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[2,3], Sara Muller[1], Christian Mallen[1,9], Caroline Mitchell[4], Samantha Hider[1,9], Catherine Hill[5,6], Beverley Shea[7], Sarah Mackie[8]

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

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. h.j.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 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

$CRP + MSTx0.1 + VAS_{pain} + VAS_{physician} + EUL_{0-3}$

(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

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. 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. The OMERACT core domain set for outcome measures for clinical trials in polymyalgia rheumatica. J Rheumatol 2017;44:1515–21.
6. Duarte C, Ferreira RJ d. O, Mackie SL, Kirwan JR, Pereira da Silva JA. 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;46:1021-1027.

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. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343.
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;37:1639–1644.
11. Salvarani C, Cantini F, Niccoli L, Catanoso MG, I PM, Pulsatelli LIA, 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;46:1360-1364.

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 25:130–6.

27. OMERACT. OMERACT Handbook Instrument Selection Chapter 5 Mar 2019. In: OMERACT handbook 2019. Available from: https://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. Rheumatol (United Kingdom) 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;10:1–13.

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 JEJ, 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.

40. Alvarez-Rodriguez L, Lopez-Hoyos M, Mata C, Jose Marin M, Calvo-Alen J et al. Circulating cytokines in active polymyalgia rheumatica. Ann Rheum Dis 2010;69:263–269.

41. Benucci M, Olivito B, Manfredi M, Meacci F, Infantino M et al. Polymyalgia rheumatica: inflammation suppression with low dose of methylprednisolone or modified-release prednisone. Eur Rev Med Pharmacol Sci 2015;19:745-751

42. Binard A, De Bandt M, Berthelot JM, Saraux A. Performance of the polymyalgia rheumatica activity score for diagnosing disease flares. Arthritis Rheum 2008;59:263-269

43. Boiardi L, Casalli B, Farnetti E, Pipitone N, Nicoli D et al. Relationship between interleukin 6 promotor polymorphism at position -174, IL-6 serum levels and risk of relapse / recurrence in polymyalgia rheumatica. J Rheumatol 2006;33:703-708

44. Caporali R, Cimmino G, Ferraccioli G, Gerli R, Klersy C et al. Prednisolone plus methotrexate for polymyalgia rheumatica: A randomised, double blind, placebo-controlled trial. Ann intern med 2004;141:493-500

45. Catanoso M, Macchioni P, Boiardi L, Pipitone N, Salvarani C. Treatment of refractory polymyalgia rheumatica with etanercept: an open pilot study. Arthritis Rheum 2007;57:1514-1519

46. Cimmino M, Parodi M, Caporali R, Montecucco C. Is the course of steroid-treated polymyalgia rheumatica more severe in women? Ann NY Acad Sci 2006;1069:315-321

47. Cimmino M, Parodi M, Motencucco C, Caporali R. The correct prednisolone starting dose in polymyalgia rheumatica is related to body weight but not to disease severity. BMC Musculoskeletal Dis 2011;12:1-4

48. Cimmino M, Salvarani C, Macchioni P, Gerli R, Bartolonni Bocci E et al. Long-term follow-up of polymyalgia rheumatica patients treated with methotrexate and steroids. Clin Exp Rheum 2008;26:395-400

49. Cleuziou C, Binard A, De Bandt M, Berthelot JM, Saraux A. Contribution of the polymyalgia rheumatica activity score to glucocorticoid dosage adjustment in everyday practice. J Rheum 2012;39:310-313

50. Corrao S, Pistone R, Scaglione R, Colomba D, Calvo L et al. Fast recovery with etanercept in patients affected by polymyalgia rheumatica and decompensated diabetes: a case-series study. Clin Rheum 2009;28:89-92

51. Cutolo M, Hopp M, Liebscher S, Dasgupta B, Buttgeriet F. Modified-release prednisone for polymyalgia rheumatica: A multicentre randomised, active-controlled, double-blind, parallel-group study. RMD Open 2017;3:1-7

52. Dasgupta B, Dolan A, Panayi G, Fernandes L. An initially double-blind controlled 96 week trial of depot methylprednisolone against oral prednisolone in the treatment of polymyalgia rheumatica. Br J Rheum 1998;37:189-195.

53. Dasgupta B, Gray J, Fernandes L, Olliff C. Treatment of polymyalgia rheumatica with intramuscular injections of depot methylprednisolone. Ann Rheum Dis 1991;50:942-945

54. Devauchelle-Pensec V, Berthelot J, Cornec D, Renaudineau Y, Marhadour T et al. Efficacy of first-line tocilizumab therapy in early polymyalgia rheumatica: A prospective longitudinal study. Ann Rheum Dis 2016;75:1506-1510

55. Diamantopoulos AP, Hetland H, Myklebust G. Leflunomide as a corticosteroid-sparing agent in giant cell arteritis and polymyalgia rheumatica: a case series. La Presse Medicale 2013;42:726-727

56. Feinberg H, Sherman J, Shrepferman C, Dietzen C, Feinberg GD. The use of methotrexate in polymyalgia rheumatica. J Rheum 1996;23:1550-1552

57. Ferraccioli G, Salaffi F, Salvatore D, Casatta L, Bartoli E. Methotrexate in PMR: Preliminary results of an open randomised study. J Rheum 1996;23:624-628

58. Hutchings A, Hollywood J, Lamping D, Pease C, Chakravarty K et al. Clinical outcomes, quality of life and diagnostic uncertainty in the first year of polymyalgia rheumatica. Arthritis Rheum 2007;57:803-809

59. Izumi K, Kuda H, Ushikubo M, Kuwana M, Takeuchi T et al. Tocilizumab is effective against polymyalgia rheumatica: Experience in 13 intractable cases. RMD Open 2015 ;1:e000162

60. Jimenez-Palop M, Naredo E, Humbrado L, Medina J, Uson J et al. Ultransongraphic monitoring of response to therapy in polymyalgia rheumatica. Ann Rheum Dis 2010;69:879-882

61. Krogsgard M, Lund B, Johnsson B. A longterm prospective study of the equipotency between deflazacort and prednisolone in the treatment of patients with polymyalgia rheumatica. J Rheum 1995;22:1660-1662

62. Lally L, Forbess L, Hatzis C, Spiera R. Brief report: A prospective open-label phase IIa trial of tocilizumab in the treatment of polymyalgia rheumatica. Arthritis Rheum 2016;68:2550-2554

63. Leeb B, Rintelen B, Sautner J, Fassl C, Bird H. The polymyalgia rheumatica activity score in daily use: proposal for a definition of remission. Arthritis Rheum 2007;57:810-815

64. Littman B, Bjarnason D, Bryant G, Engelbrecht J, Cohen M et al. Steroid sparing activity of tenidap in patients with polymyalgia rheumatica: a multicenter double blind randomized placebo controlled study. J Rheum;22:1097-1103.

65. Macchioni P, Catanoso M, Pipitone N, Boiardi L, Salvarani C. Longitudinal examination with shoulder ultrasound of patients with polymyalgia rheumatica. Rheumatol 2009;48:1566-1569

66. Mackie S, Pease C, Fukuba E, Harris E, Emery P et al. Whole-body MRI of patients with polymyalgia rheumatica identifies a distinct subset with complete patient-reported response to glucocorticoids, Ann Rheum Dis 2015;74:2188-2192

67. Migliore A, Massafra U, Carloni E, Padalino C, Martin Martin S et al. TNF-alpha blockade induce clinical remission in patients affected by polymyalgia rheumatica associated to diabetes mellitus and/or osteoporosis: a seven cases report. Eur Rev Med Pharmacol Sci 2005;9:373-378

68. Pallard-Novello X, Querellou S, Gouillou M, Saraux A, Marhadour T et al. Value of 18F-FDG PET/CT for therapeutic assessment of patients with polymyalgia rheumatica receiving tocilizumab as first-line treatment. Eur J Nucl Med Mol Imaging 2016;43:773–779

69. Pulsatelli L, Boiardi L, Pignotti E, Dolzani P, Silvestri T et al. Serum interleukin-6 receptor in polymyalgia rheumtica: a potential marker of relapse / recurrence risk. Arth care and res 2008;59:1147-1154

70. Pulsatelli L, Peri G, Macchioni P, Boiardi L, Salvarani C et al. Serum levels of long pentraxin PTX3 in patients with polymyalgia rheumatica. Clin Exp Rheum 2010;28:756-758

71. Salvarani C, Cantini F, Niccoli L, Macchioni P, Consonni D et al. Acute-phase reactants and the risk of relapse / recurrence in polymyalgia rheumatica: a prospective follow up study. Arthritis Rheum 2005;53:33-38

72. Salvarani C, Cantini F, Olivieri I, Barozzi L, Macchioni L et al. Corticosteroid injections in polymyalgia rheumatica: a double-blind, prospective, randomized, placebo controlled study. J Rheum 2000;27:1470-1476

73. Salvarani C, Macchioni P, Manzini C, Paolazzi G, Trotta A et al. Infliximab plus Prednisone or Placebo plus Prednisone for the Initial Treatment of Polymyalgia Rheumatica. Ann Int Med 2007;146:631-640

74. Viapiana O, Gatti D, Troplini S, Adami S, Fracassi E et al. Prednisone compared to methylprednisolone in the polymyalgia rheumatica treatment. Rheumatol Int 2015;35:735-739

75. Weyand C, Fulbright J, Evans J, Hunder G, Goronzy J. Corticosteroid requirements in polymyalgia rheumatica. Arch Int Med 1999;159:577-584

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

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 re-test reliability	Patients stable in the interim period Time interval appropriate Test conditions similar for the measurements Correct statistic used (intra-class correlation coefficient for continuous data, Kappa 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, standardised response mean, 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 one or both groups *A priori* hypotheses stated regarding the anticipated mean differences in change scores between sub-groups (positive, negative or no change expected) Statistical methods adequate for the hypotheses tested (relative efficiencies, pooled treatment effect sizes, standardised 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 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)

Domain	Number of studies assessing this domain	Most frequent instrument used (number of studies)	Other instruments used (number of studies)
Laboratory markers of inflammation	43 / 46 (93%)	ESR / CRP (42)	IL-6 (10) Fibrinogen (6) TNF-alpha (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)	Physician global assessment (6) Patient global assessment (9)
Imaging	9 / 46 (2%)	Ultrasound (6)	MRI (3) FDG PET-CT (2)
Other: Physical examination, presence of synovitis, fever or weight loss Number of relapses, duration of treatment or cumulative steroid dose	 10 7		

Other blood parameters (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	Generalised Anxiety Disorder-7 (GAD-7) (38) (1) Patient Health Questionnaire-8 (PHQ-8) (39) (1)	

VAS = visual analogue scale

NRS = numeric rating scale

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

Instrument	Measurement property	Studies	Quality assessment	Findings	Rating
Pain VAS	Construct validity	Leeb 2003 (22)	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.	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 pre-treatment 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	Global pain ICC = 0.82	Amber

			conditions were stable. Statistical methods were appropriate (ICC)		
	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	SDD and % MDC = 28.9.	Red
Duration of morning stiffness	Construct validity	Leeb 2003 (22)	Comparison made to pre-treatment 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 pre-treatment 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;	ICC 0.11	Red

			the time period was appropriate and test conditions were stable. Statistical methods were appropriate (ICC)		
	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	SDD = 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 post-treatment measurements Linear regression coefficient with duration MS, pain VAS and CRP was 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 pre-treatment levels No *a priori* hypotheses about magnitude of change or correlation with other instruments stated.	Statistically significant improvement at all measurement time points	Red
		McCarthy 2014 (25)	Each instrument was compared to its pre-treatment levels. Comparator measures were not evaluating the same construct.	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	Red

			No *a priori* hypotheses about magnitude of change or correlation with other instruments stated.	0.39 respectively	
	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	SDD = 0.78 % MDC = 25.9	Red
ESR / CRP	Construct validity	Leeb 2003 (22)	Each instrument was compared to its pre-treatment 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 pre-treatment levels No *a priori* hypotheses about magnitude of	Statistically significant improvement between baseline and W1 and W1 and W4 but not between W4 and W26.	Red

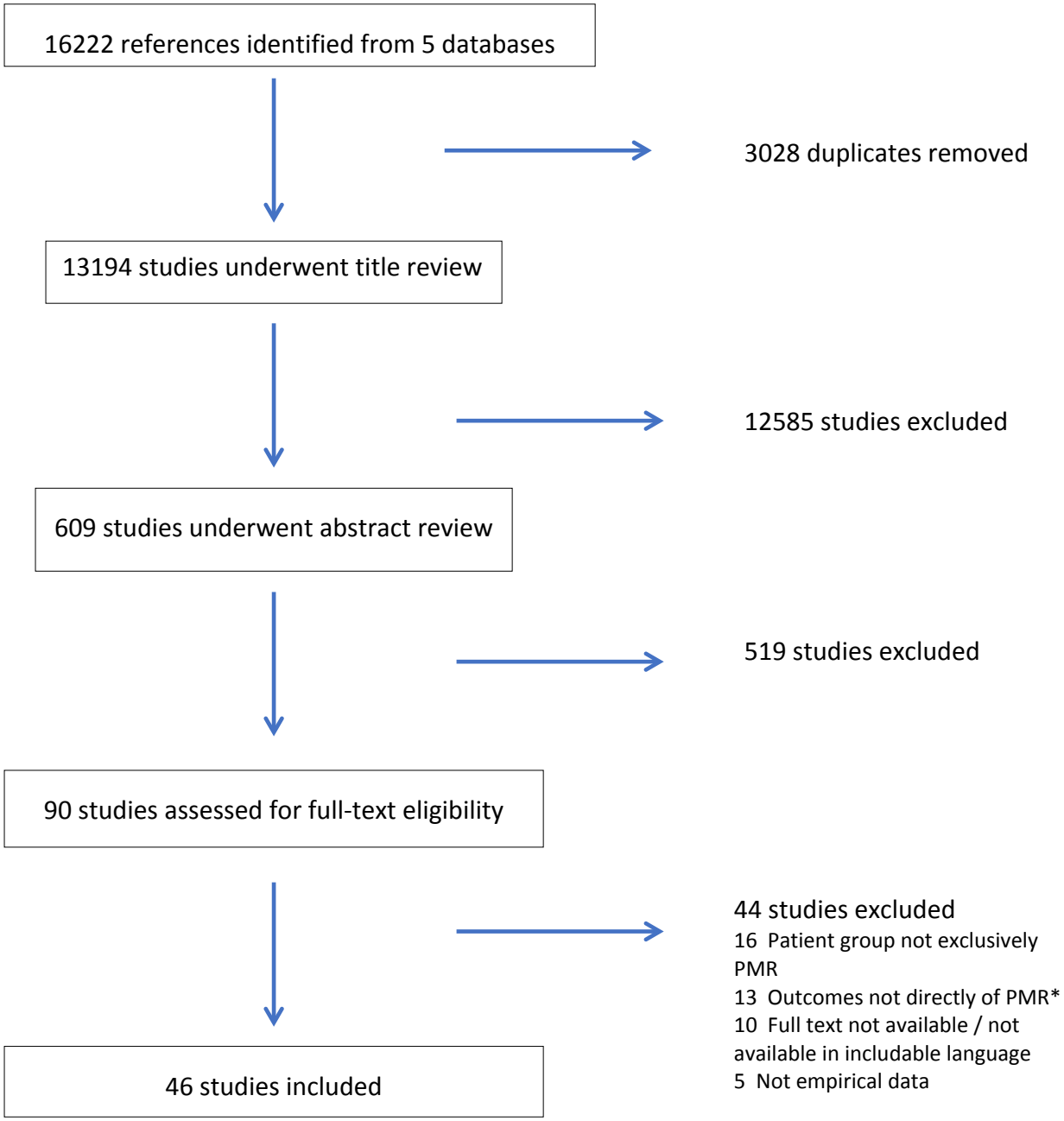
			change or correlation with other instruments stated.		
		McCarthy 2014 (25)	Each instrument was compared to its pre-treatment levels. Comparator instrument for correlation was the mHAQ which measures a different construct. No explicit *a priori* hypotheses about magnitude of change or correlation with 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 / ES = 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 cut offs justified from the literature. Statistical methods satisfactory but did not use multiple methods to triangulate findings.	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

				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

	Evaluation of evidence supporting use of this instrument in PMR **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			
	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**	**-**	**+**

Figure 1: PRISMA diagram of study selection process



*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