%, PPV 0.72, LR 2.8 · CRP: sensitivity 100%, specificity 70%, PPV 0.77, LR 3.33 Ability of ESR < 20 mm/h or CRP < 6 mg/L to detect disease remission: · ESR: sensitivity 43%, specificity 75%, PPV 0.87, LR 1.7 · CRP: sensitivity 58%, specificity 67%, PPV 0.88, LR 2.04	Amber

CRP: C-reactive protein; ESR: erythrocyte sedimentation rate; ESS: effect size statistic; HAQ: Health Assessment Questionnaire; ICC: intraclass correlation coefficient; LR: likelihood ratio; MDC: minimal detectable change; mHAQ: modified Health Assessment Questionnaire; PMR: polymyalgia rheumatica; PMR-AS: Polymyalgia Rheumatica Activity Score; PPV: positive predictive value; SDC: smallest detectable change; SRM: standardized response mean; VAS: visual analog scale; W: week.

Table 5. Summary of quality of evidence on measurement properties of outcome measurement instruments in PMR.

	Construct validity	Test-retest Reliability	Responsiveness	Thresholds of Meaning
Pain VAS	–	+	–	–
Stiffness VAS	N/A	N/A	N/A	N/A
Duration of morning stiffness	–	–	–	–
HAQ–DI	–	N/A	–	N/A
mHAQ	–	+	–	–
ESR and CRP	–	N/A	–	+

Evaluation of evidence defined as N/A: not evaluated; –: evaluated but insufficient evidence to support use in clinical studies; +: evaluated and some evidence to support use; ++: good evidence to support use in clinical studies. CRP: C-reactive protein; ESR: erythrocyte sedimentation rate; HAQ-DI: Health Assessment Questionnaire–Disability Index; mHAQ: modified Health Assessment Questionnaire; PMR: polymyalgia rheumatica; VAS: visual analog scale.

in PMR or they measured a different construct altogether. Both were rated red against the Good Methods checklist.

Responsiveness of the pain VAS was evaluated in 2 studies[24,25]. Neither study stated hypotheses about the anticipated change in response to treatment or the magnitude of the anticipated effect size *a priori*, and again, both were rated red for this measurement property.

Test-retest reliability of a pain VAS was evaluated by Matteson, *et al*[23]. The methods were appropriate and the result suggests good reliability, but the small sample size (n = 14) meant that this study was rated amber.

The % minimal detectable change (MDC) for pain VAS was calculated in the same small subgroup in this study (n = 14)[23]. This was the only study looking at any thresholds of meaning for a pain VAS in PMR. The authors did not evaluate what a minimally important change might be for patients, and the study was rated red for this measurement property as well.

Duration of morning stiffness. The 4 studies that evaluated

measurement properties of pain VAS all also evaluated duration of morning stiffness[22,23,24,25]. The limitations to the methods discussed above also applied for this outcome measure, and test-retest reliability was poorer. All were rated red for all measurement properties.

HAQ-DI. Kalke, *et al*[24] evaluated the construct validity and responsiveness of the HAQ as an assessment of function in PMR, but significant limitations meant it was rated red for both measurement properties.

Construct validity was evaluated by studying correlation of the HAQ with duration of morning stiffness, pain VAS, and CRP, none of which are measures of function. The correlation was good (> 0.6 in each case), but no hypotheses about the magnitude of change or strength of correlation were stated. Responsiveness was evaluated using the standardized response mean (SRM). The SRM was higher for the HAQ than for the other measures in this study, suggesting greater responsiveness to change, but no *a priori* hypotheses were stated.

mHAQ. Two studies evaluated the mHAQ, covering the full range of measurement properties between them[23,25], but they were rated red for all measurement properties except test-retest reliability.

Both studies provide some evidence toward the construct validity of the mHAQ through demonstrating its improvement in response to treatment[23,25]. McCarthy, *et al* also demonstrated correlation of the mHAQ with other outcome measures[25], but the comparator measures were not measures of function.

Responsiveness of the mHAQ was evaluated by McCarthy, *et al* using appropriate statistical methods, but no hypothesis about the magnitude of change was given[25].

Test-retest reliability of the mHAQ was evaluated by Matteson, *et al*[23]. The ICC was 0.72, but the small sample size prevented the study being rated green[27]. The %MDC was calculated in the same study, but there was limited information on the methods and no attempt to determine a minimally important difference to patients.

ESR/CRP. Construct validity was supported by 3 studies[22,23,25], which all confirmed that ESR and CRP improved with treatment of PMR. McCarthy, *et al* found moderate correlation between ESR/CRP and the mHAQ[25], but these instruments do not measure the same construct. None of the studies set out hypotheses about expected relationships, and all 3 studies were rated red.

Responsiveness was evaluated in 2 studies[24,25], but neither set out hypotheses about magnitude of change *a priori*. One study[28] addressed thresholds of meaning for ESR and CRP, and was rated amber. This study found that CRP was superior to ESR in detecting active disease and disease remission.

DISCUSSION

We identified all the outcome measures and instruments used to date in studies of PMR and categorized them using the PMR Core Domain Set endorsed by OMERACT in 2016. Results from the first part of the review informed the decision on which instruments to evaluate as candidates for inclusion in a core instrument set. Only 5 studies evaluating measurement properties of candidate instruments in populations with PMR were identified. Crucially, none of the studies were rated green for any of the measurement properties when assessed against the COSMIN-OMERACT Good Methods criteria. For pain VAS and the mHAQ, there was 1 study of test-retest reliability, which achieved amber, and there was 1 study considering thresholds of meaning for ESR/CRP, which was also rated amber.

The majority of PMR studies included in this review were cohort studies, with only 10 RCT. Almost all had sample sizes of fewer than 150 participants. We found that outcome measures used in studies of PMR varied widely and were often poorly defined. This makes comparing results across studies very difficult and prevents synthesis of current data to improve the evidence base.

Systemic inflammation was most frequently assessed of the 4 PMR core domains, followed by pain and stiffness. Physical function was measured least often. This contrasts with findings from qualitative studies where patients with PMR have highlighted disability and stiffness as having significant effect on their quality of life[29,30].

Pain was the most commonly assessed patient-reported outcome, with VAS being the most frequently used measurement instrument. However, as noted in previous reviews[6,31], there is little consistency in the question and scales used or on the time frame being considered. Each measurement property of pain VAS has been evaluated in PMR, but there is only sufficient evidence on its test-retest reliability.

Stiffness was measured in 28/46 studies in this review. Given that it is a cardinal symptom of PMR, this is notably low. No studies evaluated a stiffness severity VAS despite the widely acknowledged limitations of "duration of morning stiffness" as an outcome measure[30,32,33]. We did not find sufficient evidence for any measurement property of duration of morning stiffness to support its use in PMR.

Physical function was assessed in the least consistent way of the core domains. Most frequently, it was measured as part of the PMR-AS, an overall assessment of disease activity that includes evaluation of "elevation of the upper limbs" on a 0–3 scale. This is a very limited assessment of overall function and is insufficient to represent this domain[29,30]. Therefore, the measurement properties of mHAQ and HAQ-DI were reviewed. We found that neither mHAQ nor HAQ-DI had high-quality evidence to support their use as an outcome measure in PMR. Since physical function is of prime importance to people's daily lives, the failure to measure it in a meaningful, reliable way that allows comparison across studies of PMR needs addressing.

Where inflammatory markers are used in studies of PMR, ESR and CRP are usually both measured. In studies that chose one over the other, more recent studies tended to use CRP. ESR and CRP are used to evaluate many rheumatological conditions and are frequently incorporated into disease activity scores. Certain properties of biomarkers, such as face validity and feasibility, are likely to be transferrable across conditions. However, properties such as responsiveness and test-retest reliability may vary between conditions, and the limited evaluation in patients

with PMR is therefore of note. Indeed, up to 20% of people with PMR may have normal ESR or CRP before treatment; the relationship between these biomarkers and PMR disease activity is not straightforward[34].

A small number of studies measured domains that were outside of the core set but included in the "important" or "research agenda" list by the OMERACT 2016 group[35]. These include fatigue, psychological effect, and overall health status. Although these constructs are heavily intertwined, with each other and with pain, stiffness, and function, this may signify a gap in the core domain set. An overall measure of PMR-related quality of life could be of value in addressing this gap.

The exclusion of papers considering PMR and GCA as a single group is a potential source of bias. However, the risk of bias from including participants with GCA is high and outweighs the small risk of having missed any outcome measure of relevance. One exception to this rule was made in 2 papers (arising from 1 study) by McCarthy, *et al*, in which 1 participant out of 60 had biopsy-proven GCA as well as PMR[25,28]. This decision was made by the team because there were so few studies on measurement properties of instruments in PMR that these 2 papers contributed substantially to the available data, and it was felt that there was minimal risk of bias from 1 participant having a dual diagnosis.

Risk-of-bias assessment of included studies added value in this review, as it had not been done previously, to our knowledge. This is a subjective process but was carried out using an established tool and verified by a second assessor. That only 13 of the 46 studies demonstrated low risk of bias shows the limitations of the evidence base in PMR and has implications for the ability to draw firm conclusions from this review. This highlights the need to identify high-quality, well-documented datasets from modern clinical studies of PMR for further evaluation of instrument properties, as well as the need for a core outcome measurement set incorporating the best-performing instruments in order to standardize secondary outcomes across future trials.

Measurement of outcomes in studies of PMR lacks consistency. The critical patient-centered domain of physical function is the least frequently measured of the OMERACT core domains and, when it is measured, is often assessed only by ability to elevate the upper limbs. Overall, none of the candidate instruments considered for inclusion in the core outcome set had high-quality evidence, from studies in populations with PMR, on their full range of measurement properties. This is in part because there are very few published instrument validation studies. We are planning further studies reexamining individual patient data to determine whether the selected instruments are suitable for a core outcome measurement set for PMR.

ACKNOWLEDGMENT

We would like to thank the wider OMERACT PMR Working Group for their contribution to this study.

RERERENCES

1. Crowson CS, Matteson EL, Myasoedova E, Michet CJ, Ernste FC, Warrington KJ, et al. The lifetime risk of adult-onset rheumatoid arthritis and other inflammatory autoimmune rheumatic diseases. Arthritis Rheum 2011;63:633-9.
2. Salvarani C, Cantini F, Hunder GG. Polymyalgia rheumatica and giant-cell arteritis. Lancet 2008;372:234-45.
3. Partington RJ, Muller S, Helliwell T, Mallen CD, Abdul Sultan A. Incidence, prevalence and treatment burden of polymyalgia rheumatica in the UK over two decades: a population-based study. Ann Rheum Dis 2018;77:1750-6.
4. Dejaco C, Singh YP, Perel P, Hutchings A, Camellino D, Mackie S, et al. Current evidence for therapeutic interventions and prognostic factors in polymyalgia rheumatica: A systematic literature review informing the 2015 European League Against Rheumatism/American College of Rheumatology recommendations for the management of polymyalgia rheumatica. Ann Rheum Dis 2015;74:1808-17.
5. Mackie SL, Twohig H, Neill LM, Harrison E, Shea B, Black RJ, et al; OMERACT PMR Working Group. The OMERACT core domain set for outcome measures for clinical trials in polymyalgia rheumatica. J Rheumatol 2017;44:1515-21.
6. Duarte C, de Oliveira Ferreira RJ, Mackie SL, Kirwan JR, Pereira da Silva JA; OMERACT Polymyalgia Rheumatica Special Interest Group. Outcome measures in polymyalgia rheumatica. A systematic review. J Rheumatol 2015;42:2503-11.
7. Boers M, Beaton DE, Shea BJ, Maxwell LJ, Bartlett SJ, Bingham CO 3rd, et al. OMERACT Filter 2.1: Elaboration of the Conceptual Framework for Outcome Measurement in Health Intervention Studies. J Rheumatol 2019;46:1021-7.
8. Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing bias in studies of prognostic factors. Ann Intern Med 2013;158:280-6.
9. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al; Cochrane Bias Methods Group; Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928.
10. Cawley A, Prior JA, Muller S, Helliwell T, Hider SL, Dasgupta B, et al. Association between characteristics of pain and stiffness and the functional status of patients with incident polymyalgia rheumatica from primary care. Clin Rheumatol 2018;37:1639-44.
11. Salvarani C, Cantini F, Niccoli L, Catanoso MG, Macchioni P, Pulsatelli LI, et al. Treatment of refractory polymyalgia rheumatica with infliximab: a pilot study. J Rheumatol 2003;30:760-3.
12. Healey LA. Long-term follow-up of polymyalgia rheumatica: evidence for synovitis. Semin Arthritis Rheum 1984;13:322-8.
13. Chuang TY, Hunder GG, Ilstrup DM, Kurland LT. Polymyalgia rheumatica: a 10-year epidemiologic and clinical study. Ann Intern Med 1982;97:672-80.
14. Dasgupta B, Cimmino MA, Maradit-Kremers H, Schmidt WA, Schirmer M, Salvarani C, et al. 2012 provisional classification criteria for polymyalgia rheumatica: a European League Against Rheumatism/American College of Rheumatology collaborative initiative. Ann Rheum Dis 2012;71:484-92.
15. Bird HA, Esselinckx W, Dixon AS, Mowat AG, Wood PH. An evaluation of criteria for polymyalgia rheumatica. Ann Rheum Dis 1979;38:434-9.
16. Jones JG, Hazleman BL. Prognosis and management of polymyalgia rheumatica. Ann Rheum Dis 1981;40:1-5.
17. Di Munno O, Imbimbo B, Mazzantini M, Milani S, Occhipinti G, Pasero G. Deflazacort versus methylprednisolone in polymyalgia rheumatica: clinical equivalence and relative antiinflammatory potency of different treatment regimens. J Rheumatol 1995;22:1492-8.
18. Kreiner F, Galbo H. Effect of etanercept in polymyalgia rheumatica: a randomized controlled trial. Arthritis Res Ther 2010;12:R176.

19. Leeb BF, Bird HA. A disease activity score for polymyalgia rheumatica. Ann Rheum Dis 2004;63:1279-83.
20. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.
21. Owen CE, Yates M, Twohig H, Muller S, Neill LM, Harrison E, et al. Toward a Core Outcome Measurement Set for polymyalgia rheumatica: report from the OMERACT 2018 Special Interest Group. J Rheumatol 2019;46:1360-4.
22. Leeb BF, Bird HA, Nesher G, Andel I, Hueber W, Logar D, et al. EULAR response criteria for polymyalgia rheumatica: results of an initiative of the European Collaborating Polymyalgia Rheumatica Group (subcommittee of ESCISIT). Ann Rheum Dis 2003;62:1189-94.
23. Matteson EL, Maradit-Kremers H, Cimmino MA, Schmidt WA, Schirmer M, Salvarani C, et al. Patient-reported outcomes in polymyalgia rheumatica. J Rheumatol 2012;39:795-803.
24. Kalke S, Mukerjee D, Dasgupta B. A study of the health assessment questionnaire to evaluate functional status in polymyalgia rheumatica. Rheumatology 2000;39:883-5.
25. McCarthy EM, MacMullan PA, Al-Mudhaffer S, Madigan A, Donnelly S, McCarthy CJ, et al. Plasma fibrinogen along with patient-reported outcome measures enhances management of polymyalgia rheumatica: a prospective study. J Rheumatol 2014;41:931-7.
26. Dasgupta B, Matteson EL, Maradit-Kremers H. Management guidelines and outcome measures in polymyalgia rheumatica (PMR). Clin Exp Rheumatol 2007;6 Suppl 47:130-6.
27. OMERACT. OMERACT Handbook Instrument Selection Chapter 5 Mar 2019. In: OMERACT handbook 2019. [Internet. Accessed November 19, 2020.] Available from: omeracthandbook.org/handbook
28. McCarthy EM, MacMullan PA, Al-Mudhaffer S, Madigan A, Donnelly S, McCarthy CJ, et al. Plasma fibrinogen is an accurate marker of disease activity in patients with polymyalgia rheumatica. Rheumatology 2013;52:465-71.
29. Twohig H, Mitchell C, Mallen C, Adebajo A, Mathers N. "I suddenly felt I'd aged": a qualitative study of patient experiences of polymyalgia rheumatica (PMR). Patient Educ Couns 2015;98:645-50.
30. Mackie SL, Hughes R, Walsh M, Day J, Newton M, Pease C, et al. "An impediment to living life": why and how should we measure stiffness in polymyalgia rheumatica? PLoS One 2015;10:e0126758.
31. Huang A, Castrejon I. Patient-reported outcomes in trials of patients with polymyalgia rheumatica: a systematic literature review. Rheumatol Int 2016;36:897-904.
32. Halls S, Sinnathurai P, Hewlett S, Mackie SL, March L, Bartlett SJ, et al. Stiffness is the cardinal symptom of inflammatory musculoskeletal diseases, yet still variably measured: Report from the OMERACT 2016 Stiffness Special Interest Group. J Rheumatol 2017;44:1904-10.
33. Halls S, Dures E, Kirwan J, Pollock J, Baker G, Edmunds A, et al. Stiffness is more than just duration and severity: a qualitative exploration in people with rheumatoid arthritis. Rheumatology 2014;54:615-22.
34. Cantini F, Salvarani C, Olivieri I, Macchioni L, Ranzi A, Niccoli L, et al. Erythrocyte sedimentation rate and C-reactive protein in the evaluation of disease activity and severity in polymyalgia rheumatica: a prospective follow-up study. Semin Arthritis Rheum 2000;30:17-24.
35. Helliwell T, Brouwer E, Pease CT, Hughes R, Hill CL, Neill LM, et al. Development of a provisional core domain set for polymyalgia rheumatica: report from the OMERACT 12 Polymyalgia Rheumatica Working Group. J Rheumatol 2016;43:182-6.
36. Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473-83.
37. Hochberg MC, Chang RW, Dwosh I, Lindsey S, Pincus T, Wolfe F. The American College of Rheumatology 1991 revised criteria for the classification of global functional status in rheumatoid arthritis. Arthritis Rheum 1992;35:498-502.
38. Spitzer RL, Kroenke K, Williams JBW, Lowe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166:1092-7.
39. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16:606-13.