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

Helen Twohig¹, Claire Owen²³, Sara Muller¹, Christian Mallen¹⁹, Caroline Mitchell⁴,
Samantha Hider¹⁹, Catherine Hill⁵⁶, Beverley Shea⁷, Sarah Mackie⁸

Running header: Outcome measures in PMR

Key indexing terms: polymyalgia rheumatica, outcome measures, systematic review,
OMERACT

Affiliations

1. Primary Care Centre Versus Arthritis, School of Primary, Community and Social
    Care, Keele University, UK
2. Department of Rheumatology, Austin Health, Melbourne, Australia
3. Department of Medicine, University of Melbourne, Melbourne, Australia
4. Academic Department of Primary Medical Care, University of Sheffield, UK
5. Rheumatology Unit, The Queen Elizabeth and Royal Adelaide Hospitals, Adelaide,
    Australia
6. Discipline of Medicine, The University of Adelaide, Adelaide, Australia
7. Ottawa Hospital Research Institute, University of Ottawa, Ottawa, Canada
8. Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds,
    UK; NIHR Leeds Biomedical Research Centre, Leeds Teaching Hospitals NHS Trust,
    UK
9. Midlands Partnership Foundation Trust, Staffordshire, UK

Downloaded on August 17, 2020 from www.jrheum.org
Funding
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 (ARC) 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.

Conflicts of interest:
Dr Sarah Mackie declares consultancy to Roche, Chugai, 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, Sanofi; EULAR2019 attendance supported by Roche.

Dr Sara Muller is a trustee of PMRGCAuk

Initials, surnames, appointments, and highest academic degrees of all authors:
Dr H.J.Twohig, Wellcome Trust Doctoral Fellow and General Practitioner. MRCP, MRCGP
https://orcid.org/0000-0001-8781-1268

Dr C.E.Owen, Consultant Rheumatologist. MBBS (Hons) FRACP https://orcid.org/0000-0002-2694-5411

Dr S.Muller, Senior Lecturer in Epidemiology and Applied Statistics, PhD

Professor C.D.Mallen, Professor of Primary Care. FRCGP PhD https://orcid.org/0000-0002-2677-1028

Dr C.Mitchell, Senior Clinical Lecturer in Primary Care. FRCGP MD https://orcid.org/0000-0002-4790-0095

Dr S.Hider, Reader & Honorary Consultant Rheumatologist. FRCP PhD
Dr C.Hill, Clinical Professor and Consultant Rheumatologist, FRACP, MD

Dr B.Shea, Clinical Investigator, PhD https://orcid.org/0000-0002-7686-2585

Dr S.L.Mackie, Associate Clinical Professor and Honorary Consultant Rheumatologist, MRCP, PhD, orcid.org/0000-0003-2483-5873

Corresponding Author: Helen Twohig, Primary Care Centre Versus Arthritis, School of Primary, Community and Social Care, Keele University, UK. hj.twohig@keele.ac.uk
Abstract

Objectives: 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 five databases. Two researchers were involved in screening, data extraction and risk of bias assessment. Once outcomes and instruments used were identified and categorised, key instruments were selected for further review through a consensus process. Studies on measurement properties of these instruments were appraised against the COSMIN-OMERACT (Outcome Measures in Rheumatology) checklist to determine the extent of evidence supporting their use in PMR.

Results: 46 studies were included. In decreasing order of frequency, the most common outcomes (and instruments) used were: markers of systemic inflammation (ESR/CRP), pain (visual analogue 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 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-centred 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.
Introduction

Polymyalgia rheumatica (PMR) is the most common inflammatory rheumatic condition of older people (Crowson et al., 2011) and is characterised by proximal pain and stiffness, raised inflammatory markers and a therapeutic response to glucocorticoids (Salvarani, Cantini and Hunder, 2008). A recent UK study using the Clinical Practice Research Datalink found an annual incidence of 96 per 100000 people aged over 40, with incidence rising markedly with increasing age (Partington et al., 2018).

Although it is common, PMR remains under-researched and there are many unanswered questions about its management (Dejaco et al., 2015). A Core Outcome Measurement set of standardised instruments for use in clinical studies of PMR would make it easier to synthesise 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 (Mackie et al., 2017). We now need to establish ‘how’ to best measure these domains. A previous systematic review (Duarte et al., 2015) 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. Furthermore, 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

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 (Boers et al., 2019).
**Materials and Methods**

**Protocol and registration**

The review protocol was registered in Prospero,

[https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=80058](https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=80058) Registration number CRD42017080058.

**Ethics Approval**

No ethical approval was necessary for this systematic review.

**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 randomised controlled trials, 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 via OVID, CINAHL via EBSCO, Embase via HDAS, Web of Science and the Cochrane Library) were searched from inception until September 30th 2017.
Clinical trials 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.

Study selection

Identified studies were imported into Endnote X8 (https://endnote.com) and duplicates removed. HT screened these titles and uploaded eligible studies to Covidence (https://www.covidence.org/home). HT screened all abstracts and full texts against the inclusion and exclusion criteria, and each was independently screened by one 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 gender 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 judgement of overall study quality, risk of bias was assessed using criteria from three domains of the Quality In Prognosis Studies (QUIPS) tool (Hayden et al., 2013); domains 1 (study participation), 2 (study attrition) and 4 (outcome measurement). The other three 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 (Higgins et al., 2011) were applied to included randomised controlled trials (adequacy of the randomisation 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 categorised 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 reached.

The assessment of risk of bias for each study was used in critical judgement of the weight given to the study when deciding which outcome measures to take forwards 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 judgement on whether there was sufficient evidence to support the use of the instrument in PMR.

**Planned methods of analysis**

Outcomes and instruments were categorised according to the core domain set agreed by the OMERACT PMR Working Group in 2016 (Mackie et al., 2017). 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 synthesised into an overall rating of the body of evidence for each measurement property of each instrument in PMR.

**Results**

**Study selection**

46 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 on-going or unpublished studies were identified from clinical trials registries.

**Study characteristics**

The 46 included studies were carried out between 1995 and 2017. 40 were carried out in Europe, five in North America and one in Japan. Only one study recruited exclusively from primary care (Cawley et al., 2017).

**Study types:**

The most frequent study type was prospective cohort study (n=23), followed by randomised controlled trial (n=10). There were five pilot efficacy / safety studies, three non-randomised, non-controlled intervention studies, three case series and two case control studies.

**Numbers of participants and follow up:**
The sample size of individual studies ranged from four (Salvarani et al., 2003) to 652 (Cawley et al., 2017). Aside from the study by Cawley et al., all studies had <150 participants. In longitudinal studies, follow up duration ranged from four weeks to four years.

**Age and gender 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 (Healey, 1984) and Chuang (Chuang et al., 1982) criteria (9 studies and 8 studies respectively). Five studies used the 2012 EULAR / ACR criteria (Dasgupta et al., 2012), six used Bird criteria (Bird et al., 1979) and six used Jones and Hazleman criteria (Jones and Hazleman, 1981). 12 studies used clinician diagnosis or a specified combination of clinical features.

**Risk of bias within studies**

13/46 studies were judged to have low risk of bias using the study participation domain as a marker of overall risk of bias. 25 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 ultimately 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.

18/46 studies measured an outcome representing each of the core OMERACT domains, of which only two were randomised controlled trials ((Di Munno et al., 1995) and (Kreiner and Galbo, 2010)).

**Laboratory markers of inflammation**

Laboratory markers of inflammation were reported in 43/46 studies. Most studies measured both ESR and CRP (n=32). The five measuring only ESR were all from before the year 2000 whereas the five measuring only CRP were all published after the year 2000.

**Pain**

32/46 studies assessed pain. The most common instrument used (n=29) was a pain severity visual analogue scale (VAS) but the anchor question was rarely stated.

**Stiffness**

28/46 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 NRS.

**Physical function**

22/46 studies assessed physical function, with eight of these using more than one 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*(Leeb and Bird, 2004)). 12 studies used the Health Assessment Questionnaire (HAQ(Fries et al., 1980) in some form, either the HAQ-DI (n=9) or the mHAQ (n=3).

**Disease activity / global assessment**
13/46 studies recorded PMR-AS (Leeb and Bird, 2004). 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**

9/46 studies included a form of imaging in their outcome set. In five of these, assessment of the utility of the imaging technique in PMR was part of the study’s aims.

**On-going or unpublished studies**

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

**Evaluation of measurement properties**

The OMERACT PMR-SIG, comprised of 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 (Owen et al., 2019). 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 and NRS and duration of morning stiffness, function – mHAQ and HAQ-DI.

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

The standardised OMERACT Summary of Measurement Properties tables were also completed for each instrument and the example for pain VAS is available as supplementary information.

**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 2003 (Leeb *et al.*, 2003) and Matteson 2012 (Matteson *et al.*, 2012) 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 in PMR or measured a different construct altogether. Both were rated red against the Good Methods Checklist.

Responsiveness of the pain VAS was evaluated in two studies (Kalke, Mukerjee and Dasgupta, 2000; McCarthy *et al.*, 2014). 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.* (Matteson *et al.*, 2012). The methods were appropriate, and the result suggests good reliability but the small sample size (14) meant that this study was rated amber.

The percentage minimal detectable change (MDC) for pain VAS was calculated in the same small sub-group in this study (*n*=14)(Matteson *et al.*, 2012). 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 too.

**Duration of morning stiffness**

The four studies that evaluated measurement properties of pain VAS all also evaluated
duration of morning stiffness (Dasgupta, Matteson and Maradit-Kremers, no date; Kalke,
Mukerjee and Dasgupta, 2000; Leeb et al., 2003; McCarthy et al., 2014). 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. (Kalke, Mukerjee and Dasgupta, 2000) 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 standardised
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 (Matteson et al., 2012; McCarthy et al., 2014), but they were rated red for all measurement properties except test-retest reliability.

Both studies provide some evidence towards the construct validity of the mHAQ through demonstrating its improvement in response to treatment (Matteson et al., 2012; McCarthy et al., 2014). McCarthy et al. also demonstrated correlation of the mHAQ with other outcome measures (McCarthy et al., 2014) 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 (McCarthy et al., 2014).

Test-retest reliability of the mHAQ was evaluated by Matteson et al. (Matteson et al., 2012). The ICC was 0.72 but the small sample size prevented the study being rated green (OMERACT, 2019). The percentage minimal detectable change 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 three studies (Leeb et al., 2003; Matteson et al., 2012; McCarthy et al., 2014), which all confirmed that ESR and CRP improved with treatment of PMR. McCarthy et al. found moderate correlation between ESR/CRP and the mHAQ (McCarthy et al., 2014) but these instruments do not measure the same construct.

None of the studies set out hypotheses about expected relationships and all three studies were rated red.
Responsiveness was evaluated in two studies (Kalke, Mukerjee and Dasgupta, 2000; McCarthy et al., 2014) but neither set out hypotheses about magnitude of change a priori. One study (McCarthy et al., 2013) addressed thresholds of meaning for ESR and CRP was rated amber. This study found that CRP was superior to ESR in detecting active disease and disease remission.

**Discussion**

We identified all of the outcome measures and instruments used to date in studies of PMR and categorised 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 five 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 one study of test-retest reliability which achieved amber and there was one 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 ten randomised controlled trials. Almost all had sample sizes of less 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 four PMR core domains, followed by pain and stiffness. Physical function was least often measured. This contrasts with findings from qualitative studies where patients with PMR have highlighted disability...
and stiffness as having significant impact on their quality of life (Mackie et al., 2015; Twohig et al., 2015).

Pain was the most commonly assessed patient-reported outcome with a VAS being the most frequently used measurement instrument. However, as noted in previous reviews (Duarte et al., 2015; Huang and Castrejon, 2016), 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 (Halls et al., 2014, 2017; Mackie et al., 2015). 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 which 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 (Mackie et al., 2015; Twohig et al., 2015). 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 its 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 rather than ESR. 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 (Cantini et al., 2000).

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 (Helliwell et al., 2016). These include fatigue, psychological impact 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.

Strengths and limitations

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 including two papers (arising from one study) by McCarthy et al., in which one participant out of 60 had biopsy-proven GCA as well as PMR (McCarthy et al., 2013, 2014). This decision was made by the team because there were so few studies on measurement properties of instruments in PMR that these two papers contributed substantially to the available data and it was felt that there was minimal risk of bias from one participant having a dual diagnosis.
Assessment of risk of bias of included studies added value in this review as it had not been
done previously. 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 standardise secondary outcomes
across future trials.

Conclusions
Measurement of outcomes in studies of PMR lacks consistency. The critical patient-centred
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 re-examining individual patient data to determine whether
the selected instruments are suitable for a Core Outcome Measurement Set for PMR.

Footnotes
*The PMR-AS is defined as

\[ \text{CRP} + \text{MST} \times 0.1 + \text{VAS}_{\text{pain}} + \text{VAS}_{\text{physician}} + \text{EUL} \times 0.3 \]

(where CRP is C-reactive protein (mg/dL), MST is morning stiffness duration in minutes, VAS is visual
analogue scale (possible range: 0-10) and EUL is elevation of the upper limbs (possible range: 0-3)).
Acknowledgements

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

References


Table 1: Search strategy for OVID Medline

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>polymyalgia rheumatica.mp.</td>
</tr>
<tr>
<td>2</td>
<td>Polymyalgia Rheumatica/</td>
</tr>
<tr>
<td>3</td>
<td>rheumatic polymyalgia.mp</td>
</tr>
<tr>
<td>4</td>
<td>polymyalgia arteritica.mp.</td>
</tr>
<tr>
<td>5</td>
<td>forestier certonciny syndrome.mp.</td>
</tr>
<tr>
<td>6</td>
<td>rheumatic myalgia.mp.</td>
</tr>
<tr>
<td>7</td>
<td>rhizomelic pseudopolyarthritis.mp.</td>
</tr>
<tr>
<td>8</td>
<td>polymyalgi*.mp.</td>
</tr>
<tr>
<td>9</td>
<td>senile gout.mp.</td>
</tr>
<tr>
<td>10</td>
<td>1 -9 combined with OR</td>
</tr>
</tbody>
</table>
Table 2: Quality criteria for each measurement property, taken from the COSMIN-OMERACT Good Methods Checklist (27)

<table>
<thead>
<tr>
<th>Measurement Property</th>
<th>Quality criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Construct validity (hypothesis testing)</td>
<td>Clear description given of the construct measured by the comparator instrument</td>
</tr>
<tr>
<td></td>
<td>Measurement properties of the comparator instrument described and adequate</td>
</tr>
<tr>
<td></td>
<td>Design and statistical methods adequate for the hypothesis to be tested</td>
</tr>
<tr>
<td></td>
<td>Otherwise free of any important flaws</td>
</tr>
<tr>
<td>Test re-test reliability</td>
<td>Patients stable in the interim period</td>
</tr>
<tr>
<td></td>
<td>Time interval appropriate</td>
</tr>
<tr>
<td></td>
<td>Test conditions similar for the measurements</td>
</tr>
<tr>
<td></td>
<td>Correct statistic used (intra-class correlation coefficient for continuous data, Kappa for dichotomous / ordinal / nominal scores)</td>
</tr>
<tr>
<td></td>
<td>Otherwise free of important flaws</td>
</tr>
<tr>
<td>Responsiveness (longitudinal construct validity)</td>
<td>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</td>
</tr>
<tr>
<td></td>
<td>Measurement properties of the comparator standard described and adequate</td>
</tr>
<tr>
<td></td>
<td>Statistical methods appropriate for the testing situations:</td>
</tr>
<tr>
<td></td>
<td>• For comparison to gold standard – ROC, AUC, predicative values, sensitivity and specificity, correlation of change with external anchor</td>
</tr>
<tr>
<td></td>
<td>• For constructs – effect size, standardised response mean, correlation</td>
</tr>
<tr>
<td></td>
<td>Otherwise free of important flaws</td>
</tr>
<tr>
<td>Clinical trial discrimination</td>
<td>Time interval between testing stated and appropriate</td>
</tr>
<tr>
<td></td>
<td>A proportion of people were expected to change in one or both groups</td>
</tr>
<tr>
<td></td>
<td>A priori hypotheses stated regarding the anticipated mean differences in change scores between sub-groups (positive, negative or no change expected)</td>
</tr>
<tr>
<td></td>
<td>Statistical methods adequate for the hypotheses tested (relative efficiencies, pooled treatment effect sizes, standardised mean differences)</td>
</tr>
<tr>
<td></td>
<td>Otherwise free of important flaws</td>
</tr>
<tr>
<td>Thresholds of meaning</td>
<td>Patient group similar to target population</td>
</tr>
<tr>
<td></td>
<td>Criterion (external anchor, benchmarks, comparable population) selected in a credible manner</td>
</tr>
<tr>
<td></td>
<td>Analysis done separately for improvement and deterioration or only in direction anticipated in the target application</td>
</tr>
</tbody>
</table>
Multiple criteria used and results triangulated
Analysis includes either a Youden index threshold from ROC or another cut off on a ROC approach. If a threshold approach was used, was it tested for diagnostic utility (sensitivity and specificity)? Otherwise free of any flaws

ROC = receiver operating characteristic curve
AUC = area under the curve
Table 3: Summary of outcomes measured by domain (OMERACT core set domains in bold)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Number of studies assessing this domain</th>
<th>Most frequent instrument used (number of studies)</th>
<th>Other instruments used (number of studies)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laboratory markers of inflammation</td>
<td>43 / 46 (93%)</td>
<td>ESR / CRP (42)</td>
<td>IL-6 (10)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fibrinogen (6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TNF-alpha (1)</td>
</tr>
<tr>
<td>Pain</td>
<td>32 / 46 (70%)</td>
<td>VAS (29)</td>
<td>NRS (2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Physician assessment of pain (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Pain site manikins (2)</td>
</tr>
<tr>
<td>Stiffness</td>
<td>28 / 46 (63%)</td>
<td>Morning stiffness duration in minutes (26)</td>
<td>Stiffness severity VAS / NRS (4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Physician assessment of stiffness (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Stiffness site manikins (2)</td>
</tr>
<tr>
<td>Physical function</td>
<td>22 / 46 (48%)</td>
<td>Elevation of upper limbs on 0-3 scale (13)</td>
<td>HAQ (12)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SF-36 physical component (36) (3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>American Rheumatism Association functional class assessment (37) (1)</td>
</tr>
<tr>
<td>Global assessment / disease activity</td>
<td>21 / 46 (46%)</td>
<td>PMR-AS (13)</td>
<td>Physician global assessment (6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Patient global assessment (9)</td>
</tr>
<tr>
<td>Imaging</td>
<td>9 / 46 (2%)</td>
<td>Ultrasound (6)</td>
<td>MRI (3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>FDG PET-CT (2)</td>
</tr>
<tr>
<td>Other:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical examination, presence of synovitis, fever or weight loss</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of relapses, duration of treatment or cumulative steroid dose</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other blood parameters (e.g. FBC, HbA1c, ACTH / cortisol)</td>
<td>17</td>
<td>NRS (1)</td>
<td>Time to onset of fatigue for daily chores (1)</td>
</tr>
<tr>
<td>----------------------------------------------------------</td>
<td>----</td>
<td>---------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td>Fatigue</td>
<td>6</td>
<td>VAS (4)</td>
<td>Back to normal question (1)</td>
</tr>
<tr>
<td>Health status</td>
<td>5</td>
<td>Unspecified questionnaire / VAS (4)</td>
<td></td>
</tr>
<tr>
<td>Mood / anxiety</td>
<td>1</td>
<td>Generalised Anxiety Disorder-7 (GAD-7) (38) (1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Patient Health Questionnaire-8 (PHQ-8) (39) (1)</td>
<td></td>
</tr>
</tbody>
</table>

VAS = visual analogue scale  
NRS = numeric rating scale
<table>
<thead>
<tr>
<th>Instrument</th>
<th>Measurement property</th>
<th>Studies</th>
<th>Quality assessment</th>
<th>Findings</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain VAS</td>
<td>Construct validity</td>
<td>Leeb 2003 (22)</td>
<td>Comparison made to pre-treatment 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.</td>
<td>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.</td>
<td>Red</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Matteson 2012 (23)</td>
<td>Comparison made to pre-treatment levels No a priori hypotheses about magnitude of change or correlation with other instruments stated.</td>
<td>Statistically significant improvement between baseline and W1 and W1 and W4 but not between W4 and W26.</td>
<td>Red</td>
</tr>
<tr>
<td>Responsiveness</td>
<td>McCarthy 2014 (25)**</td>
<td>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.</td>
<td>SRM = 0.89 ESS = 0.96</td>
<td>Red</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kalke 2000 (24)</td>
<td>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.</td>
<td>SRM = 1.7</td>
<td>Red</td>
</tr>
<tr>
<td>Test-retest reliability</td>
<td>Matteson 2012 (23)</td>
<td>Small sample size, n=14 Patients were stable in the interim time period; the time period was appropriate and test</td>
<td>Global pain ICC = 0.82</td>
<td>Amber</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Study Details</td>
<td>Findings</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Thresholds of meaning</strong></td>
<td><strong>Matteson 2012 (23)</strong></td>
<td>Patient group is sufficiently similar to target population. Not enough information on methods given. No attempt to calculate minimally important difference to patients. SDD and % MDC = 28.9.</td>
<td>Red</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Duration of morning stiffness</strong></td>
<td><strong>Leeb 2003 (22)</strong></td>
<td>Comparison made to pre-treatment levels. No <em>a priori</em> hypotheses about magnitude of change or strength of correlation with other instruments stated. Highly significant improvement at W24 compared to baseline.</td>
<td>Red</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Matteson 2012 (23)</strong></td>
<td>Comparison made to pre-treatment levels. No <em>a priori</em> 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.</td>
<td>Red</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Responsiveness</strong></td>
<td><strong>McCarthy 2014 (25)</strong></td>
<td>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</td>
<td>Red</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Kalk 2000 (24)</strong></td>
<td>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</td>
<td>Red</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Test-retest reliability</strong></td>
<td><strong>Matteson 2012 (23)</strong></td>
<td>Small sample size, n=14. Patients were stable in the interim time period; Statistical methods were appropriate (ICC)</td>
<td>Red</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Methodology</td>
<td>Findings</td>
<td>Rating</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------</td>
<td>--------</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Thresholds of meaning</strong></td>
<td>Matteson 2012 (23)</td>
<td>Patient group is sufficiently similar to target population</td>
<td>SDD = 231 %MDC = 16.1</td>
<td>Red</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not enough information on methods given.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>No attempt to calculate minimally important difference to patients</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>HAQ-DI</strong></td>
<td>Kalke 2000 (24)</td>
<td>Small sample size, n = 18</td>
<td>Significant improvement in HAQ score between pre- and post-treatment measurements</td>
<td>Red</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>No clear description of the construct measured by the comparator instrument (not measures of function).</td>
<td>Linear regression coefficient with duration MS, pain VAS and CRP was 0.66, 0.72 and 0.63 respectively</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>No <em>a priori</em> hypotheses about magnitude of change or correlation with other instruments stated.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Responsiveness</strong></td>
<td>Kalke 2000 (24)</td>
<td>Small sample size, n = 18</td>
<td>SRM = 3</td>
<td>Red</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Situation of change clear – newly diagnosed, started on treatment.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Statistical methods are appropriate but no hypotheses about direction of change or strength of correlation between instruments were made.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>mHAQ</strong></td>
<td>Matteson 2012 (23)</td>
<td>Each instrument was compared to its pre-treatment levels</td>
<td>Statistically significant improvement at all measurement time points</td>
<td>Red</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>a priori</em> hypotheses about magnitude of change or correlation with other instruments stated.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>McCarthy 2014 (25)</td>
<td>Each instrument was compared to its pre-treatment levels.</td>
<td>Statistically significant improvement between W1 and W6 in the active group.</td>
<td>Red</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparator measures were not evaluating the same construct.</td>
<td>Correlation coefficients between mHAQ and PMR-AS, ESR and CRP were 0.68, 0.45 and</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **SDD** = Standardized Difference in Differences
- **%MDC** = Percentage Minimal Detectable Change
- **SRM** = Standardized Response Mean
<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Responsiveness</strong></td>
<td>McCarthy 2014 (25)</td>
<td>No <em>a priori</em> hypotheses about magnitude of change or correlation with other instruments stated.</td>
<td>0.39 respectively</td>
<td>Red</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Situation of change clear in active group – newly diagnosed, started on treatment.</td>
<td>SRM = 1.36</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PMR-AS used as gold standard for assessment of remission – accepted as a validated measure.</td>
<td>ESS = 1.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Statistical methods were appropriate but no hypotheses about magnitude of change were made.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Test-retest reliability</strong></td>
<td>Matteson 2012 (23)</td>
<td>Small sample size, n=14</td>
<td>ICC = 0.72</td>
<td>Amber</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Patients were stable in the interim time period; the time period was appropriate and test conditions were stable.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Statistical methods were appropriate (ICC)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Thresholds of meaning</strong></td>
<td>Matteson 2012 (23)</td>
<td>Patient group is sufficiently similar to target population</td>
<td>SDD = 0.78</td>
<td>Red</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not enough information on methods given.</td>
<td>% MDC = 25.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>No attempt to calculate minimally important difference to patients</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ESR / CRP</strong></td>
<td>Leeb 2003 (22)</td>
<td>Each instrument was compared to its pre-treatment levels and correlation between VAS pain and ESR / CRP was assessed.</td>
<td>Highly significant improvement at W24 compared to baseline.</td>
<td>Red</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>No a priori</em> hypotheses about magnitude of change or strength of correlation with other instruments stated.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matteson 2012 (23)</td>
<td>Each instrument was compared to its pre-treatment levels</td>
<td>Statistically significant improvement between baseline and W1 and W1 and W4 but not between W4 and W26.</td>
<td>Red</td>
</tr>
</tbody>
</table>
| Responsiveness                                                                 | McCarthy 2014 (25)                                                                 | Statistically significant improvement from W1 to W6 in the active group  
Correlation coefficient between mHAQ and ESR / CRP = 0.45 / 0.39 | Red |
|--------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|--------------------------------------------------------------------------|-----|
| 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 / ES = 1.05 / 1.14 | | Red |

<table>
<thead>
<tr>
<th>Kalke 2000 (24)</th>
<th>Small study, n=18</th>
<th>CRP SRM 1.6</th>
<th>Red</th>
</tr>
</thead>
</table>
| Situation of change clear – newly diagnosed, started on treatment.  
Statistical methods are appropriate but no hypotheses about magnitude of change were made. |                                                                                   |             |     |

| Thresholds of meaning                                                          | McCarthy 2013 (28)**                                                             | Ability of ESR >40mm/h / CRP >6mg/l to detect active disease:  
Values for ESR: sensitivity 92%, specificity 66%, PPV 0.72, Likelihood ratio 2.8.  
Values for CRP: sensitivity 100%, specificity 70%, PPV 0.77, Likelihood ratio 3.33  
Ability of ESR <20mm/h / CRP <6mg/l to detect disease remission: | Amber |
|--------------------------------------------------------------------------------|----------------------------------------------------------------------------------|-------------|-----|
| Appropriate patient group.  
Criteria for assessment of disease activity and definition of remission satisfactory.  
Thresholds for ESR and CRP cut offs justified from the literature.  
Statistical methods satisfactory but did not use multiple methods to triangulate findings. |                                                                                   |             |     |
| Values for ESR: sensitivity 43%, specificity 75%, PPV 0.87, Likelihood ratio 1.7. |
| Values for CRP: sensitivity 58%, specificity 67%, PPV 0.88, Likelihood ratio 2.04. |
Table 5: Summary of quality of evidence on measurement properties of outcome measurement instruments in PMR

<table>
<thead>
<tr>
<th></th>
<th>Evaluation of evidence supporting use of this instrument in PMR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N/A = not evaluated, - = evaluated but insufficient evidence to support use, + = evaluated and some evidence to support use, ++ = good evidence to support use in clinical studies</td>
</tr>
<tr>
<td></td>
<td>Construct validity</td>
</tr>
<tr>
<td>Pain VAS</td>
<td>-</td>
</tr>
<tr>
<td>Stiffness VAS</td>
<td>N/A</td>
</tr>
<tr>
<td>Duration of morning stiffness</td>
<td>-</td>
</tr>
<tr>
<td>HAQ-DI</td>
<td>-</td>
</tr>
<tr>
<td>mHAQ</td>
<td>-</td>
</tr>
<tr>
<td>ESR and CRP</td>
<td>-</td>
</tr>
</tbody>
</table>
16222 references identified from 5 databases

3028 duplicates removed

13194 studies underwent title review

12585 studies excluded

609 studies underwent abstract review

519 studies excluded

90 studies assessed for full-text eligibility

44 studies excluded
16 Patient group not exclusively PMR
13 Outcomes not directly of PMR*
10 Full text not available / not available in includable language
5 Not empirical data

46 studies included

*Studies in this group included those that examined outcomes such as rates of cardiovascular disease or fractures in PMR or that analysed biochemical markers involved in the pathogenesis of the disease