Keele
UNIVERSITY

This work is protected by copyright and other intellectual property rights and duplication or sale of all or part is not permitted, except that material may be duplicated by you for research, private study, criticism/review or educational purposes. Electronic or print copies are for your own personal, non-commercial use and shall not be passed to any other individual. No quotation may be published without proper acknowledgement. For any other use, or to quote extensively from the work, permission must be obtained from the copyright holder/s.

Development and psychometric evaluation of a patient reported outcome measure for polymyalgia rheumatica

Helen Jane Twohig

Submission for a Doctorate of Philosophy

March 2022

Keele University

Acknowledgements

This thesis would not have been possible without the fantastic support I have received from my supervisory team, Dr Sara Muller, Dr Caroline Mitchell and Professor Christian Mallen. Throughout my fellowship they have given me encouragement and freedom to pursue my ideas, whilst providing wise guidance and rapid feedback at all stages. Dr Sara Muller deserves special gratitude for her enduring patience in guiding me through the statistical processes necessary for the analyses – thank you.

I would also like to thank the clinicians and researchers who collaborated on the formative research studies that led to the programme of work constituting this PhD. These include Professor Georgina Jones, Dr Ade Adebajo and Professor Nigel Mathers as well as Dr Caroline Mitchell and Professor Christian Mallen. Dr Mitchell and Professor Mallen have supported and encouraged me from the outset, and it is thanks to them that I reached the stage of applying for a Wellcome Trust Primary Care Doctoral Fellowship. I am grateful to them for taking on the role of supervising my PhD and to the Wellcome Trust for my fellowship award.

The OMERACT PMR-Special Interest Group have been instrumental to my understanding of the broader context of the field of outcome measures and research into PMR, and regular meetings with the group have kept my focus on the patient perspective. Clinical academics from the SIG participated in the systematic review reported in Chapters 4 and 5, assuring the protocol, providing second reviews of data extraction and risk of bias assessments as well as co-authoring the resultant paper. My thanks especially to Dr Claire Owen and Dr Sarah Mackie for their support with this process.

The two cross-sectional surveys carried out for this PhD were supported by the West Midlands NIHR Clinical Research Network and I would particularly like to thank Samantha Hunt for guiding me through the study set up and recruitment process. The second of these studies was interrupted by the COVID-19 pandemic. I am hugely grateful to Sarah Lawton and the team at Keele, who enabled the study to resume as soon as it was possible and went out of their way to support me by collecting the replies from the department and entering the data during the time that I was unable to work on campus.

Finally, I want to thank my husband Henry, who encourages me and supports me in everything I do. My sons, Thomas and Matthew, have grown from babies to schoolboys during the time I have been doing my PhD and they and Henry continually provide humour, distraction and perspective. Juggling clinical work, home-schooling and this PhD during the pandemic was certainly a challenge, but with the support of everyone mentioned, we somehow did it.

Abstract

Background

Polymyalgia rheumatica (PMR) causes pain, stiffness and disability in older adults. Measuring the impact of the condition from the patient's perspective is vital to high-quality research and patient-centred care, yet there are no validated patient reported outcome measures (PROMs) for PMR. The aims of this PhD are, i) to identify outcome measures and instruments used in clinical studies of PMR and evaluate the evidence supporting their use in the condition, and ii) to develop and evaluate a PMR-specific PROM.

Methods

Outcomes and instruments used in PMR research were systematically identified and categorised. Studies on their measurement properties were appraised.
Two primary research studies were then undertaken:
1) 256 people with PMR completed a draft PROM providing data for item reduction, formation of dimension structure and scoring system development.
2) 179 people with PMR completed the PROM at two time points along with comparator questionnaires and anchor questions. Test-retest reliability, construct validity, and responsiveness were evaluated.

Results

The most frequent outcomes (and instruments) identified in the literature were: markers of systemic inflammation (ESR/CRP), pain (visual analogue scale), stiffness (duration) and physical function (elevation of upper limbs). No instruments had high-quality evidence to support their use in PMR.
Results from the first study led to the development of a PROM, the PMR-impact scale (PMR-IS), comprising symptoms, function, emotional and psychological well-being and steroid side-effects domains.
Construct validity and test-retest reliability were good for each domain of the PMR-IS. It was responsive to improvement in the condition but there was insufficient evidence to determine its ability to detect flares.

Conclusions

Current outcome measures used in PMR are not adequate. The PMR-IS provides a real opportunity to improve patient-centred research and care, but further work is needed to more fully establish its responsiveness and interpretability parameters.

Table of Contents

List of Included Tables

List of Included Figures

Chapter 1: Polymyalgia Rheumatica

1.1 Introduction

Polymyalgia rheumatica (PMR) is an inflammatory musculoskeletal condition affecting older adults. It causes pain and stiffness, mainly around the shoulder and hip girdles, and significant associated functional impairment (González-Gay et al., 2017). It is a heterogenous condition, with features that overlap with many other conditions and has a variable disease course (Michet & Matteson, 2017). The mainstay of treatment is glucocorticoid medication, which usually has to be taken for around two years and can cause significant side effects (Dejaco, Singh, Perel, Hutchings, Camellino, Mackie, Matteson, et al., 2015).

PMR is thought to have first been described in 1888 by William Bruce in an article entitled "Senile Rheumatic Gout" in which he outlined five cases of an inflammatory musculoskeletal condition distinct from rheumatoid arthritis or gout (Bruce, 1888). The term 'polymyalgia rheumatica' however, did not appear in the literature until 1957 when it was used by Barber (Barber, 1957). He outlined the key clinical features by which the condition is still recognised today, including the rapid response to treatment with glucocorticoids. A review of the early writings about the condition notes that even in the earliest reports, its diversity and diagnostic difficulties were highlighted (Hunder, 2006). Despite the significant advances in medicine since its first description, there are still many unanswered questions about PMR, making it a challenging condition to experience either as a patient or health care professional.

In this chapter, I will summarise what is currently known about the condition and where the existing uncertainties lie, to set the scene for the research that makes up this thesis.

1.2 Incidence and epidemiology

1.2.1 UK perspective

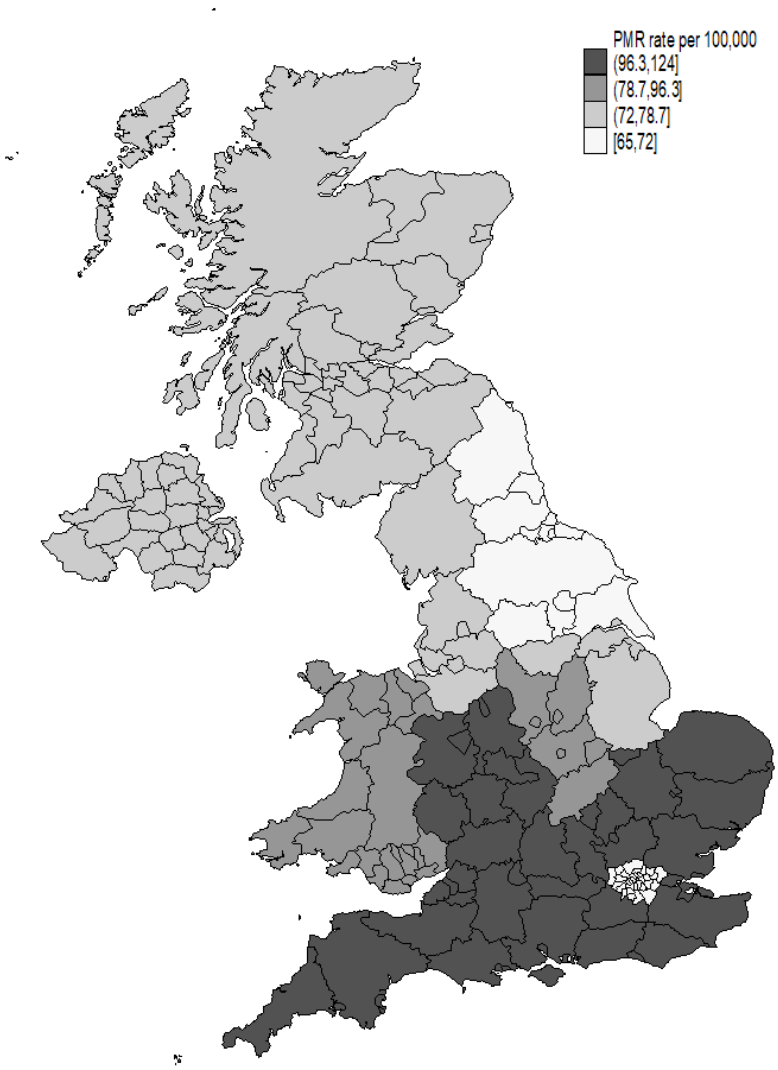
The most recent assessment of the incidence and prevalence of PMR in the UK was carried out by Partington et al. (2018). They used data from the Clinical Practice Research Datalink (CPRD), which contains electronic, coded routine primary healthcare data collected from around 17 million patients. CPRD is representative of the UK population in terms of ethnicity, gender and age. As most patients with PMR in the UK are managed exclusively in primary care (Barraclough et al., 2008; Yates et al., 2016), using a high-quality established database from this setting is likely to give the most accurate estimate of incidence and prevalence. Partington reports that between 1990 and 2016, the incidence of PMR in people aged over 40 years was 95.9 per 100 000 person years. The incidence was higher at older ages; in people over 80 years it was 314.9 per 100 000 person years. Women were 67% more likely to develop PMR than men. The point prevalence of PMR in 2015 amongst people aged over 55 years was 1.7%. There was marked geographical variation with rates of the condition highest in the South-West and lowest in the North-East (Figure 1.1).

The findings from this latest study were consistent with findings from previous UK and international studies in terms of variation in incidence by sex and geographical location. The one previous General Practice Research Database (GPRD, the precursor to CPRD) study was by Smeeth et al. (2006) looking at data from 1990-2001. They found the

overall UK age adjusted incidence rate to be slightly lower at 8.4 per 10 000 person years but showed that the incidence steadily increased from 1990 to 1999. The Partington et al. (2018) study, in contrast, found that incidence rates were stable between 2003 and 2014.

Figure 1.1: Incidence rates of PMR by region, 1990-2015

(reproduced with permission from Partington et al. (2018))



A study by Yates et al (2016) looked at prevalence of PMR in one Norfolk practice with a list size of 13 000 patients. They found the prevalence of GP diagnosed PMR was 2.27% in over 55s. This is higher than that found by Partington et al. but is based on a smaller geographical area in the South East of the UK, which, as the map above shows, is known to have a higher incidence of PMR.

1.2.2 Global perspective

Worldwide, there is considerable variation in the incidence and prevalence of PMR. The highest incidence rates of PMR are in the Northern latitudes, in Scandinavia and people of Northern European descent, with a much lower incidence reported in the Mediterranean basin (Gonzalez-Gay et al., 2009). In Norway the annual incidence rate in people aged over 50 is 113 per 100 000 (Crowson & Matteson, 2017) whereas a rate of 3.15 per 100 000 person years has been reported from Turkey (Pamuk et al., 2009).

In the U.S. the incidence and prevalence of PMR has been studied in a population-based cohort set in Olmstead County, Minnesota. The data are from the Rochester Epidemiology Project (https://rochesterproject.org/), which is a medical records-linkage system established in 1966 to capture health care information for the entire population of Olmsted County, USA to support population-based health research (St Sauver et al., 2012).

Using this database, an inception cohort of all cases of PMR between 1970 and 1999 was identified (Salvarani et al., 1995) and more recently, expanded to include all cases up until 2014 (Raheel et al., 2017). The age and sex-adjusted annual incidence rate of PMR in 2000-2014 was 63.9 (95% CI 57.4, 70.4) per 100 000 population aged 50 years and older,

which was significantly higher than the annual age and sex-adjusted rate found in the same population in 1970-1999 (55.8 per 100 000).

Using the same cohort data, Crowson and Matteson (2017) calculated an estimated point prevalence of PMR on Jan 1st 2015 in the US, finding it to be 701 (95%CI: 651–750) per 100 000 population aged 50 years and older. In line with previous studies, the prevalence was two to three times higher in women and increased with age.

Based on the initial Olmstead County cohort data, the lifetime risk of developing PMR has been calculated to be 2.4% in women and 1.7% in men. In comparison to other inflammatory rheumatic diseases, this makes PMR the second most common inflammatory rheumatic condition over a lifetime (behind rheumatoid arthritis) and it has the highest incidence of any of the inflammatory rheumatic diseases in people over 70 years old (Crowson et al., 2011).

1.3 Aetiology

The aetiology of PMR is uncertain but both environmental and genetic factors are thought to play a role (Gonzalez-Gay et al., 2009; Michet & Matteson, 2017). The increased incidence of PMR in Scandinavian countries and in communities elsewhere with strong Scandinavian ethnic background suggests a genetic link. The related condition of giant cell arteritis (GCA) is known to be associated with several genes in the major histocompatibility complex (HLA-DRB1 alleles) but there are no known conclusive associations between specific genes and PMR (González-Gay et al., 2003).

The observation that there are cyclical peaks and seasonal variation in incidence of PMR led to investigations into an association with infectious agents including *M.pneumoniae*,

parvovirus B19 and *C. pneumoniae,* but attempts to prove this have been inconclusive (González-Gay et al., 2017; Narvaez et al., 2000; Perfetto et al., 2005).

1.4 Link with Giant Cell Arteritis

Giant cell arteritis (GCA) is a large vessel vasculitis that shares some pathophysiological and phenotypical features with PMR, suggesting that the two conditions may be manifestations of the same disease process (Gonzalez-Gay, 2004).

GCA and PMR have long been known to frequently occur together, affect a similar age and sex distribution and both typically demonstrate high acute phase reactants at onset and responsiveness to steroids. Both conditions can cause a polymyalgic syndrome with constitutional symptoms including fever and weight loss, but GCA characteristically causes headache with features of cranial artery inflammation (e.g., temporal artery tenderness and jaw claudication) and visual symptoms, which are not features of PMR.

Polymyalgic symptoms are present in 40-60% of people with biopsy proven GCA and 16-21% of people with PMR may develop GCA, particularly if left untreated (Dejaco et al., 2011).

Recent advances in imaging techniques have enabled studies using vascular ultrasound and 18-fluorine fluorodeoxyglucose PET/CT (18F-FDG PET), which have demonstrated that up to 80% of GCA patients and approximately one-third of patients with PMR have subclinical large vessel inflammation at diagnosis (Blockmans et al., 2007; Schmidt et al., 2008).

GCA is now considered to be a spectrum of disease with variants encompassing cranial arteritis and extracranial large vessel vasculitis (Dejaco, Brouwer, et al., 2017). In their

2017 paper on the spectrum of GCA and PMR, Dejaco, Duftner et al. (2017) assert that '*It is conceivable that PMR is a limited or aborted form of GCA in which overt vasculitis has either not yet started or has been prevented by unclear regulatory mechanisms'.*

1.5 Pathology

PMR is a non-erosive arthropathy in which disease activity in and around the joints is high but permanent joint damage does not occur. Imaging studies of patients with PMR have shown inflammation both in the affected joints (capsular) and in the periarticular structures (extra-capsular), but the precise pathology of the condition remains unclear (Dejaco, Duftner, et al., 2017).

Ultrasound findings typical in PMR include subacromial bursitis, biceps tenosynovitis and glenohumeral synovitis at the shoulders and synovitis and trochanteric bursitis at the hips (Cantini et al., 2001, 2005). A Magnetic Resonance Imaging (MRI) study of 32 Japanese patients with PMR found that all had subacromial and subdeltoid bursitis, whilst 93% had glenohumeral joint synovitis and 57% had biceps tenosynovitis. Similar changes were demonstrated in the knee joints and inflammatory changes in the soft tissues around the joint capsule were also marked (Mori et al., 2007). McGonagle et al. (2001) found that the presence of extracapsular soft tissue oedema distinguished patients with PMR from those with rheumatoid arthritis whereas bursitis, tenosynovitis and joint effusion occurred in either condition.

A whole-body MRI study aimed at identifying clinically relevant sub-groups of PMR, found that 14/22 patients had a pattern of symmetrical extracapsular inflammation around the shoulders and hips and that this pattern was associated with a significantly higher

reported complete response to glucocorticoids (Mackie, Pease, et al., 2015). More recently, newer MRI techniques were used in the TENOR study of tocilizumab in PMR and for the first time, localised myofascial lesions were demonstrated in patients with active PMR, which improved with treatment (Laporte et al., 2019). This finding adds to the growing body of evidence of involvement of musculotendinous structures in PMR and led Owen, Liew et al. (2019) to claim in their associated editorial that "musculotendinous inflammation represents the defining pathology of PMR".

The existence of musculotendinous inflammation in PMR still needs to be confirmed histologically. Synovitis in PMR has been histologically studied however and has been shown to be characterised by vascular proliferation and leucocyte infiltration with predominance of macrophages and CD4+ T cells (Meliconi et al., 1996).

Approximately one-third of people with PMR will have large artery vasculitis on biopsy. Even in those with apparently normal temporal arteries histologically, adventitial dendritic cells can be seen to be in an activated state which leads to the production of inflammatory cytokines such as interleukin-1 (IL-1) and interleukin-6 (IL-6). In GCA, these activated dendritic cells recruit CD4+ T-cells into the arterial walls where the inflammatory mediators cause disruption of the internal elastic lamina and activation of repair mechanisms such as intimal hyperplasia and neoangiogenesis. In those with PMR, even though many of the same inflammatory mediators can be detected, CD4+ T-cells are not recruited into the vascular walls and the inflammation remains subclinical (Salvarani et al., 2008).

The systemic effects of PMR are caused by the circulation of inflammatory cytokines, including IL-1 and IL-6, derived from macrophages (Michet & Matteson, 2017).

1.6 Clinical features

PMR is characterised by pain and stiffness in the shoulder and pelvic girdles. Shoulder pain is the presenting symptom in 70–95% of patients, whereas the hips and neck are less frequently affected (50–70%) (Salvarani et al., 2008). Onset of symptoms is usually fairly rapid, typically occurring over days (González-Gay et al., 2017). Pain and stiffness can be unilateral initially but quickly spread to become bilateral. Examination typically shows painful restriction of active, and sometimes passive, movements of the shoulders, neck and hips but without detectable proximal joint swelling (Salvarani et al., 2008).

Constitutional symptoms including low grade fever, fatigue, low mood and weight loss have been found to be present in 30-40% of patients with PMR (Chuang et al., 1982; Gonzalez-Gay et al., 1998).

Many patients with PMR also have distal musculoskeletal involvement. Studies have identified peripheral arthritis, which is typically asymmetrical, mono- or pauci-articular, transient and not destructive, in 20-45% of people with PMR during the course of their disease (Gran & Myklebust, 2000; Narvaez et al., 2001; Salvarani et al., 1998). Carpal tunnel syndrome and distal extremity swelling with pitting oedema can also be present, although less frequently than peripheral arthritis (Narvaez et al., 2001; Salvarani et al., 1998).

The combination of pain and stiffness and the constitutional symptoms of PMR can all lead to reduced functional ability with associated reduced quality of life (Hutchings et al., 2007; Mackie, Hughes, et al., 2015).

1.6.1 Assessment and diagnosis

PMR is a clinical diagnosis, meaning it is diagnosed on the basis of reported symptoms and clinical signs rather than laboratory tests. It can be a difficult diagnosis to make as several other autoimmune, infectious, endocrine and malignant disorders can present with similar symptoms (Dasgupta et al., 2012). A rapid response to treatment with glucocorticoid medication can contribute to the diagnostic process.

Several classification criteria have been proposed and these are summarised in Table 1.1.

Inflammatory markers are usually raised in PMR and are part of many of the published classification criteria. However, up to 22% of people with PMR have a low ESR (<40mm/hr) at diagnosis (González-Gay et al., 2017). A study of PMR diagnosis and monitoring in UK general practice found 11.8% of those diagnosed and managed as PMR had a normal ESR at baseline (defined as ESR <20mm/hr) (Helliwell et al., 2013). Those with low ESRs at diagnosis appear to have similar joint symptoms and disease course as those with high ESRs but are more likely to be male, tend to be younger and have fewer constitutional symptoms (Gonzalez-Gay et al., 1997; Helfgott & Kieval, 1996). The finding of a low CRP at diagnosis is less common but there are rare cases of PMR where both CRP and ESR are normal (Manzo & Milchert, 2018) and there does not appear to be a correlation between ultrasound findings and level of ESR / CRP (Mackie, Koduri, et al., 2015). Other inflammatory markers such as plasma viscosity, IL-6 and fibrinogen have been suggested as alternatives to ESR / CRP but are not readily available to most clinicians (González-Gay et al., 2017).

Table 1.1: Classification criteria for PMR

(reproduced from Yates et al. (2016))

Author and year	Proposed Criteria	Requirement for classification
Bird, 1979	Age ≥65 years Bilateral shoulder pain and stiffness; acute or subacute onset (<2 weeks); morning stiffness >1h depression and/or weight loss; bilateral tenderness in upper arm muscles ESR >40 mm/h	Any three, or any one plus temporal artery abnormality (including decreased pulsation, tenderness, beading or bruit).
Jones and Hazelman, 1981	Shoulder and pelvic girdle pain; morning stiffness >1 h; exclusion of rheumatoid arthritis or other inflammatory arthropathy, myopathy, malignancy ESR >30 mm/h or CRP >6 mg/L Rapid response to corticosteroids	All criteria must be met
Chuang, 1982	Age ≥50 years >1 month bilateral aching and stiffness of at least two of the following areas: Neck or torso, shoulders or proximal arms, hips or proximal thighs; exclusion of other causes, ESR >40 mm/h	All criteria must be met
Healey, 1984	Age ≥ 50 years >1 month of neck, shoulder, or pelvic girdle pain (any two areas); morning stiffness >1 h; exclusion of other diagnoses, ESR ≥40 mm/h Rapid response to daily, low-dose steroid therapy (i.e., prednisolone ≤20 mg)	All criteria must be met

Doran, 2002	Age ≥ 50 years Bilateral aching and morning stiffness (lasting ≥ 30 min) persisting for at least 1 month and involving 2 of the following areas: neck or torso, shoulders or proximal regions of the arms, and hips or proximal aspects of the thighs, ESR > 40 mm/h OR rapid response to corticosteroids	All criteria must be met
EULAR, 2012	Morning stiffness ≥ 45 min (2 points); Hip pain, limited range of movement (1 point); Absence of other joint pain (1 point); Normal RhF or ACPA (2 points) Ultrasound criteria: at least 1 shoulder with subdeltoid bursitis and/or biceps tenosynovitis and/or glenohumeral synovitis AND at least 1 hip with synovitis and/or trochanteric bursitis (1 point); both shoulders with subdeltoid bursitis, bicep tenosynovitis or glenohumeral synovitis (1 point)	All patients must be: Age ≥50 years, have bilateral shoulder aching and abnormal ESR/CRP Scoring algorithm - without ultrasound score of 4 needed – with ultrasound score of 5 needed

The most recent European League Against Rheumatism (EULAR) / American College of Rheumatology (ACR) guidance on management of PMR (Dejaco, Singh, Perel, Hutchings, Camellino, Mackie, Abril, et al., 2015) advises that *'clinical evaluation should be directed toward exclusion of relevant mimicking (e.g. non-inflammatory, inflammatory (such as giant cell arteritis or rheumatoid arthritis), drug-induced, endocrine, infective and neoplastic) conditions.'* They suggest a number of baseline investigations to exclude mimicking conditions and establish a baseline for monitoring of therapy. These are summarised in Table 1.2.

One of the most difficult conditions to distinguish from PMR is late onset sero-negative rheumatoid arthritis (RA), which can present as polymyalgia with a mild synovitis and a good response to treatment with glucocorticoids (Cutolo et al., 2009). A study of the performance of the various classification criteria in discriminating PMR from other mimicking conditions found that the EULAR 2012 classification criteria had the highest sensitivity, but their ability to distinguish PMR from sero-negative RA was still suboptimal (Ozen et al., 2016). The distinction between the two conditions is important as they require different drug treatments and have differing prognoses (Manzo & Emamifar, 2019).

In one longitudinal study of 231 patients diagnosed with PMR or GCA, 11 (4.8%) were diagnosed with RA during the follow up period (of whom ten were originally diagnosed as having pure PMR and one had PMR and GCA) (Gran & Myklebust, 2000). The mean duration of PMR at diagnosis of RA was 62.3 months. In another study of a cohort of 116 patients presenting with polymyalgic symptoms, 22 were initially diagnosed with RA and of the remaining 94 who were diagnosed with PMR, 19 (20.2%) went on to be reclassified as having RA during the first 12 months of follow up (Caporali et al., 2001).

There are some clinical features (such as arthritis at the wrist associated with at least one metacarpophalangeal or proximal interphalangeal joint involved at disease onset) which can be helpful in distinguishing late onset sero-negative RA from PMR (Pease et al., 2009), but in general close monitoring and frequent review are needed to delineate the two diagnoses.

Table 1.2: Investigations in suspected PMR (EULAR / ACR 2015 guidelines)

Basic laboratory dataset	Additional investigations to consider	Important comorbidities to identify / consider
Rheumatoid factor and/or anti–cyclic citrullinated peptide antibodies (ACPA)* C-reactive protein and/or erythrocyte sedimentation rate (ESR) Full blood count Glucose Creatinine Liver function tests Bone profile (including calcium, alkaline phosphatase) Dipstick urinalysis	Protein electrophoresis Thyroid stimulating hormone (TSH) Creatine kinase Vitamin D Chest X-ray Anti-nuclear antibodies (ANA) Anti–cytoplasmic neutrophil antibodies (ANCA) Tuberculosis tests	Hypertension Diabetes / glucose intolerance Cardiovascular disease Dyslipidemia Peptic ulcer Osteoporosis / recent fractures Cataract or glaucoma Chronic or recurrent infections Co-medication with non-steroidal anti-inflammatory drugs (NSAIDs), other relevant medications and risk factors for steroid-related side effects.

*Not available to request by GPs in most areas of the UK currently

1.7 Management

The most recent management guidelines from the European League Against Rheumatism and American College of Rheumatology, published in 2015 (Dejaco, Singh, Perel, Hutchings, Camellino, Mackie, Abril, et al., 2015), were developed using the Grading of

Recommendations, Assessment, Development and Evaluation (GRADE) methodology as a framework. These guidelines set out some general principles on the importance of patient centred, individualised care for people with PMR and then make a series of specific recommendations, which are summarised in Table 1.3.

Most clinical practice guidance in the U.K (General Practice Notebook, 2021; NICE, 2021; Tidy & Knott, 2021), is based on the British Society of Rheumatology and British Health Professionals in Rheumatology guidelines (Dasgupta, Borg, Hassan, et al., 2010), which recommend that if PMR is suspected, a trial of 15mg once daily prednisolone is commenced. Response to this should be evaluated after 2-4 weeks and a significant improvement in symptoms can be taken as support for the diagnosis. A lack of response should lead to the diagnosis being re-evaluated. It is known however, that up to approximately 30% of patients with PMR do not adequately respond to steroid treatment within 4 weeks (Dasgupta et al., 2012). Additionally, as discussed above in relation to late onset RA, response to steroid treatment can occur in mimics of PMR.

Once an initial response to prednisolone treatment has been achieved, the dose should be tapered as per the guidance in Table 1.3. Faster tapering regimes than those recommended have been shown to have higher risk of relapse (Dejaco, Singh, Perel, Hutchings, Camellino, Mackie, Matteson, et al., 2015). Recommendations on steroid reduction regimes however, are based on expert opinion because evidence from randomised controlled trials is scarce (González-Gay et al., 2017).

Table 1.3: Summary of the 2015 European League Against Rheumatism (EULAR)/American College of Rheumatology (ACR) recommendations for the management of polymyalgia rheumatica (PMR)

(reproduced from (Dejaco, Singh, Perel, Hutchings, Camellino, Mackie, Abril, et al., 2015))

1	The panel strongly recommends using GC* instead of NSAIDs in patients with PMR, with the exception of possible short-term use of NSAIDs and/ or analgesics in PMR patients with pain related to other conditions. No specific recommendation can be made for analgesics.
2	The panel strongly recommends using the minimum effective individualized duration of GC therapy in PMR patients.
3	The panel conditionally recommends using the minimum effective GC dose within a range of 12.5–25 mg prednisone equivalent daily as the initial treatment of PMR. A higher initial prednisone dose within this range may be considered in patients with a high risk of relapse and low risk of adverse events, whereas in patients with relevant comorbidities (e.g. diabetes, osteoporosis, glaucoma, etc.) and other risk factors for GC-related side effects, a lower dose may be preferred. The panel discourages conditionally the use of initial doses <7.5 mg/day and strongly recommends against the use of initial doses >30 mg/day.
4	The panel strongly recommends individualizing dose tapering schedules, predicated to regular monitoring of patient disease activity, laboratory markers and adverse events. The following principles of GC dose tapering are suggested: A. Initial tapering: Taper dose to an oral dose of 10 mg/day prednisone equivalent within 4–8 weeks. B. Relapse therapy: Increase oral prednisone to the pre-relapse dose and decrease it gradually (within 4–8 weeks) to the dose at which the relapse occurred.

	C. Tapering once remission is achieved (following initial and relapse therapies): Taper daily oral prednisone by 1 mg every 4 weeks (or by 1.25 mg decrements using schedules such as 10/7.5 mg alternate days, etc.) until discontinuation given that remission is maintained.
5	The panel conditionally recommends considering intramuscular (i.m.) methylprednisolone as an alternative to oral GCs. The choice between oral GCs and i.m. methylprednisolone remains at the discretion of the treating physician.
6	The panel conditionally recommends using a single rather than divided daily doses of oral GCs for the treatment of PMR, except for special situations such as prominent night pain while tapering GCs below the low-dose range (prednisone or equivalent <5 mg daily).
7	The panel conditionally recommends considering early introduction of methotrexate (MTX) in addition to GCs, particularly in patients at a high risk for relapse and/or prolonged therapy as well as in cases with risk factors, comorbidities and/or concomitant medications where GC-related adverse events are more likely to occur. MTX may also be considered during follow-up of patients with a relapse, without significant response to GC or experiencing GC-related adverse events. MTX has been used at oral doses of 7.5–10 mg/week in clinical trials.
8	The panel strongly recommends against the use of TNFa blocking agents for treatment of PMR.
9	The panel conditionally recommends considering an individualized exercise program for PMR patients aimed at the maintenance of muscle mass and function and reducing risk of falls especially in older persons on long-term GCs as well as in frail patients.
10	The panel strongly recommends against the use of the Chinese herbal preparations Yanghe and Biqi capsules in PMR patients.

*GC = glucocorticoid

1.7.1 Disease monitoring and follow up

The 2015 guidelines (Dejaco, Singh, Perel, Hutchings, Camellino, Mackie, Abril, et al., 2015) suggest reassessment 1-3 weeks after starting steroid treatment and then review every 4-8 weeks within the first year and every 8-12 weeks in the second year. This may be varied if there are relapses or as treatment is tapered and discontinued. At each clinical review, patients should be evaluated for signs and symptoms of active PMR, disease complications including steroid-related side effects and atypical manifestations, or features suggesting an alternative diagnosis. Risk factors for relapse or need for prolonged therapy should also be assessed.

1.7.2 Relapse

Relapses are defined as the recurrence of polymyalgia rheumatica symptoms that are generally associated with rise in erythrocyte sedimentation rate and C-reactive protein concentration (González-Gay et al., 2017). Most patients with PMR will relapse at some point during the course of their disease and the relapse rate is highest during the first year (Kyle & Hazleman, 1993; Shbeeb et al., 2018). If relapse occurs, the dose of prednisolone should be increased to the pre-relapse dose and then gradually tapered again over the next 4-8 weeks. The occurrence of frequent relapses is an indication to consider methotrexate treatment (Dejaco, Singh, Perel, Hutchings, Camellino, Mackie, Abril, et al., 2015).

1.7.3 Referral to secondary care

The majority of patients with PMR in the UK are managed exclusively in primary care (71% in the most recent assessment of this by Yates et al.) (Barraclough et al., 2008;

Helliwell et al., 2013; Yates et al., 2016). Referral to a rheumatologist is recommended for atypical presentations (including marked peripheral inflammatory arthritis or systemic symptoms, low inflammatory markers, age <60 years), experience of, or high risk of, therapy-related side effects, PMR refractory to prednisolone treatment, and/or relapses or prolonged therapy (Dejaco, Singh, Perel, Hutchings, Camellino, Mackie, Abril, et al., 2015).

1.7.4 Non-steroid drug treatments

People with PMR who have severe side effects from steroids, have risk factors or co-morbidities making steroid treatment less desirable, have recurrent episodes of relapse or who are likely to require long term steroid treatment may be treated with the disease-modifying antirheumatic drug (DMARD), methotrexate (González-Gay et al., 2017). The conditional use of methotrexate is endorsed in the recent EULAR / ACR guidelines because it has been shown in some studies to produce small benefits in reducing rates of relapse and lowering the cumulative dose of glucocorticoids (Caporali et al., 2004; Cimmino et al., 2008; Ferraccioli et al., 1996). Studies which have shown no benefit of methotrexate have been smaller and of low quality (Dejaco, Singh, Perel, Hutchings, Camellino, Mackie, Matteson, et al., 2015). No other DMARDs (e.g., azathioprine, leflunamide) have sufficient evidence to recommend their use in PMR (Dejaco, Singh, Perel, Hutchings, Camellino, Mackie, Abril, et al., 2015).

There has been recent interest in biological agents (TNF-alpha blocking agents) as treatments for PMR, following their successful use in other inflammatory rheumatological conditions such as rheumatoid and psoriatic arthritis. However, trials so far of infliximab

(Salvarani et al., 2007) and etanercept (Kreiner & Galbo, 2010) have failed to show significant benefits and these drugs are not currently recommended for use in PMR. There have been studies of the IL-6 receptor blocker, tocilizumab, that have reported since the EULAR / ACR guidelines, which have shown possible benefits of this drug in PMR (Devauchelle-Pensec et al., 2016; Izumi et al., 2015) and further trials of tocilizumab are ongoing.

1.8 Prognosis

The disease course of PMR is very variable. Some patients may be able to discontinue treatment within one or two years of disease onset whilst others have a much more prolonged illness with recurrent episodes of relapse.

The recent UK study by Partington et al. (2018) found the median duration of time to cessation of continuous glucocorticoid treatment was 1.31 years (IQR 0.65-2.6) but the median duration of total glucocorticoid treatment was 1.93 years (0.95-4.03) with 25% of patients receiving more than 4 years of treatment. A similar study in the U.S. using the Olmstead County cohort found that only 37% (95% CI 31-42) discontinued glucocorticoid treatment by two years and the median time to permanent discontinuation of therapy was 5.95 years (95% CI 3.37-8.88) (Shbeeb et al., 2018). There were differences in these two studies in the way they classified 'end of treatment' and in their inclusion criteria, which may account for some of the difference in results, but both suggest that a significant proportion of patients with PMR are treated with glucocorticoid medication for longer than was previously recognised.

The heterogeneity of PMR has led to the hypothesis that there may be distinct subgroups within the diagnostic classification who have different disease trajectories and prognosis. Muller et al. (2019) explored this in data from their large primary care based PMR cohort study using latent class growth analysis to identify clusters of individuals with differing trajectories of pain and stiffness. They identified five different symptom trajectories, one of which was rapid and sustained recovery ('classic PMR') but others that were more varied (sustained symptoms, partial recovery with sustained moderate symptoms, recovery before worsening, slow continuous recovery).

1.8.1 Predictors of relapse

Many clinical and laboratory parameters have been studied to see if they can predict the course of PMR. High inflammatory markers at baseline, older age, female gender and faster tapering of prednisolone medication have all been found to be associated with longer duration of steroid treatment (Gonzalez-Gay et al., 1999; Narvaez et al., 1999). Persistence of high CRP or IL-6 and faster glucocorticoid tapering regimes have been associated with higher risk of relapse (Kyle & Hazleman, 1993; Salvarani et al., 2005).

A more recent five-year prospective cohort study by Mackie et al. (2010) found a strong association between a starting dose of >15mg prednisolone and a longer course of prednisolone treatment, which they postulated may be due to adrenal suppression by the exogenous steroids. In contrast to previous studies though, this cohort study failed to find any significant predictors of relapse.

1.8.2 Glucocorticoid related adverse effects

Adverse effects of treatment with glucocorticoid medication are common and make a significant contribution to long term morbidity in people with PMR (Dejaco et al., 2011). A systematic review and meta-analysis of adverse effects of low to medium-dose steroid treatment in different inflammatory diseases found an adverse event rate of 80 per 100 person years (CI 15 to 146) based on inclusion of four studies of PMR (Hoes et al., 2009). For comparison the adverse event rate in patients with rheumatoid arthritis in the same study was 43 per 100 person years (CI 30-55). The most commonly reported adverse effects in PMR were gastrointestinal, endocrine and metabolic, cardiovascular and infectious.

A retrospective study from 2012 designed specifically to look at rates of fragility fractures, osteoporosis, hypertension, myocardial infarction, stroke or peripheral arterial disease during glucocorticoid treatment in PMR, found that 43% had at least one of these serious adverse events during a mean duration of therapy of 31 (+/-22) months (Mazzantini et al., 2012). Longer duration of steroid treatment and higher cumulative steroid dose were significantly associated with higher risk of these events occurring. However, a study using the Olmstead County cohort, which compared the cumulative incidence of diabetes, hypertension, hyperlipidaemia, cataracts and fractures over five years in the PMR cohort with a group of non-PMR controls found that only the incidence of cataracts was significantly greater in the PMR group (Shbeeb et al., 2018).

A questionnaire survey of patients' perspectives on steroid related adverse effects found that 100% of a group of 55 people with a variety of inflammatory rheumatic disease diagnoses (including PMR) reported at least one adverse effect from their steroid

medication out of a checklist of 19 (Black et al., 2017). The adverse event prevalence per person was 7.7. The most frequent adverse events reported in this study were thin skin/easy bruising, weight gain, sleep disturbance, and stomach upset/gastric reflux. Interestingly, those ranked by patients as the 'worst' were thin skin/easy bruising, weight gain and sleep disturbance which contrasts with the adverse events typically measured in clinical studies.

1.8.3 Mortality

Survival rates in people with PMR are the same as for the general population and there is no evidence that PMR affects mortality (Gran et al., 2001; Partington et al., 2020; Raheel et al., 2017).

1.9 Areas of uncertainty

There are still many unanswered questions about PMR and its management. The 2015 EULAR / ACR guidelines include a research agenda which is summarised in Table 1.4. A recent Lancet review also highlighted the need, not only for better diagnostic and disease activity biomarkers and trials of new drug treatments, but for patient perspectives of the condition and its impact to be taken into account (González-Gay et al., 2017).

Table 1.4: PMR research agenda

Adapted from the 2015 EULAR / ACR guidelines (Dejaco, Singh, Perel, Hutchings, Camellino, Mackie, Abril, et al., 2015)

1	Which outcome measures including patient-related outcomes, and response, remission and relapse criteria should be used in PMR? What is the value of a composite score? What are the most relevant treatment targets in PMR?
2	What is the efficacy and safety of different routes of glucocorticoid (GC) administration (oral, intramuscular, intra-articular), different initial GC doses, various GC tapering regimens, and different GC flare doses?
3	What is the efficacy and safety of DMARDs (non-TNFa biologic, conventional synthetic and conventional targeted) in PMR? What is the optimal strategy for using DMARDs in PMR: monotherapy versus combination therapy, early versus late introduction, and (particularly for biologics) use with or without GCs?
4	What is the minimal/optimal duration of therapy and which strategies for withdrawing GCs and/or DMARDs yield the best efficacy/safety profile?
5	What is the optimal strategy for shared primary and specialty care including recommendations for specialist referral? How can patients be better involved in treatment decisions, and are there any decision aids? What is the role of self-management?
6	What is the value of tight control (i.e., treat to target) versus conventional management strategies in PMR?
7	How should patients with long-standing disease and long-term low-dose GC therapy be managed?
8	What is the cost utility and effectiveness of DMARD use in PMR (versus GC use alone)?
9	What is the value of non-pharmacological therapies in PMR? Particularly, it is assumed but not yet demonstrated that physiotherapy may support preservation of function and reduce the risk of adverse events related to GC

	use. Patients may benefit from exercise by maintaining muscle mass and function as well as by fall prevention especially in the frail. What is the role of diet in PMR and nutrition supplements (e.g., fish oil) related to outcomes?
10	What is the efficacy and safety of herbal preparations in PMR?
11	What is the role of imaging (particularly ultrasound) for the assessment and monitoring of PMR, identification of overlap with other diseases (e.g., large vessel vasculitis or inflammatory arthritis) alongside clinical and patient reported outcomes?
12	Which biomarkers may be useful in PMR? Why do some patients do better than others? How can we identify these groups and what is the biological mechanism behind it? Should different drugs be applied to different PMR subgroups?
13	What is the morbidity and mortality of PMR patients (with a particular focus on cardiovascular risk) in long-term observational studies?
14	What is the etiopathogenesis of PMR? Which targeted therapies could be developed based on new knowledge of disease mechanisms?

The first point on the research agenda concerns outcome measures, highlighting the lack of consensus about assessment of PMR as a key issue. Indeed, it could be argued that if progress is to be made with any of the items on the research agenda, establishing a way to measure the impact of the condition, and of its treatments, on the people affected by it is essential.

1.10 Summary and conclusions

In this opening chapter, I have provided an overview of what is currently known about the condition polymyalgia rheumatica and the state of current clinical practice. It is clear that

it is a complex condition with uncertainties persisting at every level, from its aetiology and pathology through to its diagnosis and management.

One of the important gaps that is evident in our understanding of the condition is the question of how to measure its impact and evaluate the effects of treatment. In subsequent chapters of this thesis, I will address this issue further by discussing the uses and development of patient reported outcome measures, exploring the measures we currently have for assessing PMR, and going on to describe the development and evaluation of a new patient-reported outcome measure designed specifically for PMR.

Chapter 2: Patient Reported Outcome Measures

2.1 Introduction

In order to discuss outcome measurement in PMR, it is first necessary to consider the importance of outcome measurement in medicine in general. In this chapter I will discuss outcome measurement with a focus on patient reported outcome measures (PROMs), which are central to this thesis.

I will define patient reported outcome measures and outline the different types of PROMs that exist and their respective uses. I will then discuss the background to PROM development and the origins of quality standards for studies evaluating outcome measures. This will set the context for discussion of specific PROM development methodology in later chapters.

2.2 Outcome measurement in medicine

Measurement is central to clinical practice and research, helping inform diagnosis, prognosis and evaluation of interventions. Many types of outcome can be measured, ranging from pathophysiological assessments such as specific blood tests or blood pressure measurement, through to symptoms such as pain or more complex constructs such as quality of life. The way in which a measurement is obtained is dependent on the nature of the outcome; certain measurements have to be obtained by external processes whereas others can only be directly reported by the individual themselves.

To produce meaningful information for patients, clinicians and researchers, a measurement instrument needs to be appropriate for the condition, context and question. However, in many situations there is a lack of evidence and / or consensus on which are the most appropriate outcomes to measure, which are the best instruments to use and when to use them. The use of different measurement instruments, which may vary in content, purpose and quality, to measure the same construct (subject of measurement e.g., depression or pain) limits comparison between studies, inhibits meta-analysis and can hamper development of the evidence base for a condition.

To combat the problems of clinical trials measuring different outcomes for a particular condition (or selectively reporting outcomes for publication), the concept of core outcome sets (COS) arose. A COS for a condition is a standardised set of outcomes that should be measured and reported as a minimum in any clinical trial for that condition, enabling results to be more easily compared, contrasted and combined (Gargon et al., 2014). The COMET initiative (Core Outcome Measures in Effectiveness Trials, http://www.cometinitiative.org/) was founded in 2010 and aims to encourage high quality COS development and support methodological research in this area. They carry out an annual systematic review of COS to populate an on-line database and have shown that year on year, more COS are being developed, they are being developed to a higher standard and involvement of patients and public in COS development continues to increase (Gargon et al., 2019).

2.3 Definition of patient reported outcome measures

"Patient reported outcomes (PROs) are any report of health status that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else. Patient reported outcomes measures (PROMs) are indexes, scales or questionnaires that aim to measure one or more aspects of patient reported outcomes." (International Society for Quality of Life Research, 2019).

In other words, PROMs are a series of structured questions that ask patients about **their** health from **their** perspective.

The value of PROMs in a wide variety of health-care settings and in clinical research is increasingly recognised (Devlin et al., 2010). There are now organisations devoted to evaluating quality of PROMs, providing guidance on PROM design methodology and optimising the use of PROMs for research, service delivery and routine care e.g. the Centre for Patient Reported Outcome Research (CPROR), based at the University of Birmingham (https://www.birmingham.ac.uk/research/activity/mds/centres/cpror/index.aspx), and the COSMIN initiative (COnsensus-based Standards for the selection of health Measurement Instruments, https://www.cosmin.nl/). Traditional ways of measuring health and the effects of treatment are increasingly accompanied by, and in some cases being replaced by, PROMs.

2.4 Types of PROM and modes of administration

PROMs can be generic or condition-specific. Generic PROMs measure the impact of a person's state of health on their overall quality of life and allow comparison of changes to

health across different patient and population groups (Devlin et al., 2010). Commonly used examples include the SF-36 (Ware & Sherbourne, 1992) and the EuroQoL-5D (EQ5D) (Brooks, 1996). Generic measures are particularly useful in assessments of value for money and analysis of productivity and performance.

Condition or disease-specific measures focus on particular attributes related to the condition in question. They can discriminate between people with different degrees of severity of a condition and are more sensitive to specific clinical outcome (Bowling, 2001, Chapter 1). They are therefore useful in clinical research and clinical practice and there are thousands of different condition-specific measures available (Devlin et al., 2010). However, they do not allow comparisons of health across patients with different types of condition. Often therefore, information from a condition-specific and generic PROM will be collected together in trials to allow clinical and broader policy questions to be addressed (Bowling, 2001, Chapter 1).

PROMs data can be collected through the use of paper questionnaires, interviews or, increasingly commonly, PROMs can be completed electronically. The advantages and disadvantages of different modes of administration are outlined in Table 2.1. A systematic review and meta-analysis of studies comparing different modes of administration of health-related PROMs (Rutherford et al., 2016) found no difference in responses between paper based and electronic methods when participants self-completed questionnaires, suggesting that PROMs developed in paper form could be safely transferred to electronic forms. However, for administered questionnaires, responses varied with the largest discrepancy being between clinic administration and home self-completion.

Table 2.1 Advantages and Disadvantages of different modes of PROM administration

Method of PROM data collection	**Advantages**	**Disadvantages**
Pen and paper – by post	Convenient for participant to complete at a time that suits them May be more willing to divulge sensitive information No need for computer literacy	Cost of printing, postage etc. Requires participants to actively send back Data needs entering which is costly and may incur error
Pen and paper – in clinic	High participation rates for those that attend No postage costs No need for computer literacy	If a clinic visit is missed when the PRO completion is scheduled, the opportunity for data collection is missed Participant may feel pressured or rushed to complete the PROM Data needs entering which is costly and may incur error
Face to face / interview administration	May improve response rates May enable people who struggle with reading or writing, those with a disability or who are more unwell, to participate	Requires additional staff time May be less willing to divulge sensitive information

Electronic administration	Reduced researcher time for data entry and no postage or printing costs May be preferred by some groups as convenient and quick May enable housebound people / those with specific disabilities to more easily participate Allows computer adaptive testing to be implemented to reduce patient burden and increase precision	Requires resource to ensure compatibility of databases across different devices / software. May not be accessible to some patient populations. Risk of technical fault / data protection / connectivity issues.
Allowing flexibility of collection method according to patient choice	May improve response rates	Requires additional resource

2.5 Uses of PROMs

2.5.1 PROMs in healthcare evaluation

Healthcare accounts for a large proportion of UK government spending, and resources need to be used in a way that maximises their value to patients and to society.

The term 'value' relates the outcomes of health care interventions to their cost (Gray & el Turabi, 2012). In recent years there has been a shift from an emphasis on quality improvement to the different paradigm of value-based healthcare (Brook, 2010). Value-

based healthcare is a broad concept, going beyond aims of improving efficiency (maximising outcomes for the least cost), practising evidence-based medicine and quality improvement (which are all still highly important for any particular service) and focussing on optimising value for the whole population (Gray, 2017).

Traditional measures of healthcare tended to focus on adverse outcomes such as death, infections and re-admission rates. Over the last few decades, focus has shifted to more comprehensive measures of health status encompassing physical, mental and social components (Brook, 2010). This has allowed different questions to be asked and answered e.g., rather than solely assessing the effect of an intervention on life expectancy, the ability of the intervention to improve function and social participation could be assessed. These measures of health status have also allowed comparison of value across different conditions and different communities i.e., whether one particular intervention produces more health per pound invested than another.

PROMs are one way in which patients' views of changes in their own health status can be captured, in a structured manner, to help improve the quality and effectiveness, but also the value, of health care.

Routine use of PROMs was introduced into the NHS in 2009 under the auspices of the National PROMS programme (https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/patient-reported-outcome-measures-proms), which aims to provide patient-centred information on outcomes of NHS care at provider level, with a view to encouraging poorer performing organisations to implement changes to improve their standard of care.

Up until 2017, the programme involved the routine measurement of PROMs for all NHS patients in England before and after surgery for hernia repair, varicose vein surgery and hip and knee replacement (Department of Health, 2009). From October 2017 onwards, data collection for varicose vein and hernia repair surgery was ceased as it was felt that the value of this information was outweighed by the burden of additional data collection, but the programme continues for hip and knee replacement surgery. The programme uses the EQ5D3-L (comprised of the EQ5D descriptive system and the EQ visual analogue scale, https://euroqol.org/eq-5d-instruments/eq-5d-3l-about/) alongside condition-specific PROMs (the Oxford Hip and Knee scores (Dawson et al., 1996, 1998)) to provide data on symptoms and quality of life.

Whilst the National PROMs programme has been hailed as an important first step towards putting patient-centred outcomes at the heart of healthcare delivery, it has also been criticised for its significant failings. The taskforce that reviewed the programme in 2016 found little evidence to suggest that it had improved performance of outliers (Kyte et al., 2016). In addition, organisations that were not outliers were not using the data to influence their resource priorities (despite spending considerable time and resource collecting it) and the data was not being used by patients to influence their choices. Despite this, the taskforce strongly supported the continued collection of PROMs in the NHS. They suggested however, that the focus should be shifted from the current 'top down' approach to clinic-based collection of PROMs data that could be used for multiple purposes including monitoring outcomes, big-data research and prognostic modelling whilst also allowing instantaneous feedback of PROMs data to patients and clinicians to improve shared decision making (Kyte et al., 2016).

2.5.2 PROMs in research

PROMs are now widely recognised as valid primary or secondary outcomes in research studies to evaluate disease or treatment outcomes (Devlin et al., 2010). The use of PROMs in addition to traditional clinical indicators allows the patient perspective of the physical, functional and psychological impact of the disease and / or treatment to be systematically captured.

In clinical trials, the use of validated PROMs alongside biomedical outcomes allows the impact of an intervention to be more comprehensively assessed (Mercieca-Bebber et al., 2018). PROMS can be used as primary outcomes, if there is a suitable PROM available that appropriately assesses the hypothesised outcome of the intervention, or as secondary outcomes to deepen the understanding of the benefits and risks of the intervention. They can also be used to gather information to support prognostic counselling and to help improve future clinical trials methods e.g., through better understanding of compliance or the burden of measurements (Au et al., 2010). PROM data can therefore provide valuable evidence to inform clinical guidelines, prognostic modelling, labelling claims and health policy (Calvert et al., 2018).

A systematic review of trials registered on the on-line registry www.ClinicalTrials.gov between 2007 and 2013 found that 27% of trials included one or more PROM (Vodicka et al., 2015). It is likely that this number will have increased since, given the number of initiatives to increase use of PROMs in trials (e.g. the European Medicines Agency guidance on the use of PROs in the evaluation of anti-cancer medicinal products (European Medicines Agency, 2016), the United States Food and Drug Administration

guidance on use of PROMs to support medical product labelling claims (U.S Dept of Health and Human Services, 2009) and the Medicare Evidence Development and Coverage Advisory Committee guidance that the quality of life measures should be included in heart failure studies (Centres for Medicare and Medicaid Services., 2017)).

As well as being used in clinical trials, PROMs could provide important data in other types of research, ensuring the patient perspective is captured in these studies too. Routine collection of PROMs into the Electronic Health Record (EHR) could enable inclusion of this information into pragmatic real-world trials carried out using the EHR and into 'big data' longitudinal and cross-sectional observational research (Calvert et al., 2015).

The PROM or PROMs chosen for inclusion in any particular study will depend on the context of the research – the study type, its aims, the population of interest and the research methods (Fitzpatrick et al., 1998). Selecting the PROM which best captures the target outcomes, whilst minimising burden for participants, is of critical importance to collecting high quality publishable data. With increasing numbers of instruments available, the ability to select the right PROM for the study requires good understanding of measurement properties (such as reliability, validity, responsiveness) and of the methods of PROM development and evaluation. These concepts are discussed in more depth in Chapter 6.

Despite the potential benefits of collecting PROMs data in trials, PROM data is frequently wasted through being poorly collected, poorly reported or not reported at all (Bylicki et al., 2015; Friedlander et al., 2016). Attempts to combat this include the development of extensions to both the Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) statement and the Consolidated Standards of Reporting Trials (CONSORT)

statement. The SPIRIT-PRO Extension (Calvert et al., 2018) and the CONSORT-PRO Extension (Calvert et al., 2013) set out recommendations for items that should be addressed and included in clinical trial protocols and reports in which PROs are a primary or key secondary outcome.

2.5.3 PROMs in clinical practice

As well as the macro-level uses of PROMs described above, they can be used at an individual level in clinical practice as part of patient assessment and management. In this context, PROMs can be considered as a tool to help bridge the gap between clinicians' and patients' perspectives of reality and they have been heralded as a way of improving communication, patient satisfaction and even clinical outcomes (Nelson et al., 2015).

In the care of an individual, PROMs may be used as one-off assessments, with scores determining follow up schedules or referral decisions, or used serially in long-term conditions to monitor disease progression and the effects of treatment. Aggregated PROMs data can also be used in consultations with individual patients to support decision making through providing insight into the effects of treatment, impact on quality of life or likely trajectories of illness from a patient perspective (Lewis, 2019).

PROMs can be used to monitor symptoms remotely between clinic reviews, helping to identify problems early, plan appointment schedules and reduce unnecessary appointments (Calvert et al., 2019). A randomised controlled study of using such remote monitoring in people receiving chemotherapy for solid tumours found that people in the intervention arm had better health-related quality of life, fewer emergency admissions and better overall survival than the control group (Basch et al., 2016, 2017).

A realist review by Greenhalgh et al. (Greenhalgh et al., 2017), considered the theories and mechanisms by which PROMs use in clinical practice supports patient-clinician communication and subsequent care processes and outcomes. They identified, considered the mechanisms for, and tested two theories; 1) PROMs completion supports patients to raise issues with clinicians and 2) PROMs scores raise clinicians' awareness of patients' problems. They found support for both these theories and conclude that PROMs completion can change how patients think about their condition and is not merely a neutral act of information retrieval. They also identified complex contextual issues affecting how PROMs are used and their impact on patients and clinicians, suggesting that the clinician-patient relationship shapes the way PROMs are used in consultations and that use of these tools is entangled with the process of maintaining these relationships.

There are however multiple barriers to use of PROMs in clinical practice including clinicians' scepticism about their value, fears about increased workload and burden on patients and worry about disruption to the consultation (Lohr & Zebrack, 2009). A qualitative study of general practitioners' views of using the PHQ-9 assessment tool for depression (Mitchell et al., 2011), carried out at the time when the PHQ-9 had recently been introduced as a requirement of the Quality and Outcomes Framework, found that GPs perceived the mandated use of the tool as mechanistic, intrusive and unnecessary. They also highlighted the difficulties of using such tools in ethnically diverse patient groups and revealed multiple ways in which the tool was used in practice, which differed from the methods by which it had been validated.

A mixed methods study of views of a group of patients with Parkinson's disease and a group of neurologists and physiotherapists on the use of PROMs in consultations, found

that the healthcare professionals (HCPs) felt that use of PROMs could give patients insight into their disease progression, facilitate monitoring of treatment effects and facilitate personalised care and shared decision making (Damman et al., 2019). However, some HCPs expressed the view that many patients might not want PROMs information. In addition, there was scepticism as to the reliability of PROMs as healthcare measurements and HCPs felt they needed more training to be able to use them properly. The patient participants in this study were generally positive about the use of PROMs but HCPs and patients differed in their opinions on which form of PROMs data was most useful (professionals preferring individual data over time and patients preferring aggregated data) and in their views on who should initiate discussion about PROMs data, with each feeling it was the others' responsibility (Damman et al., 2019).

In order to be used in clinical practice, PROMs data must add value to the clinical encounter and be affordable and practical to collect without disrupting clinical workflow (Snyder & Aaronson, 2009). Health care professionals must be trained and supported to use them and be convinced that the tools are robust (Boyce et al., 2014). This means that the selection of instruments must be supported by clear evidence of feasibility, validity and reliability and that any PROMs chosen for use should ideally be integrated into existing clinical systems.

2.5.4 Limitations of PROMs

Whilst PROMs can be valuable in some circumstances as outlined above, they do not necessarily provide all the information needed for decision-making in either research or practice. Clinical findings and measurements are also needed and PROMs data should

therefore be used to complement clinical and other information about patients (Devlin et al., 2010).

Relying on self-report of health status can be difficult, particularly in some groups e.g. young children or those with cognitive impairment. Whilst there are some ways to facilitate use of PROMs in these groups such as specifically designing and validating PROMs in the population concerned, varying the methods of data collection, use of carers to support data collection etc., in some situations use of PROMs is not practical.

Another important consideration is that, as with any data, meaningful and comprehensible presentation of PROMs results is essential to help with correct interpretation. This is especially important if the data is to be relied on to make treatment decisions.

A summary of challenges to the use of PROMs is presented in Table 2.2.

Table 2.2 Current challenges in the uses of PROMs

(adapted from Calvert et al. (2019))

Area	**Challenges**
PROM selection	PROMs are not always designed and selected with patient input to ensure that they measure what matters Measurement properties, patient acceptability and burden, cultural validity, and interpretation guidelines are not always considered Inconsistency in PROMs used within and across disease specialties make comparisons difficult
Ethical considerations	Patients may be unsure why they are being asked to complete a PROM, who will access their responses, and how the data will be used Patient burden of completing multiple questionnaires Inconsistent management of situations where PROM data show levels of symptoms that require an urgent response Poor quality or no reporting of PROM data means that patients may complete multiple questionnaires for no discernible purpose
Data collection, analysis, reporting and interpretation	Engagement and acceptance from stakeholders for PROM collection may be lacking Many clinical trials do not provide a clear rationale for PROM assessment How the data will be used to maximise patient care has not always been fully considered PROM data in research is commonly collected from a relatively small subset of the population, hindering wider

	applicability of findings. Appropriate, culturally validated, alternative language PROMs are often not available Missing data hinder reporting and use, and approaches to minimising missing data are highly variable Lack of consensus regarding analytical approach Many clinical applications of PROMs have been developed in silos and remain unpublished, limiting sharing of implementation strategies, good practice, and results PROM results are often poorly reported and are difficult to access and interpret by patients and clinicians
Data logistics problems	Incompatible IT systems without integration with electronic health records and use across service providers Data stored in different formats Lack of relevant IT/health informatics expertise
Inefficient, uncoordinated approach	Development in silos leads to duplication of effort and inconsistency in collection methods, measures used, and data collected Lack of integration between routine data collected for population level initiatives and individual symptom monitoring, and between routinely collected PROMs and research data Missed opportunity to upscale datasets and enhance efficiency

2.6 Quality criteria for evaluation of healthcare measurement instruments

For a researcher, clinician or patient to know whether an instrument is well designed and appropriate for its intended purpose, there must be systematic methods to evaluate the content and measurement properties of the instrument. The field of evaluation of health outcomes measurement has rapidly developed over the last 30-40 years and has drawn on research from various scientific disciplines including epidemiology, healthcare, psychology and biostatistics to become an international discipline.

Many attempts have been made to define and synthesise criteria that ought to be considered when evaluating a measurement instrument for use for a specific purpose. The timeline shown in Figure 2.1 highlights some of the important stages in the development of this field, showing the formation of key organisations and publications. It is not intended to provide comprehensive review of all of the literature in this area but features key organisations and documents, including COSMIN and OMERACT, which are relevant to this thesis.

The majority of literature discussing methodological issues is focussed specifically on PROMs but can be adapted to apply to other types of measurement instrument. The literature is also predominantly concerned with the evaluation of measurement instruments, but concepts discussed in this literature also inform the methods for developing measures de novo.

Figure 2.1 Timeline of development of quality criteria for evaluating studies of measurement instruments

2.6.1 OMERACT

OMERACT was originally an acronym for an alliance of professionals concerned with Outcome Measures in Rheumatoid Arthritis Clinical Trials. It arose out of a growing awareness through the 1980s that there was a lack of consensus and standardisation of outcome measures used in rheumatoid arthritis (RA) clinical trials. The multiplicity of outcomes and instruments reported in clinical trials of RA made interpretation of findings difficult and there was considerable variation in the way rheumatologists made judgements about individual responses to treatment in those with the condition (Boers et al., 2007).

The first OMERACT conference was held in 1992 and brought together groups of interested professionals (rheumatologists, methodologists, drug regulatory officials and pharmaceutical physicians) from around the world with the aim of developing consensus on the minimum set of outcome measures to be included in all clinical trials of RA and to develop criteria for minimally clinically important improvement and minimum important difference between treatment groups (Kirwan, 2013).

Since then, the organisation has broadened its remit to include the consideration of other rheumatological conditions (OMERACT today stands for Outcome Measures in Rheumatology, www.omeract.org) and importantly now involves patient partners who participate on an equal basis with professional groups. There is a conference every two years with Working Groups for each condition carrying out development and validation work between these meetings and reporting back to a Special Interest Group (SIG) during the biennial conference. For each condition under consideration, the aim is to produce a core set of domains (what to measure) and then agree the instruments to be used (how

to measure) for each domain (Boers et al., 2007). The principles of instrument selection are described by the 'OMERACT filter' which refers to the 'truth, discrimination and feasibility' of any measure (Boers et al., 1998).

I will discuss OMERACT processes, and in particular the work of the OMERACT PMR Special Interest Group, in more detail in Chapter 4 (Section 4.1.1).

2.6.2 COSMIN

The COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) initiative was founded in 2005 with its stated aim being:

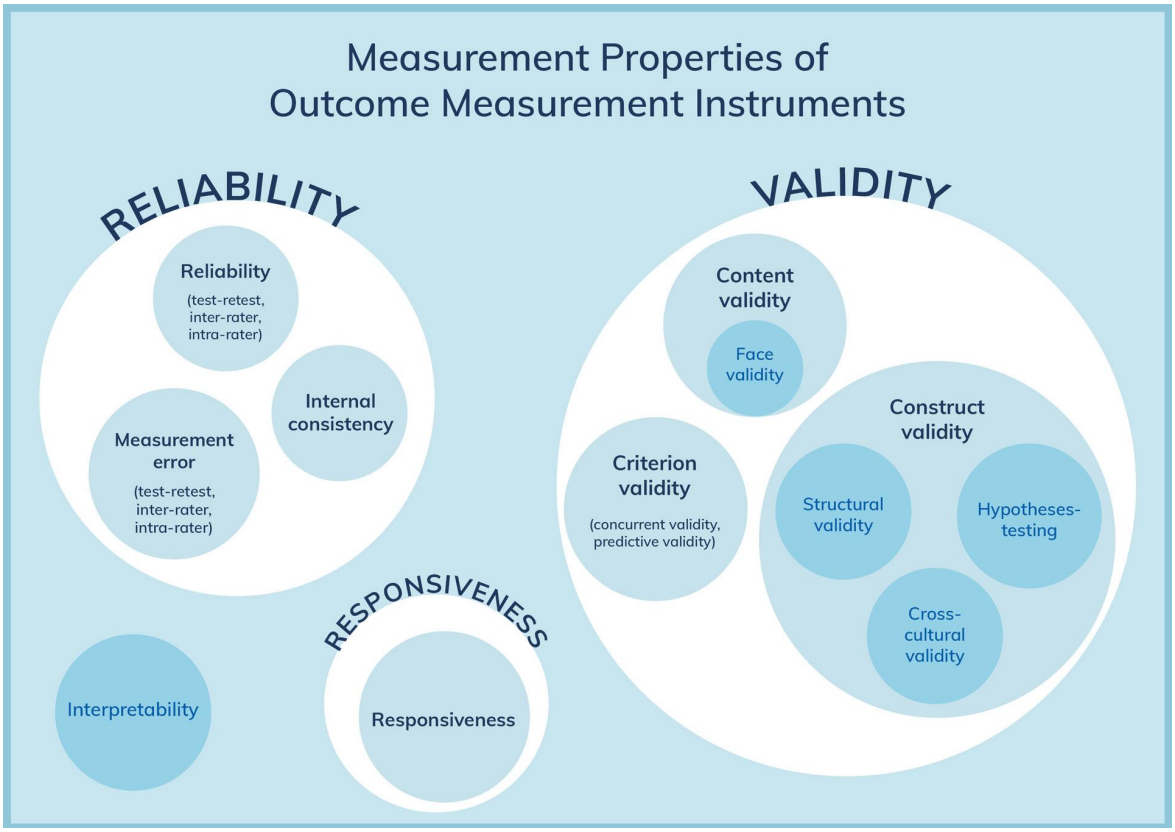
"to improve the selection of outcome measurement instruments both in research and in clinical practice by developing methodology and practical tools for selecting the most suitable outcome measurement instrument." (COSMIN, 2019a).

COSMIN developed a taxonomy and list of definitions of measurement properties (Mokkink et al., 2010b) and published a checklist tool to evaluate the methodological quality of studies on measurement properties of Health-Related PROs (Mokkink et al., 2010a). Both the taxonomy and checklist were developed by consensus through international Delphi Studies.

The COSMIN taxonomy is presented in Figure 2.2. It sets out three quality domains; reliability, validity and responsiveness. Each of these contains a number of measurement properties which all need to be evaluated for any measurement instrument in any application.

Figure 2.2 The COSMIN Taxonomy

(taken from (COSMIN, 2019b)



Definitions for each of these measurement properties are presented in Table 2.3. Each measurement property is discussed in depth in Chapter 6, along with discussion of methods for its evaluation.

Table 2.3 COSMIN definitions of domains, measurement properties and aspects of measurement properties

(taken from (COSMIN, 2019b))

Term			Definition
Domain	Measurement property	Aspect of a measurement property	
Reliability			The degree to which the measurement is free from measurement error
Reliability (extended definition)			The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: e.g. using different sets of items from the same health related-patient reported outcomes (HR-PRO) (internal consistency); over time (test-retest); by different persons on the same occasion (inter-rater); or by the same persons (i.e. raters or responders) on different occasions (intra-rater)
	Internal consistency		The degree of the interrelatedness among the items
	Reliability		The proportion of the total variance in the measurements which is due to 'true'† differences between patients
	Measurement error		The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured
Validity			The degree to which an HR-PRO instrument measures the construct(s) it purports to measure
	Content validity		The degree to which the content of an HR-PRO instrument is an adequate reflection of the construct to be measured
		Face validity	The degree to which (the items of) an HR-PRO instrument indeed looks as though they are an adequate reflection of the construct to be measured
	Construct validity		The degree to which the scores of an HR-PRO instrument are consistent with hypotheses *(for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups)* based on the assumption that the HR-PRO instrument validly measures the construct to be measured
		Structural validity	The degree to which the scores of an HR-PRO instrument are an adequate reflection of the dimensionality of the construct to be measured
		Hypotheses testing	Idem construct validity
		Cross-cultural validity	The degree to which the performance of the items on a translated or culturally adapted HR-PRO instrument are an adequate reflection of the performance of the items of the original version of the HR-PRO instrument
	Criterion validity		The degree to which the scores of an HR-PRO instrument are an adequate reflection of a 'gold standard'
Responsiveness			The ability of an HR-PRO instrument to detect change over time in the construct to be measured
	Responsiveness		Idem responsiveness
Interpretability*			Interpretability is the degree to which one can assign qualitative meaning - that is, clinical or commonly understood connotations – to an instrument's quantitative scores or change in scores.

† *The word 'true' must be seen in the context of the CTT, which states that any observation is composed of two components – a true score and error associated with the observation. 'True' is the average score that would be obtained if the scale were given an infinite number of times. It refers only to the consistency of the score, and not to its accuracy (ref Streiner & Norman)*

* *Interpretability is not considered a measurement property, but an important characteristic of a measurement instrument*

2.7 Summary and conclusions

In this chapter, I have discussed the importance of measurement in medicine in general and then focussed on one specific type of measurement tool, patient-reported outcome measures. I have defined PROMs and outlined their uses in three broad areas – economic evaluation of healthcare, clinical research and clinical practice. I have then given an overview of the background to development of quality criteria for evaluation of measurement instruments with particular focus on two organisations in this field, OMERACT and COSMIN.

All of this gives context for the body of work in this thesis which is concerned with outcome measurement in polymyalgia rheumatica. As outlined in Chapter 1, PMR is a heterogenous condition causing symptoms including pain and stiffness (which are necessarily 'patient reported') and varying degrees of functional impact over a prolonged period of time. The gap in our ability to measure the impact of the condition and its treatment on those affected has been highlighted in recent guidance and reviews (Dejaco, Singh, Perel, Hutchings, Camellino, Mackie, Abril, et al., 2015; González-Gay et al., 2017). There are therefore important questions to be addressed regarding how outcomes are currently measured in PMR and whether a PROM for the condition could be a valuable tool in achieving better, person-centred assessment.

Chapter 3: Aims and objectives

3.1 Introduction

In this chapter, I will set out the aims and objectives of this thesis and present an overview of the body of work that contributes to achieving these.

3.2 Aims

1. To establish how PMR is currently assessed in clinical research and summarise the evidence that supports the use of existing outcome measures.
2. To develop a patient-reported outcome measure that assesses the impact of PMR on a person's life, to bridge the gap between patient and clinician perspectives and facilitate person-centred assessment of the condition.
3. To evaluate this new outcome measure to establish its suitability for use in clinical research.

3.3 Objectives

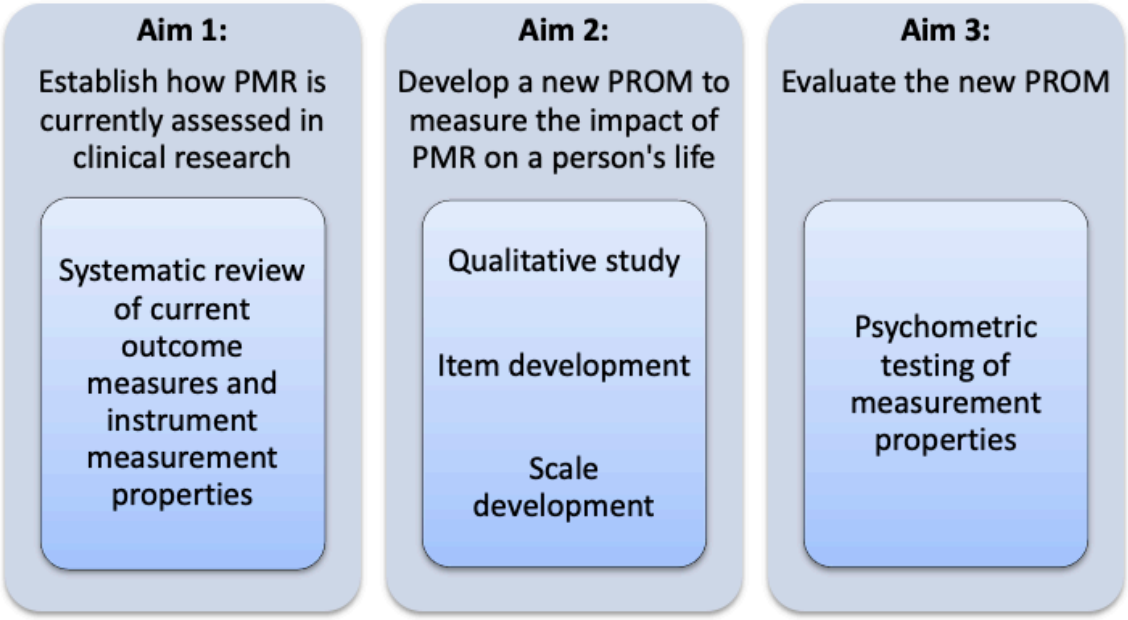
In order to achieve these aims, several objectives need to be met. Figure 3.1 shows how the objectives relate to the aims.

The objectives are:

1. To conduct a systematic review of all outcomes measured, and instruments used, in clinical studies of PMR to date.

2. To conduct a systematic review of the existing literature pertaining to measurement properties of instruments used in PMR research.
3. To develop a patient reported outcome measure to evaluate the impact of PMR on a person's life.
4. To test the measurement properties of this new patient reported outcome measure when completed by people with PMR.

Figure 3.1 Aims and objectives

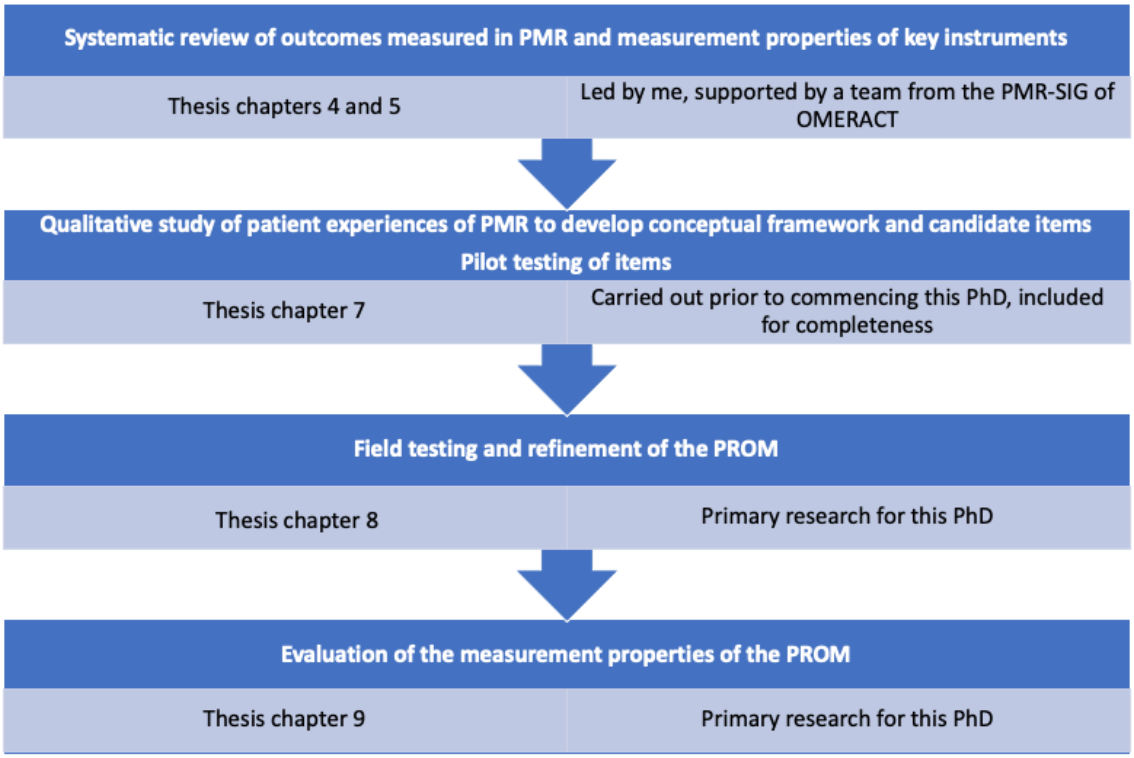


3.4 Overview of planned research

The research required to meet the stated aims and objectives is summarised in Figure 3.2. I will first conduct a systematic review of the literature on outcomes and instruments used in studies of PMR to date. I will then conduct two primary research studies; 1) to develop the PROM and 2) to evaluate its measurement properties. This research builds on work that I completed prior to commencing my PhD (described in Chapter 7) which

laid the foundations for PROM development by exploring patient experiences of the PMR and identifying a long-list of items for consideration. This preliminary research will be summarised in this thesis to demonstrate the full PROM development process.

Figure 3.2 Summary of research presented in this thesis



3.5 Summary and conclusions

This chapter sets out the aims and objectives of this thesis. Chapters 1 and 2 have provided the background and rationale for the work, discussing the current understanding of polymyalgia rheumatica and the importance of outcome measurement in this condition and more generally. Subsequent chapters will present the research studies designed to answer the aims and objectives described.

Chapter 4: A Systematic Review of Outcome Measures Used in Studies of Polymyalgia Rheumatica

4.1 Background

In this chapter, I will set the context for this systematic review by describing the OMERACT process to develop standardised sets of outcome measures for rheumatological conditions. I will then report a systematic review of outcome measures used in studies of PMR, which I carried out on behalf of the OMERACT PMR-Special Interest Group. This is a comprehensive summary of the evidence base to date in this area and ultimately underpins the rationale for the development of a new measure for this condition.

4.1.1 The OMERACT process

The context for the formation of OMERACT and the development of the organisation are described in Chapter 2 (2.6.1). OMERACT works to develop core domain sets ('what to measure') and core instrument sets ('how to measure') for clinical trials and observational studies of rheumatological conditions. It does this through a standardised process involving evidence review and consultations with key stakeholder groups.

OMERACT uses the terminology of a 'filter' to explain the process of assessment that an instrument must go through to determine whether it is included in a core instrument set. Figure 4.1 shows the current version of this, Filter 2.0 (Boers et al., 2014). The process begins with determination of at least one core domain within each of four Areas of

Outcome (life impact, pathophysiological manifestations, resource use and death). Next, instruments to measure these core domains are identified and evaluated against the three 'pillars' of the filter - truth, discrimination and feasibility - to determine the final core instrument set.

Figure 4.1 OMERACT Filter 2.0

(Copied from (Boers et al., 2014))

Characteristics of OMERACT Filter 2.0

Structure

- There are two concepts to outcome incorporating the impact of health conditions and their pathophysiological manifestations.
- There are four Core Areas of outcome: Death; Life Impact; Resource Use* and Patho-physiological Manifestations. Every clinical trial must include at least one measure under each of these headings.
- Within each Core Area there are Domains of interest to particular conditions. Experts and stakeholders should determine at least one Domain to be a core outcome within each Core Area. This is the Core Domain Set. Trial designs are not limited to the Core Domain Set but should include them in all clinical trials in that condition in addition to any other domains that might be relevant to their investigation.
- Within each Core Domain at least one valid outcome measure should be identified. Validity is assured by meeting the requirements of Truth, Discrimination and Feasibility (as described in the original OMERACT Filter).
- The resultant Core Outcome Measurement Set, which includes at least one instrument from each Core Domain, and at least one domain from each Core Area, should be included in the outcomes of all clinical trials in that condition. Trial designers may also incorporate any other outcomes of interest, including a designated primary outcome which is not part of the Core Outcome Measurement Set but is relevant to their investigation.

Process

Identify the Core Domain Sets

- A literature review of domains and instruments previously used in the condition.
- A review of the setting and any contextual factors that need to be taken into account.
- Structured enquiry with stakeholders on their views on domains of importance.
- Full participation of all stakeholders (including patients) in a consensus process to determine agreement on *what* to measure – the Core Domain Set.

Identify the Core Outcome Measurement Set

- Full literature review to identify validated and applicable outcome instruments for each Core Domain.
- Validate instruments in the condition of interest if this has not been done.
- Develop and validate new instruments for a Domain that does not have an outcome measurement instrument.
- Full participation of all stakeholders (including patients) in a consensus process to determine agreement on *how* to measure – the Core Outcome Measurement Set.

4.1.2 The OMERACT PMR working group and my role as a fellow

As discussed in Chapter 1, PMR is a heterogeneous condition causing pain, stiffness and disability in older adults who often have comorbidities and may already face the challenges of social isolation and frailty and have social care needs.

To improve the evidence base for the management of PMR, high-quality clinical studies are urgently needed. A core outcome set to guide researchers will help facilitate this and allow standardisation of future data sets, which will improve comparability and allow pooling of results.

The PMR Working Group was first formed for OMERACT 11 in 2014 with the aim of developing a core outcome set for the condition. Following the recommended OMERACT process, the initial focus of the Working Group was on developing a core domain set. Several simultaneous strands of work were carried out to prepare for the OMERACT 2016 meeting where the core domain set was debated and agreed. These work strands are outlined below.

I have been a member of the OMERACT PMR Working Group since 2015 and was a group Fellow for the 2016 and 2018 meetings. Work strands one and two occurred before I joined the group but are described here in the background to demonstrate the full process. I participated in the Delphi exercise in work strand three and carried out a qualitative study of patient experiences of PMR which contributed to work strand four. In my role as group fellow, I participated in the preliminary work to collate all the relevant evidence ahead of the meetings and contributed to the group's presentations at the meetings (poster and pre-recorded video presentation in 2016 due to being on maternity

leave). The knowledge gained and collaborations formed through this role contributed to the specific aim of this doctoral thesis of developing a new outcome measure for PMR.

Work strand 1: scoping

A scoping consultation exercise was conducted with patients using a modified nominal group technique (group discussion between patients with a health care professional facilitating) (Mackie et al., 2014). A convenience sample of patients recruited from three UK and one Belgium centre was used, involving a total of 104 patients. Discussions were based about the three pre-specified topics of symptoms, diagnosis and treatment, followed by sorting cards to identify each patient's top ten items for each topic.

Work strand 2: comparison with existing literature

The domains identified as important by patients were then compared to those measured in existing studies. It was found that there was mismatch between what was actually being measured by researchers and what patients felt was important to measure. For example, sleep disturbance and fatigue were identified as important by patients but rarely assessed in studies. There were also significant limitations in how outcomes were presented in the literature. For example, when pain was reported, the question that was asked to elicit a pain score was often not explicitly stated such that it was unclear what sites and what time period were being considered (Mackie et al., 2017).

Work strand 3: Delphi exercise

A three round Delphi study was then carried out with patient and professional groups to narrow down the domains though a consensus process (Helliwell et al., 2016). 55 patients with PMR and 85 clinicians (from a range of professional backgrounds including rheumatology and general practice) from Europe, North America and Australasia took

part. In the first round, participants were asked to select the top ten domains they felt should be included in a core domain set from a list and to suggest additional domains if they wished. Those domains selected by >70% of respondents from the previous round were included. Remaining domains identified by >20% of either group were presented for a second round. The third round sought an overall opinion on the combined outcome set. At the end of this process, a draft core domain set for proposal at the subsequent OMERACT meeting was formed.

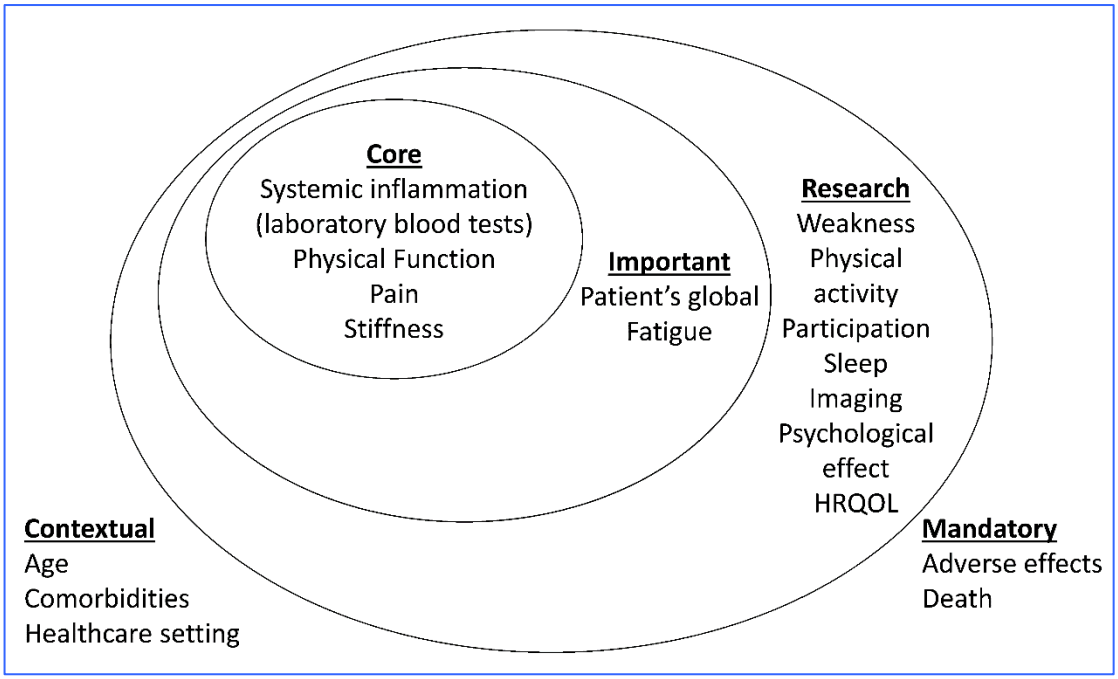
Work strand 4: Qualitative studies

Qualitative work on core symptoms experienced by patients (pain and stiffness) (Mackie, Hughes, et al., 2015), in addition to work investigating the broader patient experience of the condition (Twohig et al., 2015), was carried out concurrently with the Delphi study. The section of this work that I led is presented later on in this thesis (Chapter 7) and not detailed further here.

An onion diagram of the proposed core domain set (Figure 4.2) was produced and voted on by OMERACT participants at the final plenary session of the 2016 meeting. 93% of those present agreed with the final proposed inner core domain set – laboratory markers of systemic inflammation, pain, stiffness and physical function.

Figure 4.2 Proposed core domain set for PMR clinical trials

(published in the Journal of Rheumatology (Mackie et al., 2017))



Proposed core domain set for PMR clinical trials. This "onion" diagram uses nested circles with the innermost circle denoting the Inner Core (mandatory to measure in all clinical trials of PMR), the middle circle denoting Important Outcomes (strongly recommended to measure in PMR), and the outer circle denoting the Research Agenda (domains that require further investigation in PMR). Mandatory domains (bottom right) are those that should be reported by default in all clinical trials of any condition. The proposed contextual factors (bottom left) are suggestions we received regarding possible contextual factors and represent hypothesized factors only.
PMR: polymyalgia rheumatica; HRQOL: health-related quality of life.

4.1.3 Rationale for this review of outcomes measures in PMR

After agreement of the core domain set, the next step in the OMERACT process is to consider candidate instruments for each of the domains.

In 2015 a systematic review of 35 studies conducted by the OMERACT PMR Working Group found significant variability in the assessment of PMR in research settings and most of the instruments identified were deemed to have been insufficiently validated according to the OMERACT Filter 2.0 (Duarte et al., 2015). However, this systematic

review was limited in its scope and did not make any methodological assessment of the quality of included studies.

To identify which instruments should be considered as candidates for each of the domains, a new systematic literature review of outcomes and instruments used in studies of PMR to date was planned. This is presented below.

4.2 Aims

The aims of this review were to:

1. Identify all outcomes that have been measured in studies of PMR, and the instruments used to assess them.
2. Categorise these outcomes and instruments into the domains defined in the core domain set agreed by the OMERACT PMR Working Group in 2016 (laboratory markers of systemic inflammation, pain, stiffness, physical function).
3. Assess the number and quality of studies using each outcome / instrument to help determine which instruments to take through the full OMERACT Filter 2.1 instrument selection process.
4. Identify research gaps with respect to outcomes and instruments used in studies of PMR.

A further systematic review of the literature on the measurement properties of selected instruments will then be carried out to determine their suitability for inclusion in a core outcome set for PMR. This second, linked, review is reported in Chapter 5.

4.3 Methods

4.3.1 Protocol development and registration

The protocol for this review was written in accordance with the PRISMA-P (Preferred Reporting Items for Systematic Reviews and Meta-analyses – Protocols) 2015 guidelines (Moher et al., 2015). This provides a 17-point checklist with an associated elaboration and exploration paper (Shamseer et al., 2015) providing evidence-based explanations for each item. The PRISMA-P statement was developed with the aim of improving the "transparency, accuracy, completeness and frequency of documented systematic review and meta-analysis protocols" which in turn should reduce selective reporting, reduce duplication and encourage collaboration.

I undertook training in literature searching and systematic review techniques from information scientists at both Keele University and the University of Sheffield prior to commencing the review. The protocol was reviewed and approved by Dr Opeyemi Babatunde (Research Associate, systematic reviews, Keele University) and by the OMERACT PMR-working group. The full protocol is included in Appendix 4.1: Protocol for systematic review of outcomes measures in PMR. No amendments have been made to the protocol.

The protocol was registered on PROSPERO (https://www.crd.york.ac.uk/prospero), an international database of prospectively registered systematic review protocols in any area where there is a health-related outcome. It stores a permanent searchable record of the review protocol and therefore helps avoid duplication and also facilitates transparency in the review process by enabling comparison of the final review with the protocol. It is

funded by the National Institute for Health Research (NIHR) and is produced by the Centre for Reviews and Dissemination at the University of York.

This review is registered as: “A systematic review of outcome measures used in research studies of polymyalgia rheumatica (PMR): a review to inform the work of the OMERACT PMR Special Interest Group in determining a core outcome set for clinical trials of PMR” https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=80058 Registration number CRD42017080058.

4.3.2 Search strategy

To identify all published articles containing any research reporting outcomes of PMR, literature searches were carried out in five electronic bibliometric databases (MEDLINE via OVID, CINAHL via EBSCO, Embase via HDAS, Web of Science and the Cochrane Library (Cochrane central register of controlled trials and Cochrane database of systematic reviews)). All databases were searched from database inception until September 30th, 2017. The search strategy was based on the MeSH term “polymyalgia rheumatica” and adapted for each database with assistance from a specialist librarian.

The full search strategies are included as Appendix 4.2: Outcomes in PMR systematic review search terms. As an example, the strategy for OVID Medline is shown in Table 4.1. Reference lists from key texts and other relevant systematic reviews (identified by the search strategy, though not included in the final review) were studied to identify any additional papers missed by the primary search. Clinical trials registries (ClinicalTrials.gov, ISCTRN and the EU Clinical Trials Register) were reviewed to track any ongoing or

unpublished studies. Experts in the field were contacted to see if they were aware of any ongoing studies that may be of relevance.

Table 4.1 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

4.3.3 Eligibility criteria

Studies were eligible if they included patients with PMR and reported original quantitative data on the outcomes of PMR. A range of study types including, but not limited to, 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 without original quantitative patient data were excluded. The full text of the article had to be available, either in English or a language that could be translated with the assistance of either Google Translate or an investigator from the OMERACT PMR Working Group or affiliated academic institutions who is a competent speaker of the language concerned.

Studies that did not include patients with PMR or considered patients with PMR and GCA as a single group and did not present disease specific data, were excluded. Although PMR and GCA can be considered a continuum of the same disease process (and are often studied together) they have sufficient differences that the relevant outcomes do not entirely overlap, and as such it was felt that including studies presenting mixed data in this review would reduce the validity of the review. Diagnostic studies and studies that solely reported outcomes not pertaining to PMR directly (e.g. fractures secondary to steroid treatment, cardiovascular events in patients with PMR) were also excluded.

Table 4.2 Inclusion and exclusion criteria

Criterion	Inclusion	Exclusion
Patients	Patient population with PMR diagnosis	Mixed PMR / GCA / other inflammatory arthritides populations
Study type	Research studies reporting original data where an outcome of PMR was measured: - Randomised controlled trials - Other interventional trials - Longitudinal observational studies - Prospective cohort studies - Cross sectional studies - Case control studies Studies evaluating outcome measures in patients with PMR Published protocols with clear descriptions of the intended outcome measures	Diagnostic studies. Editorials, commentaries guidelines, review articles, case reports or letters not including original patient data or purely qualitative studies.
Availability	Full text articles English language or able to be translated by Google translate or the involvement of an investigator from the OMERACT PMR Working Group / someone in their institution who is a competent speaker of the language concerned	No access to the full text or not in a language that can be translated to be understood by the reviewers.

4.3.4 Review team

The full review team and their roles in the process are set out in Table 4.3. Where they are referred to individually subsequently in this chapter, initials are used.

Table 4.3 The review team

Name	Initials	Institution	Role in the review process
Dr Sara Muller	SM	Keele University	Protocol development, screening of full texts, checking data extraction and risk of bias assessment, analysis support
Dr Caroline Mitchell	CM	University of Sheffield	
Prof Christian Mallen	CDM	Keele University	
Dr Sarah Mackie	SLM	Leeds University	
Dr Claire Owen	CO	University of Melbourne	
Prof Samantha Hider	SH	Keele University	Arbiter for conflict resolution, checking data extraction and risk of bias assessment
Prof Catherine Hill	CH	University of Adelaide	Screening of full texts, checking data extraction and risk of bias assessment

4.3.5 Study selection

Identified studies were imported into bibliographic management software (Endnote X8, https://endnote.com) and duplicates removed. I screened the titles and then uploaded the database of de-duplicated and title-eligible studies to Covidence (https://www.covidence.org/home). Covidence is software developed by a not-for-profit team working with the Cochrane Collaboration, which aims to speed up the systematic review process by providing one platform on which to go through the whole review process from when citations are uploaded through to data extraction and quality

assessment. It allows teams to work together on the same review and provides an audit trail for each stage of the process.

CO and I independently screened abstracts against the inclusion and exclusion criteria. Disagreements were resolved by discussion and if needed by consensus with a third reviewer (SH). One other team member and I then independently screened full texts of selected articles, with conflicts resolved in the same way, to select the papers to take forwards to full review.

4.3.6 Data extraction

I developed a data collection spreadsheet specifically for this review and it was piloted and amended by members of the wider review team before use. Data extracted from each relevant study included 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.

I extracted the data for all studies and each was independently reviewed by another member of the team.

4.3.7 Risk of bias in individual studies

Although the quality of included studies is not directly relevant to the research question of which outcome measures and instruments have been used in studies of PMR, evaluating the quality of the studies could aid critical judgement about which outcome measures and instruments to prioritise. If a poorly planned and conducted study used an

outcome measure not used in other studies it is likely to be less significant than if an unusual outcome measure was used in a study judged to be of very high quality.

A modified Quality In Prognosis Studies (QUIPS) tool (Hayden et al., 2013) was used for assessing risk of bias in the included studies. This tool was selected as it was appropriate to most study types included in the review. Domains 1 (study participation), 2 (study attrition) and 4 (outcome measurement) were applied (Table 4.4). Domains 3 (prognostic factor measurement), 5 (study confounding) and 6 (statistical analysis and reporting) 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., 2016) were applied to included randomised controlled trials to ensure that these studies were appropriately assessed.

Risk of bias assessment was carried out at the same time as data extraction. Data pertaining to each domain was entered into a spreadsheet and judgement made to categorise each study into high, moderate or low risk for each domain. I carried out this process initially with the extracted data and judgement reviewed by a second member of the team. Any disagreements were discussed, and consensus reached.

Developers of the QUIPS tool do not recommend use of a summated score for overall study quality (Hayden et al., 2013). It is suggested that if an overall judgement is helpful it can be made by considering the rating of the most important domains (which should be identified as such in advance). In PMR research, the heterogeneity of the condition and the difficulties of diagnosis mean that participant characteristics, inclusion and exclusion criteria and criteria used to define PMR are all key when considering risk of bias. These are captured by the study participation domain. The study attrition domain was difficult

to apply to several of the study designs (e.g. efficacy studies, case series) included in the review and therefore is a less reliable marker of overall risk of bias. Outcomes are being considered in more detail and it is known that there are not clearly reliable and valid outcome measures available for PMR (hence the existence of this review) so this domain is arguably less informative in relation to bias. The study participation domain was therefore used as a marker of overall risk of bias.

For RCTs included in the review, additional criteria of adequacy of randomisation, adequacy of blinding and assessment as to whether the groups were considered equally throughout were applied. These were used to assign a low / medium / high rating to supplement the rating from the QUIPs domains.

The assessment of risk of bias for each study was used in critical judgement of the weight given to the study in informing discussion about outcome measures to take forwards to further assessment in the OMERACT process.

Table 4.4 Summary of the QUIPS domains and prompting items used

(modified from (Hayden et al., 2013))

Variable	Bias Domain		
	Study participation	**Study Attrition**	**Outcome Measurement**
Optimal study or characteristics of unbiased study	The study adequately represents the population of interest	The study data available (i.e., those not lost to follow up) adequately represents the study sample	The outcome of interest is measured in a similar way for all participants
Prompting items and considerations	Adequate participation by eligible persons	Adequate response rate	A clear definition of the outcome is provided
	Description of the source population or population of interest	Description of attempts to collect information on dropouts	Method of outcome measurement is adequately valid and reliable
	Description of the baseline study sample	Reasons for loss to follow up are provided	The method and setting of the outcome measurement is the same for all participants
	Adequate description of the sampling frame and recruitment	Adequate description of participants lost to follow up	
	Adequate description of inclusion and exclusion criteria	There are no important differences between those who completed the study and those who did not	

4.3.8 Analysis

Outcomes and instruments were categorised into the relevant area defined in the core domain set agreed by the OMERACT PMR Working Group in 2016 (laboratory markers of inflammation, pain, stiffness, physical function). Instruments measuring domains outside

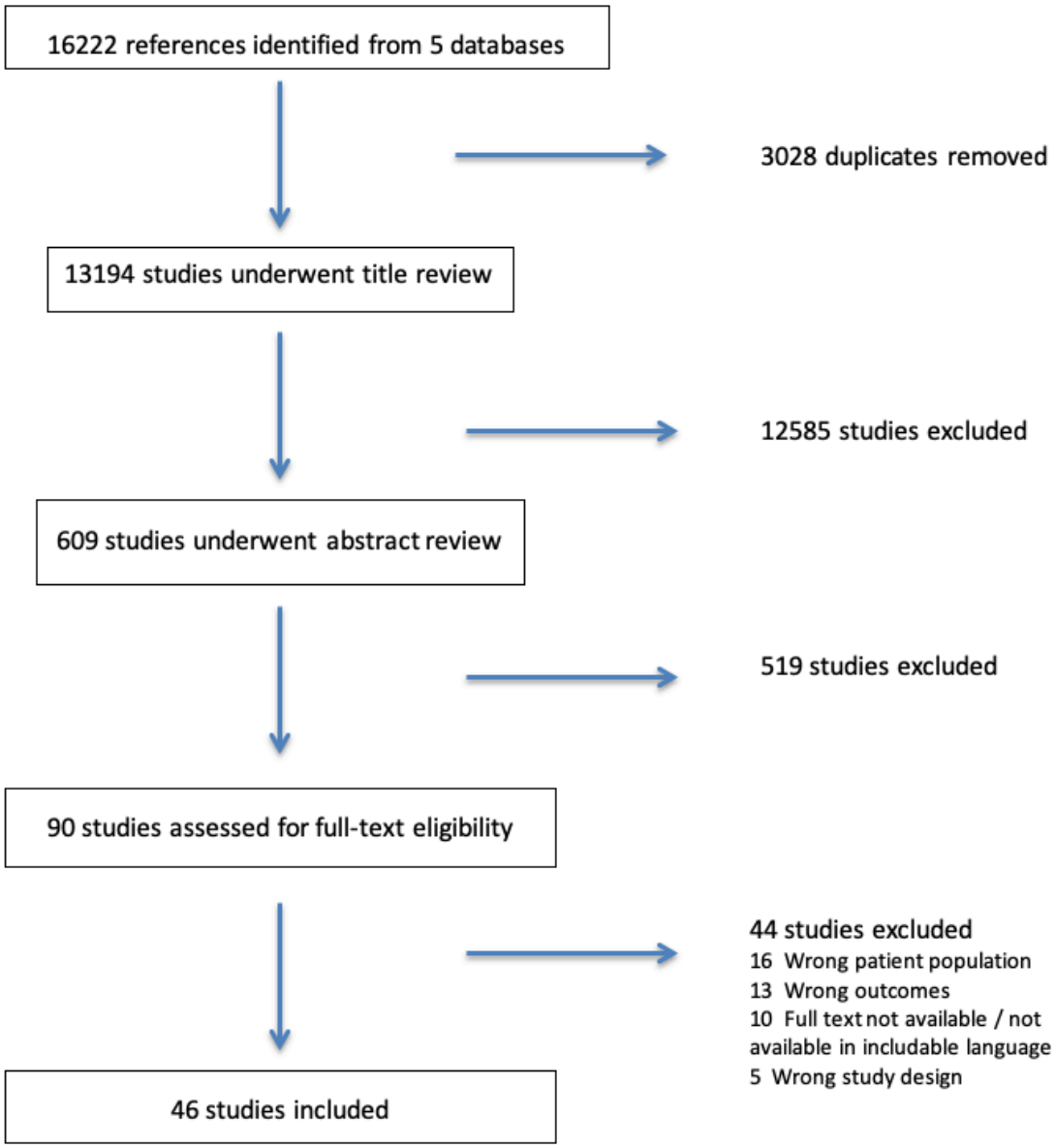
the core set were also collated to establish which other constructs are being assessed in studies of PMR as this may inform the future research agenda. A narrative review of the results was carried out. A meta-analysis was not conducted, as it was not applicable to our research aims of identifying and categorising all outcome measures and instruments used in studies of PMR.

4.4 Results

13 194 publications were identified (after duplicates were removed). This was reduced to 609 after title screening. Abstract screening resulted in 90 full text studies eligible for assessment and 46 of these were selected for inclusion in the review.

No additional studies meeting the eligibility criteria were identified from reference lists of key texts or previous systematic reviews (see details below). Key experts in PMR research were contacted to see if they knew of any additional ongoing studies of relevance, but no new studies were identified.

Figure 4.3 PRISMA flow diagram of results



4.4.1 Excluded trials and comparison with other systematic reviews

The most common reason for excluding studies at the full text review stage was that they did not meet the criteria for population studied, most frequently because they considered participants with GCA and PMR as one group. The group designated 'wrong outcomes'

included studies that focussed on associations of PMR rather than features of the disease itself e.g. fractures secondary to steroid use, development of cancer or cardiovascular disease in people with PMR. Details of the ten studies for which the full text was not available or was not available in an includable language are included in Appendix 4.3: Studies for which full text was not available.

Reference lists from two other related systematic reviews (Duarte et al., 2015; Huang & Castrejon, 2016) were cross-referenced to ensure that no key articles had been missed in our searches or excluded inappropriately.

4.4.2 Additional studies identified from trials registries

Three clinical trials registries were searched on 15/3/18, ClinicalTrials.gov (https://clinicaltrials.gov), the EU clinical trials register (https://www.clinicaltrialsregister.eu) and the ISRCTN registry (https://www.isrctn.com). Eight ongoing or unpublished studies were identified from ClinicalTrials.gov with no additional studies found on either of the other two registers. Table 4.5 details the studies found and their proposed outcome measures.

Table 4.5 Ongoing / unpublished studies identified from clinical trials registries

Study Title / Aim	Status	Proposed outcomes
PMR-SPARE - A Randomized, Double-Blind, Placebo-Controlled, Parallel Group Study to Evaluate the Efficacy of Tocilizumab as a Remission-Induction and Glucocorticoid-Sparing Regimen in Subjects With New-Onset Polymyalgia Rheumatica	Ongoing	Proportion in GC free remission at week 16, cumulative prednisolone dose at various time points, number of flares, time to 1st and 2nd flare, SF-36, FACIT-fatigue, HAQ, patient global, physician global, duration and severity of MS, EUL, ESR, CRP, adverse events (changed in vital signs, FBC, clinical chemistry parameters)
SEMAPHORE - Safety and Efficacy of tocilizuMAb Versus Placebo in Polymyalgia rHeumatica With glucocORticoid dEpendence	Ongoing	PMR-AS, cumulative doses of GCs
A Multi-Center, Randomized, Double-Blind, Placebo-Controlled, Parallel Group Study to Evaluate the Efficacy and Safety of Sirukumab in Subjects With Polymyalgia Rheumatica	Withdrawn before recruitment started	Not specified
A Randomized, Open-label, Dose-ranging Study of Oral Delayed Release Prednisone in Patients With Untreated Polymyalgia Rheumatic (PMR).	Terminated due to lack of recruitment	Severity of morning stiffness (VAS), duration of MS, pain VAS, fatigue VAS, PMR-AS
Infliximab Therapy in Patients With Refractory Polymyalgia Rheumatica: a Double Blind Placebo Controlled Trial	Unclear, last updated 2011, no corresponding paper	Proportion in GC free remission at 24 weeks, time to respond, relapses and recurrences, cumulative prednisolone dose, side effects of prednisolone, side effects of infliximab.
Efficacy of Micro-pulse Steroid Therapy as Induction Therapy in Patients With Polymyalgia Rheumatica	Unclear. Last updated 2010	Not specified
Circadian Variation in Cytokines and the Effect of Timed Release Tablet Prednisone in Polymyalgia Rheumatica	Not yet commenced	IL-6 and other cytokines, morning stiffness (mins), pain VAS, patient's opinion of condition, clinician's opinion of condition, HAQ, BRAF-MDQ fatigue scale and the Hospital Anxiety and Depression Scale
A 3-arm proof of concept study of AIN457, ACZ885 or corticosteroids in patients with PMR	Terminated early due to lack of effect	Not specified

4.4.3 Narrative data review

Data were extracted from the 46 included studies into the pre-piloted data collection spreadsheet (see Appendix 4.4: Data extraction spreadsheet). A summary table illustrating the outcome and instruments by domain for each study is included in Appendix 4.6: Summary of data extraction and risk of bias assessment of included studies.

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, followed by randomised controlled trial with a small spread of other study types.

Table 4.6 Summary of study types

Study type	**Number of studies**
Randomised controlled trials	10
Case control studies	2
Prospective cohort studies*	23
Non-randomised, non-controlled intervention studies	3
Pilot efficacy / safety studies	5
Case series	3

* one of these was reporting baseline data for a cohort so is essentially a cross-sectional study

Numbers of participants and follow up:

The sample size of individual studies ranged from 4 (Salvarani et al., 2003) to 652 (Cawley et al., 2017)). Aside from the study by Cawley et al. (2017), all studies had <150 participants, with many having considerably smaller sample sizes. In longitudinal studies, follow up duration ranged from 4 weeks to 4 years. The study presenting baseline data from an inception cohort (Cawley et al., 2017), was cross sectional so follow up duration does not apply and a study by Cimmino et al (2008) was a one off assessment 5 years after the outset of a randomised controlled trial.

Age and gender of participants:

Most studies (n=40) reported mean age. This ranged from 62 to 78 years. Four studies reported median age, which ranged from 70 to 78 years. In two studies, age of participants was not reported. Most studies (n= 42) had a greater proportion of female participants. Four had more male participants; Cimmino et al. (2008) (47% F, sample size 57), Lally et al. (2016) (50% F, sample size 10), Palard-Novello et al. (2016) (33% F, sample size 18) and Devauchelle-Pensec et al. (2016) (35% F, sample size 20). One small study (n=4) had exclusively female participants (Salvarani et al., 2003).

Criteria used for diagnosis:

Table 4.7 shows the PMR classification criteria used in the included studies. Details of the different criteria have been described in Chapter 1 (1.6.1).

Table 4.7 Summary of PMR classification criteria used

Classification criteria used	Number of studies
Healy	9
Chuang	8
2012 EULAR / ACR	5
Bird	6
Jones and Hazelman	6
Clinician diagnosis / other	12

4.4.4 Outcomes measured

Only two of the randomised controlled trials measured an outcome in each of the core OMERACT recommended domains (Di Munno et al., 1995; Kreiner & Galbo, 2010). 12 cohort studies and one other interventional study measured an outcome in each domain. Two pilot studies and one case series also measured an outcome in each domain.

Findings with respect to frequency of domain measurement and instruments used for each domain are summarised in Table 4.8.

Laboratory markers of inflammation

Laboratory markers of inflammation were reported in 43/46 studies (93%) with ESR and CRP being the most frequently measured biomarkers.

Most studies measured both ESR and CRP (n=32), five measured only CRP and five measured only ESR. The five measuring only ESR were all from before the year 2000 whereas those measuring only CRP were all published after the year 2000. Some studies additionally measured IL-6 (n=10, one of these measured IL-6 as the only inflammatory marker), TNF alpha (n=1) or fibrinogen (n=6).

Pain

32/46 studies (70%) explicitly assessed pain. Seven of the remaining 14 included statements about measuring 'signs and symptoms' or including a 'physical examination' but without giving more detail. The most common instrument (n=29) used to assess pain severity was a visual analogue scale (VAS). Two studies used a numeric rating scale (NRS) and one used a physician assessment of pain severity graded 0-3. Two studies additionally assessed pain site using manikins.

Stiffness

28/46 studies (63%) included an assessment of stiffness as an outcome measure. In the majority of studies (26) this was duration of morning stiffness measured in minutes. Four studies additionally assessed stiffness severity using either a VAS or NRS and two assessed stiffness site using manikins. One study used a physician assessment of stiffness graded 0-3.

Physical function

22/46 studies (48%) assessed physical function in some way with several studies 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 Poymyalgia Rheumatica Activity Score (PMR-AS (Leeb & 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). Three studies used mental and physical components of the Short Form 36 (SF-36), one used a 'yes, no, don't know' question about ability to lift the arms above the head and one used the American Rheumatism Association functional class assessment.

Disease activity / global assessment

13/46 studies (28%) recorded PMR-AS, which is a composite measure of disease activity developed by Leeb and Bird (2004) comprising

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

CRP = C-reactive protein
MST = morning stiffness duration in minutes
VAS = visual analogue scale
EUL = elevation of the upper limbs

Six studies that didn't use the PMR-AS included a physician global assessment VAS. Ten 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 (20%) included some form of imaging in their outcome set. In most cases this was ultrasound of the shoulders but some used magnetic resonance imaging (MRI) or fluorodeoxyglucose-positron emission tomography computed tomography (FDG PET-CT). In five of the nine, the study's aim was at least in part to assess the utility of the imaging technique in the condition.

Other outcomes measured

Several studies (n=7) included outcomes such as 'number of relapses', 'duration of treatment' or 'cumulative steroid dose'. Some (n=10) used physical examination, presence of synovitis, fever or weight loss as part of their outcome set. Some studies assessed other blood parameters as particular to their aims e.g. other cytokines, HbA1c, ACTH / cortisol. Five studies assessed fatigue (by VAS, NRS or a question regarding time

to onset of fatigue for daily chores). Six studies included a 'back to normal' question or an unspecified questionnaire to assess health status. One study measured anxiety and depression using the Generalised Anxiety Disorder-7 (GAD-7) (Spitzer et al., 2006) and Patient Health Questionnaire-8 (PHQ-8) (Kroenke et al., 2009) questionnaires.

Ongoing or unpublished studies

There were five ongoing studies identified from trials registries for which outcomes were listed. Whilst there were no new outcomes amongst these (i.e. none that had not been identified in work already published), 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 more recently.

Table 4.8 Summary of core domain outcomes measured

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 (3) American Rheumatism Association functional class assessment (1)

4.4.5 Risk of bias within studies

Data for risk of bias assessment were extracted into a specifically designed spreadsheet and a judgement made as to whether the study was high, medium or low risk of bias for each domain, guided by the statements in the QUIPS tool (Hayden et al., 2013). The full table is in Appendix 4.5: Risk of bias assessment spreadsheet, and the overall judgement of risk of bias for each criterion is included in the summary table in Appendix 4.6: Summary of data extraction and risk of bias assessment of included studies, alongside the categorisation of outcomes and instruments.

Using the study participation domain as a marker of overall risk of bias, 13 of the 46 studies were judged to have low risk of bias. 25 were judged to have moderate risk of bias and 8 were felt to show high risk of bias.

Of the included RCTs, only two studies were judged low risk of bias in the study participation domain (Caporali et al., 2004; Kreiner & Galbo, 2010). Both of these were also judged low risk by the additional RCT criteria. Kreiner and Galbo (2010) measured outcomes in each of the core domains whilst Caporali et al. (2004) only measured laboratory markers of inflammation (ESR and CRP) of the core domains (they also recorded physical examination, relapses and recurrences, cumulative prednisolone dose and duration of treatment and an unspecified questionnaire to assess health status).

None of the three other interventional studies, 8/23 cohort studies and one case control study were judged low risk of bias by the participation domain.

There does not seem to be a pattern whereby higher quality studies (as judged by low risk of participation bias) assess more of the core OMERACT domains or particular outcomes

in each domain. Similarly, those judged high risk of bias did not measure noticeably different outcomes to studies in which risk of bias was lower.

4.5 Discussion

The majority of PMR studies identified for this review were cohort studies with only ten randomised controlled trials included. Almost all had sample sizes of less than 150 participants. This supports the general assertion that PMR is an under-researched condition. A wide variety of outcome measures was found to have been used in studies of PMR and they were often poorly defined. This makes comparing results across studies very difficult and prevents synthesis of current data to improve the evidence base.

Given that the OMERACT core domain set has only recently been established, it would not be expected that researchers working prior to this would necessarily measure outcomes in each domain. However, by stratifying the outcomes measured in studies to date against the now-established core domains, the degree of mismatch can be seen. Of the OMERACT core domains, the most frequently assessed was systemic inflammation, then pain, then stiffness, with physical function being least often measured. This conflicts with what might be expected to be most important to people with the condition and indeed with what has been described as important in qualitative studies (Mackie, Hughes, et al., 2015; Twohig et al., 2015).

Where inflammatory markers are used as outcomes in studies of PMR, ESR and CRP are usually both measured. In studies that chose one over the other there was a trend for more recent studies to use CRP rather than ESR.

Pain was the most commonly assessed patient-reported outcome in studies in this review with a VAS being the most frequently used measurement instrument. However, as noted in previous reviews (Duarte et al., 2015; Huang & Castrejon, 2016), there is little consistency in the question and scales used in the VAS and often no detail on the time frame being referred to in the assessment.

Stiffness was only measured in 28/46 studies in this review. Given that it is a cardinal symptom of PMR, this seems surprisingly low. In addition, where it was measured, the instrument used most frequently was 'duration of morning stiffness', which does not capture the full impact of stiffness as identified in qualitative work (Mackie, Hughes, et al., 2015; Twohig et al., 2015). It is interesting to note that two out of five of the as yet unpublished studies identified from searches of clinical trials registries are measuring severity of stiffness as well as duration of morning stiffness.

Physical function was assessed in the least consistent way of the core domains. Most frequently it was measured as part of an overall assessment of disease activity using the PMR-AS, which includes evaluation of 'elevation of the upper limbs' on a 0-3 scale. This is a very limited assessment of overall function and could be argued to be insufficient to represent this domain. The next most common measure of function was the HAQ but the availability of various forms of this (the mHAQ and the HAQ-DI) meant there was little consistency even within this group. Given that 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 is an important finding and highlights an area in need of improvement.

This review has also identified domains that have been frequently measured in studies of PMR, but which are not included in the core domain set. A significant minority of studies evaluated overall disease activity (how 'active' or well controlled a condition is) using the PMR-AS. Although this does not map directly onto a core domain as defined by OMERACT, the individual components of this composite score (with the exception of the physician global assessment, measured using a VAS) do cover each of the core domains. Whether they are the best measures of each of the domains is questionable as already discussed, but the frequency of use of the PMR-AS certainly influences the findings of the results for other domains.

There was a small number of studies measuring domains that were classified outside of the core domains but included in the 'important' or 'research agenda' list by the OMERACT 2016 group. These include fatigue, psychological impact and overall health status (assessed by a 'back to normal' question or unspecified questionnaire in some studies). Whilst these constructs are heavily intertwined, with each other and with pain, stiffness and function, it may be that this signifies a gap in the core domain set that needs addressing. An overall measure of the impact of PMR on a person's life could be of significant value in addressing this gap.

Strengths and limitations

The inclusive and comprehensive search strategy used for this review, including comparing reference list from other similar reviews, searching trials registries and contacting key experts in the field, is a significant strength in ensuring the aim of

identifying all outcome measures and instruments used in studies of PMR to date has been achieved.

The exclusion of papers considering PMR and GCA as a single group is a potential source of bias as relevant work may have been missed. However, as outlined in the methods, the justification for excluding these studies is strong and this outweighs the very small risk of having missed any outcome measure of relevance.

Some studies were excluded on the basis of not being available in full text or in a readily translatable language. Some of these were very old studies and others were in small foreign language journals or publications that are not available on-line. It is possible that relevant work may have been missed due to this but it seems highly unlikely from the titles of the these studies.

The assessment of risk of bias of included studies was important, as this had not been done before. It was interesting to note that there was no apparent relationship between the quality of the studies, as judged by risk of bias assessment, and number and range of outcomes measured. In terms of determining which instruments to evaluate further for inclusion in the core instrument set, it does not seem that the quality of studies using them to date is particularly informative. However, the fact that only half of the included studies were judged to be of low risk of bias is an important finding in appreciating the overall state of the evidence base for this condition and highlighting that further, high-quality studies are needed.

4.6 Conclusions

This comprehensive review has identified all the outcome measures and instruments used to date in studies of PMR and categorised them by the core domain set established by the PMR working group of OMERACT in 2016. The most commonly assessed domain was markers of systemic inflammation followed by pain, then stiffness then physical function.

Assessment of risk of bias of included studies was an added strength in this review as it had not been done previously. The findings here suggest that high quality research in this condition is still lacking. The ongoing work to develop a core outcome set for use in research studies is therefore of particular importance.

Domains measured in practice which are not included in the core domain set have also been identified and this will help inform the future research agenda.

The next step is to consider the existing evidence for the measurement properties of the instruments used for each of these domains to determine their suitability for inclusion in a core instrument set.

Chapter 5: A Systematic Review of the Measurement Properties of Instruments Used to Measure Outcomes in Studies of Polymyalgia Rheumatica

5.1 Background

Chapter 4 describes a systematic review of all the outcome measures and instruments used in existing studies of PMR, categorised into the OMERACT core domains for the condition. The next step in the process towards developing a core outcome set is to select candidate instruments for each of the domains and establish what is already known about the validity of the use of these instruments in PMR. This chapter describes the process used to do this, the findings and their implications.

5.1.1 The OMERACT Filter 2.1 instrument selection process

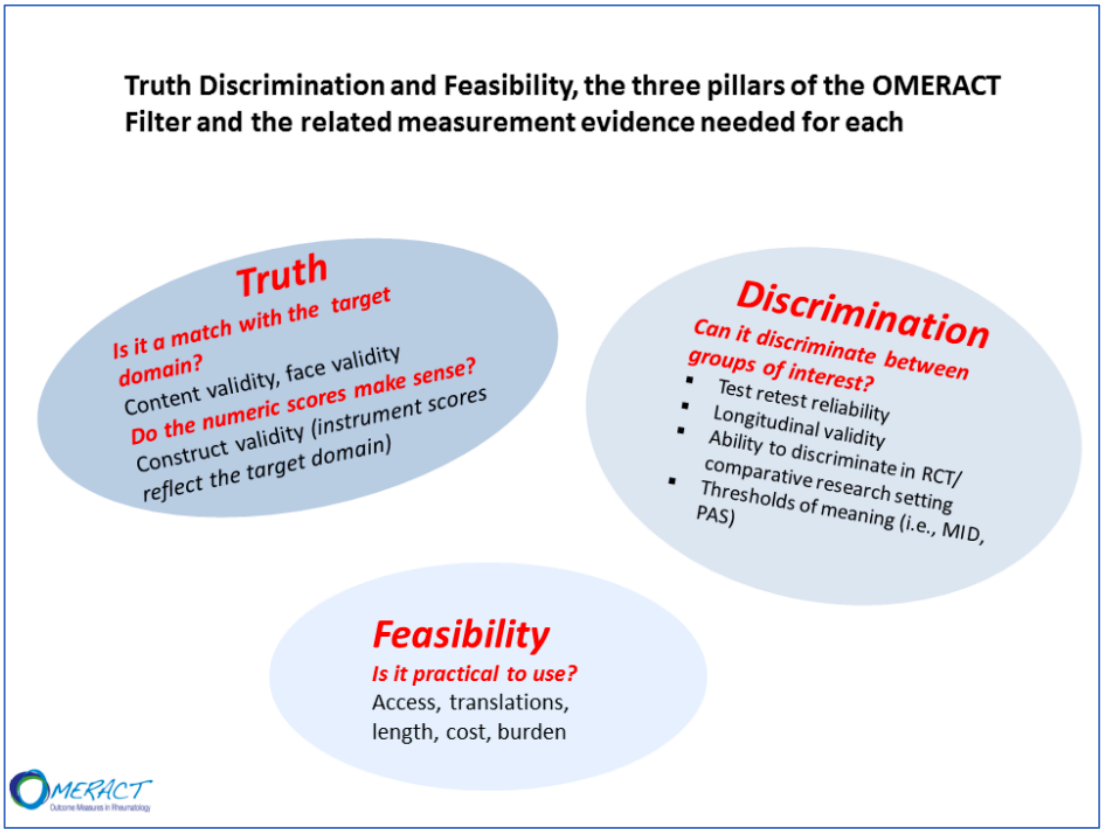
As outlined in Chapter 4 (4.1.1), OMERACT uses the three terms 'truth, discrimination and feasibility' when considering validity of a measurement instrument. The measurement properties related to each of these terms are shown in Figure 5.1.

The OMERACT Filter 2.1 instrument selection process is a framework to support finding, appraising and synthesizing the available evidence on measurement properties (Boers et al., 2019). The first part of this process involves the Working Group considering domain match (truth) and feasibility. Once they have agreed that an instrument meets these requirements and is worth considering further, a systematic literature review of the measurement properties (discriminative ability) of the candidate instrument in the

relevant condition needs to be carried out. The final stage in the process is to consider any areas where there are gaps in the evidence and design and carry out studies to provide the evidence needed.

Figure 5.1 Measurement properties considered within each of the three OMERACT pillars of evidence

(taken from the OMERACT Handbook (OMERACT, 2017, Chapter 4)



5.1.2 Selection of instruments to take forwards for review of measurement properties

Having identified instruments used to assess each domain, the PMR Working Group selected which of these to evaluate as candidates for the core outcome set. The decision-making process for this was based on the systematic review of the literature presented in

Chapter 4 along with; 1) discussion with other PMR working group members, 2) discussion with the wider rheumatology community at the American College of Rheumatology 2017 annual conference and 3) presentation of the work of the group to date and subsequent discussion at the OMERACT 2018 conference.

Laboratory markers of inflammation

ESR and CRP were selected for this domain. These are the most commonly used instruments in studies to date (used in 42/46 studies in my review) and the most widely available in research and practice. Other laboratory markers used to measure inflammation in some studies in our review, such as IL-6 and fibrinogen, are not typically used in clinical practice and not as readily available in all settings.

Pain

Of the 32 studies in our review that assessed pain, 31 used either a visual analogue scale (VAS) or numeric rating scale (NRS). These were the instruments that the group felt were the most appropriate to take forwards, given their wide acceptance and simplicity.

Stiffness

The most frequent measure of stiffness used in studies of PMR to date is assessment of duration of morning stiffness in minutes (used in 26 of the 28 studies that assessed stiffness in my review). However, qualitative work with people with PMR (Mackie, Hughes, et al., 2015; Twohig et al., 2015) has demonstrated that stiffness in PMR often does not follow the stereotyped inflammatory pattern of being worse in the morning and wearing off after a measurable time. It is actually frequently present throughout the day and has a more nuanced pattern of variability. Stiffness is closely linked to function and participants in these studies found it hard to rate severity of stiffness without a functional

context. Whilst still having limitations, an alternative measure of stiffness might therefore be a person's assessment of its overall severity (interpreted in whatever sense 'stiffness severity' means to them) measured by a VAS or NRS. The group therefore decided to evaluate both duration of morning stiffness and stiffness severity, measured using a VAS / NRS, for consideration in the core outcome measurement set.

Physical function

The two measures of function used most commonly in studies in my review were elevation of the upper limbs (EUL) on a 0-3 scale (used in 13 / 26 studies that evaluated function) and the Health Assessment Questionnaire (Fries et al., 1980) (HAQ, used in 12 / 26 studies that evaluated function).

EUL featured prominently because it is part of the composite PMR-AS, which is frequently used. As an isolated assessment of function however, it is clearly very limited as it only assesses one aspect of overall functional ability. It is not generally used as an assessment of function outside the context of the PMR-AS.

The HAQ was developed for use in in rheumatoid arthritis (Fries et al., 1980) and is the gold standard for measuring function in this condition (Maska et al., 2011). It is also widely used in other rheumatological conditions including PMR (Bruce & Fries, 2005). The full HAQ assesses five dimensions of disability, discomfort, drug side-effects, cost and death. The most commonly used version however is the HAQ-DI (disability index, often just referred to as "the HAQ" in the literature), which is the functional items and the VAS pain scale (Bruce & Fries, 2003). This HAQ-DI still comprises 41 questions, which can make clinical use difficult and so a modified HAQ (mHAQ) consisting of just eight items (one from each of the categories of the HAQ-DI) was developed (Pincus et al., 1983). This

mHAQ has also been validated in several rheumatological conditions and is widely used (Maska et al., 2011). The HAQ-DI and mHAQ were therefore chosen as the measures of physical function to evaluate further in this process.

5.2 Aim

The aim of this chapter is to systematically review the evidence in the literature on the measurement properties of candidate instruments for a core outcome set for PMR and determine whether it is sufficient to meet the OMERACT Filter 2.1 requirements of truth, discrimination and feasibility.

The candidate instruments for each domain are:

- Laboratory markers of inflammation: ESR and CRP
- Pain: visual analogue scale and numeric rating scale
- Stiffness: visual analogue scale and numeric rating scale and duration of morning stiffness
- Physical function: The Health Assessment Questionnaire (including the modified HAQ and HAQ-disability index)

5.3 Methods

5.3.1 Protocol development and registration

A single protocol (Appendix 4.1: Protocol for systematic review of outcome measures in PMR) was written to encompass both the review of outcome measures used in research studies of PMR (Chapter 4) and the review of measurement properties (current chapter).

The process and findings have been presented as two distinct chapters in this thesis for clarity.

As described in Section 4.3.1, the protocol was registered on PROSPERO with registration number CRD42017080058.

5.3.2 Search strategy

Only one database, Medline via Ovid SP, was searched for this part of the review. This was because comprehensive searches identifying all studies of PMR had already been carried out in five databases for the first part of the review and all relevant articles should have already been identified. When reviewing abstracts / full texts from the initial broad searches, articles relating to evaluation of measurement properties of an instrument were tagged in Covidence. Comparing the results of the more focussed search in Medline with the tagged studies from the first search (all those included for the full text review whether included or excluded at that stage) demonstrated that no articles had been missed. Searching the other databases again was therefore deemed unnecessary.

Search strategies for the focussed Medline search were developed following the guidance in the OMERACT instrument selection handbook (OMERACT, 2019) The common root of each of these strategies was the search strategy developed for PMR outlined in Chapter 4 (4.3.2). Different terms were then added to cover each of VAS / NRS and duration of morning stiffness, the HAQ / mHAQ and ESR and CRP.

Terms related to measurement properties were not built into the search strategy. This is because the limited amount of relevant literature meant that the articles identified from searching just on the condition and instrument could be easily screened to identify those

of relevance to measurement properties. Similarly, a single search was carried out for VAS / NRS encompassing the instruments for both pain and stiffness because it yielded few enough studies that relevant studies for each symptom could be identified easily.

The full search strategies are included in Appendix 5.1: Search strategies for evaluation of evidence regarding measurement properties of candidate instruments. As an example, the search related to VAS / NRS and duration of morning stiffness is shown in Table 5.1.

Table 5.1 Search strategy for PMR and VAS / NRS and duration of morning stiffness (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
11	((numeric* and rating and (scale or score)) or numeric scale or nrs or nprs).mp.
12	((visual and analogue and (scale or score)) or visual scale or VAS).af.
13	duration of morning stiffness.mp.
14	morning stiffness duration.mp.
15	11 OR 12 OR 13 OR 14
16	10 AND 15

5.3.3 Eligibility criteria and study selection

Studies were eligible if they included primary data evaluating one or more measurement properties of the candidate instrument in patients with PMR. For the purposes of this systematic review, I focussed on the measurement properties that OMERACT recommend are evaluated for each instrument. These are specified and defined in Table 5.2.

A more detailed discussion of properties of measurement instruments, in the context of PROM development and evaluation, is contained in Chapter 6.

Titles, abstracts and full texts were screened against the eligibility criteria to determine which studies to include in the review.

Table 5.2: Measurement properties to be considered and their interpretation according to OMERACT

(taken from the OMERACT handbook (OMERACT, 2019, Chapter 5))

Measurement property	Explanation
Construct validity	The degree to which the scores on the instrument relate to other measures (patient-report or clinical indicators) in a manner that is consistent with theoretically derived, a priori hypotheses concerning the domains that are being measured.
Test-retest reliability	A measure of the reproducibility of the instrument, that is the ability to provide consistent scores over time in a stable population.
Responsiveness / longitudinal construct validity	The extent to which an instrument can detect changes in the domain of interest over time, when they have occurred.
Thresholds of meaning	The degree to which one can assign an easily understood meaning to the scores from an instrument. Rates of achievement are compared between arms in clinical trials. This includes thresholds like a minimum important improvement, or a patient acceptable symptom state.

5.3.4 Data extraction

The following variables were extracted for each included study:

- Author, year
- Title and aim
- Measurement properties evaluated
- Methods used to evaluate the measurement properties
- Findings in relation to the measurement properties

One other review team member (CO) and I extracted the data independently for all studies and resolved any discrepancies through discussion.

5.3.5 Quality appraisal

Strengths and limitations of each study were evaluated using the OMERACT Good Methods Checklist (OMERACT, 2019). This sets out quality standards for each measurement property as summarised in Table 5.3. Each study was given a rating to signify whether it should be used as evidence for the measurement property being 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). CO and I assessed the strengths and limitations independently and then resolved any discrepancies through discussion. Results of this quality 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.

Table 5.3: Quality criteria for each measurement property

Taken from the OMERACT Good Methods Checklist (OMERACT, 2019)

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, predictive 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 characteristics curve
AUC – area under the curve

5.4 Results

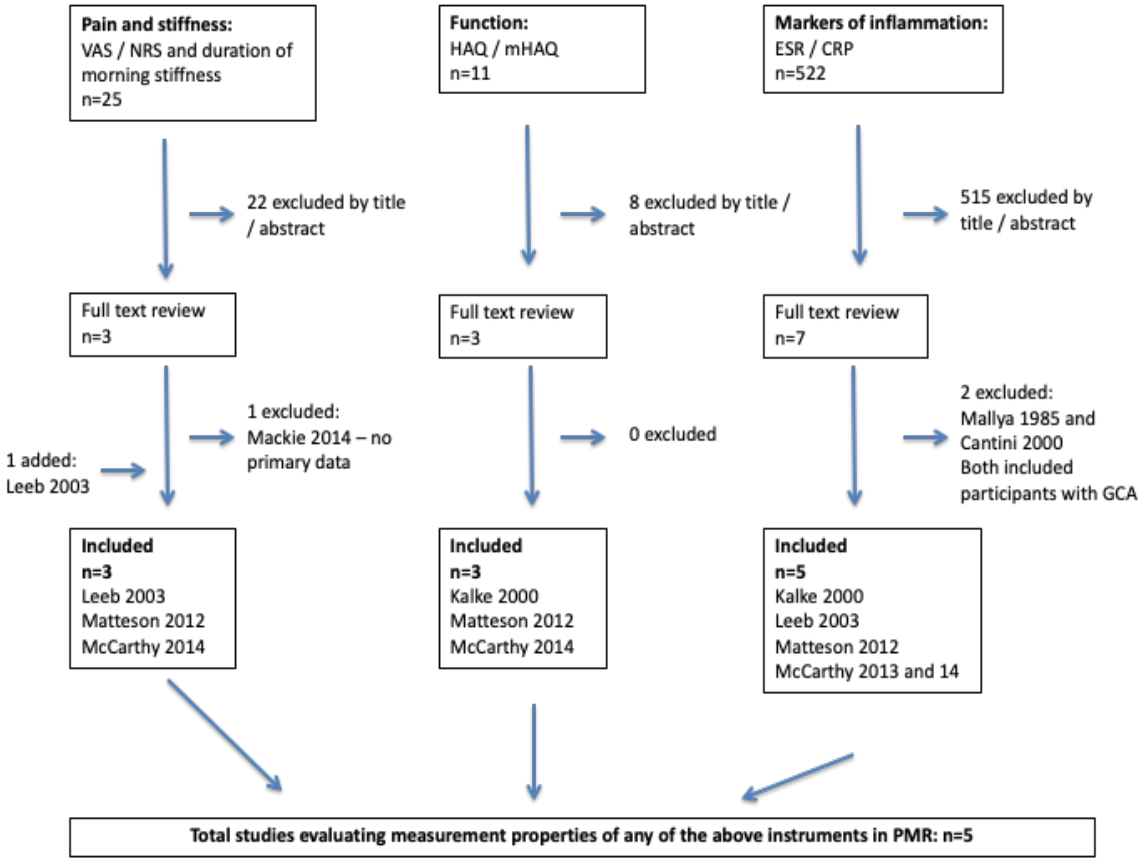
5.4.1 Search results

Searches were carried out in Medline via OVID in July 2018. A flow diagram of the search results is given in Figure 5.2.

In total, five studies were identified that met the inclusion and exclusion criteria.

One study, (Leeb et al., 2003), was not identified by the search for VAS pain or duration of morning stiffness but did appear in the results of the search pertaining to ESR and CRP (and it had also been identified in our initial comprehensive searches and flagged as a paper on measurement properties). On review of the full text, it was apparent that the study reported information about properties of the pain VAS and duration of morning stiffness in PMR as well as inflammatory markers and it was therefore added to this section.

Figure 5.2 Flow chart of the study selection process



5.4.2 Summary of included studies

A summary of the key characteristics of the five included studies is given in Table 5.4 ahead of discussion of their findings and critical appraisal.

Table 5.4 Summary of the characteristics of the included studies

Lead author / year / title	Aim	Measurement properties evaluated	Participants and methods
Kalke 2000 A study of the HAQ to evaluate functional status in PMR	To evaluate the HAQ as an assessment of functional status in PMR	**Construct validity** – change in function with treatment and correlation with other measures. **Responsiveness** – to change with treatment	Participants - new-onset PMR diagnosed using Jones and Hazelman criteria. n=18 Treated with reducing regimes of either oral prednisolone or i.m. methylprednisolone. Outcomes measured at baseline then 6, 12 and 24w after treatment started. The HAQ was compared to duration of MS, pain VAS and CRP using a linear regression coefficient, one-way analysis of variance and the standardized response mean.
Leeb 2003 EULAR response criteria for PMR: Results of an initiative of the European Collaborating PMR group	To develop response criteria for PMR for monitoring treatment and comparing alternative treatment regimens.	**Construct validity** – change in disease activity with treatment, correlation of other measures to VAS pain.	Participants – multisite study, PMR diagnosed by Bird / Wood criteria. Treated with corticosteroids in a reducing regime determined by the local investigator. n=76 Outcomes assesses at 4, 8, 16 and 24w. Change in each parameter between baseline and 24w was analysed using ANOVA for repeated measures.
Matteson 2012 Patient-reported outcomes in PMR	To report the disease course of PMR and to prospectively evaluate the performance of various clinical, laboratory and PRO measures and MSK US findings in patients with PMR with a view to	**Construct validity** – change in disease activity with treatment **Test-retest reliability** **Thresholds of meaning –** smallest detectable	Participants - new onset PMR (defined by explicitly stated criteria) and treated with oral steroids according to a standard treatment protocol. n=85 Outcomes measured at baseline and 1, 4, 12 and 26w after starting treatment. Comparison between patients at different time points was performed using paired T-tests.

	identifying a minimum set of outcomes measured to be used in practice and future clinical trials	difference (SDD) and minimal detectable change (MDC)	Test re-test reliability was assessed by calculating the intra-class correlation coefficient from 14 patients who had <10mm change on overall PMR VAS between baseline and week 1. SDD and MDC calculated in this group of 14. The SDD as a percentage of the maximum score gave the %MDC.
McCarthy 2013 Plasma fibrinogen is an accurate marker of disease activity in patients with PMR	To establish whether plasma fibrinogen was a superior marker of disease activity in active PMR than the standard biomarkers ESR and CRP.	**Construct validity** – change in disease activity with treatment **Thresholds of meaning –** discrimination between active and inactive disease	Participants - PMR diagnosed by Jones and Hazelman criteria. Defined as inactive disease if there was an absence of symptoms on a stable dose of steroids or on no treatment for 6w prior to review. n=25 active (one with GCA in addition). n=35 inactive Measurements (PMR-AS, fibrinogen, ESR and CRP) made at baseline and 6w. Comparison to pre-treatment levels calculated using Wilcoxon signed-rank test. Ability of biomarkers to detect response to treatment in the active group was calculated using ROC's, predictive values, likelihood ratios and sensitivity and specificity at different cut-off values. Ability of biomarkers to detect disease remission was calculated by dividing all participants into remission or persistent disease using the PMR-AS and then using same stats as above.
McCarthy 2014 Plasma fibrinogen along with patient-reported outcome measures enhances management of PMR: A prospective study	To assess the responsiveness of various PRO measures to changes in disease activity and their correlation with the traditional laboratory measures of disease activity, ESR and CRP, as well as plasma fibrinogen.	**Construct validity** – change in disease activity with treatment and correlation between instruments **Responsiveness** – to change with treatment	Same study protocol as McCarthy 2013. Also tested a VAS for disease activity and a VAS for QoL. Standardised response means and effect size statistics calculated for all biomarkers. Spearman's rank correlation coefficient used to compare the biomarkers to the PROs.

HAQ = health assessment questionnaire, mHAQ = modified health assessment questionnaire, w = week, i.m. = intramuscular, QoL = quality of life, VAS = visual analogue scale, PRO = patient reported outcome, ANOVA = analysis of variance

5.4.3 Results of the evaluation of measurement properties and critical appraisal of included studies

A summary of the key findings and the quality assessment of the five included studies is presented in Table 5.5. The findings for each instrument are discussed below.

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 (myalgia, elevation of the upper limbs) demonstrated by Leeb et al. (2003) and 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, McCarthy et al. (2014) and Kalke et al. (2000). In Kalke et al., the standardised response mean (SRM) was reported, calculated as mean change score divided by the standard deviation of the respondent's change score. McCarthy et al. (2014) reported the SRM along with the effect size statistic (ESS) but the methods used to calculate these were not explicitly stated. Neither study stated hypotheses about the anticipated change in response to treatment nor 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. (2012). The methods were appropriate, and the result suggests good reliability (ICC = 0.82) but the small sample size (n=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). 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.

No studies evaluated any measurement properties of a numeric rating scale for pain.

Duration of morning stiffness

The four studies that evaluated measurement properties of pain VAS (Kalke et al., 2000; Leeb et al., 2003; Matteson et al., 2012; McCarthy et al., 2014) all also evaluated duration of morning stiffness. The limitations to the methods discussed above also applied for this outcome measure and test-retest reliability was poorer (ICC = 0.11). All were rated red for all measurement properties.

The Health Assessment Questionnaire (Disability Index)

One study (Kalke et al., 2000) specifically aimed to evaluate the HAQ as an assessment of functional status in PMR but significant limitations meant that it was rated red for both measurement properties it evaluated (construct validity and responsiveness).

Construct validity was evaluated by studying correlation with 'traditional measures of disease activity in PMR'. The comparator measures were duration of morning stiffness, pain VAS and CRP, none of which are measures of function. The correlation between the HAQ and the three other outcome measures was good (>0.6 in each case) but no hypotheses about the magnitude of change or strength of correlation were stated. The statistically significant reduction in the HAQ score in response to treatment is some further evidence towards its construct validity but again, no hypotheses were set out in advance in relation to this.

Responsiveness was evaluated using the SRM. The SRM was higher for the HAQ (at 3.0) than for the other measures in this study suggesting greater responsiveness to change but no *a priori* hypotheses were stated.

The Modified Health Assessment Questionnaire (mHAQ)

Two studies evaluated measurement properties of the mHAQ (Matteson et al., 2012; McCarthy et al., 2014). Across the two studies, all the measurement properties were considered but the studies were rated red for all except test-retest reliability.

Both Matteson et al. (2012) and McCarthy et al. (2014) provide some evidence towards construct validity of the mHAQ in that they demonstrate improvement in response to treatment. McCarthy et al. also demonstrated correlation of the mHAQ with other outcome measures though again, these comparator measures were not actually measures of function.

The McCarthy et al. (2014) study evaluated responsiveness of the mHAQ, reported as SRM and ESS. The results for the mHAQ in this study population were 1.36 and 1.65

respectively but although the statistical methods were appropriate, no hypothesis about the expected magnitude of change was given.

Test-retest reliability of the mHAQ was evaluated in Matteson et al. (2012). The ICC was 0.72, which is within the range deemed as demonstrating adequate performance for this measurement property (OMERACT, 2019) but the small sample size prevented the study being rated green.

The SDD and MDC were calculated in the same study but the limited information on the methods used and the fact that there was no attempt to determine a minimally important difference to patients, meant that it did not meet the required OMERACT standards for evaluation of thresholds of meaning.

ESR and CRP

Construct validity was evaluated in three studies (Leeb et al., 2003; Matteson et al., 2012; McCarthy et al., 2014). All three studies demonstrate that ESR and CRP improved with the treatment of PMR but none of these studies set out *a priori* hypotheses about this. McCarthy et al. (2014) examined the correlation between the mHAQ and ESR / CRP and found moderate correlation (Spearman's rank correlation coefficient 0.39 / 0.42) but these instruments cannot be said to be measuring the same construct. None of the studies met all the criteria set out in the OMERACT good methods checklist for construct validity and all three were rated red.

Responsiveness was evaluated in two studies (Kalke et al., 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 of ESR and CRP and was rated amber. The study considered the ability of ESR and CRP to detect active disease and disease remission by calculating sensitivity, specificity, positive predictive values and likelihood ratios. Active PMR was defined clinically at diagnosis and disease remission was defined by the PMR-AS. CRP was found to be better than ESR in distinguishing active disease and disease remission.

Table 5.6 shows a summary of the overall evidence for the measurement properties of each instrument in PMR.

Table 5.5 Summary of measurement properties evaluated for each instrument and quality assessment of the included studies

Instrument	Measurement property	Studies	Quality assessment	Findings	Rating
Pain VAS	Construct validity	Leeb 2003	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	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 *	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	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	Small sample size, n=14 Patients were stable in the interim time period; the time period was appropriate and test conditions were stable. Statistical methods were appropriate (ICC)	Global pain ICC = 0.82	Amber
	Thresholds of meaning	Matteson 2012	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	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	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	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.	SRM = 0.89 ESS = 0.96	Red

			Statistical methods were appropriate but no hypotheses about magnitude of change were made.		
		Kalke 2000	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	Small sample size, n=14 Patients were stable in the interim time period; the time period was appropriate and test conditions were stable. Statistical methods were appropriate (ICC)	ICC 0.11	Red
	Thresholds of meaning	Matteson 2012	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	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	Small sample size, n = 18	SRM = 3	Red

			Situation of change clear – newly diagnosed, started on treatment. Statistical methods are appropriate but no hypotheses about magnitude of change were made.		
mHAQ	Construct validity	Matteson 2012	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	Each instrument was compared to its pre-treatment levels. Comparator measures were not evaluating the same construct. No *a priori* hypotheses about magnitude of change or correlation with other instruments stated.	Statistically significant improvement between W1 and W6 in the active group. Correlation coefficients between mHAQ and PMR-AS, ESR and CRP were 0.68, 0.45 and 0.39 respectively	Red
	Responsiveness	McCarthy 2014	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	Small sample size, n=14	ICC = 0.72	Amber

			Patients were stable in the interim time period; the time period was appropriate and test conditions were stable. Statistical methods were appropriate (ICC)		
	Thresholds of meaning	Matteson 2012	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	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	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 between baseline and W1 and W1 and W4 but not between W4 and W26.	Red
		McCarthy 2014	Each instrument was compared to its pre-treatment levels. Comparator instrument for correlation was the mHAQ which measures a different construct.	Statistically significant improvement from W1 to W6 in the active group Correlation coefficient between mHAQ and ESR / CRP = 0.45 / 0.39	Red

			No explicit *a priori* hypotheses about magnitude of change or correlation with other instruments stated		
	Responsiveness	McCarthy 2014	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	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 *	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.	

VAS = visual analogue scale, W = week, SRM = standardised response mean, ESS = effect size statistic, ICC = intra-class correlation coefficient, SDD = smallest detectable difference, MDC = minimum detectable change, PPV = positive predictive value

*these two papers were from one group of 60 participants, of whom one participant had biopsy-proven GCA in addition to 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**	**-**	**+**

Table 5.6 Summary of quality of evidence on measurement properties of outcome measurement instruments in PMR

5.5 Discussion

This review highlights how little work has been done to evaluate instruments used as outcome measures in studies of PMR. Only five distinct studies addressing this question were identified, some evaluating more than one instrument but in combination still only addressing a subset of the important measurement properties.

Crucially, none of the studies were rated 'green' for any of the measurement properties when assessed against the OMERACT good methods criteria. For pain VAS and the mHAQ there was one study of test-retest reliability that achieved amber and there was one study considering thresholds of meaning for ESR/CRP which was also rated amber.

Given the very limited evidence from studies of people with PMR supporting the use of these outcome measures in the condition, consideration of work done in other conditions is warranted.

5.5.1 Comparison with evidence for use of the instruments in other rheumatological conditions

Pain and stiffness

OMERACT produced a report of relevance in 2015 entitled 'Optimal Strategies for Reporting Pain in Clinical Trials and Systematic Reviews: Recommendations from an OMERACT 12 Workshop' (Busse et al., 2015). This paper reviewed the evidence about reporting of pain measurement and produced the recommendations summarised in Table 5.7. The 10cm pain VAS was suggested to be the preferred measure although it is acknowledged that pain should be reported in a context of other related domains such as

physical and emotional functioning. The main measurement property of the pain VAS discussed in this report is the Minimally Important Difference, which is reported to have been established to be approximately 1cm, regardless of pain severity (Busse et al., 2015; Dworkin et al., 2008).

Table 5.7: Summary of recommendations for reporting pain from the OMERACT 12 workshop

(taken from Busse et al. (2015))

Pain should be reported directly by patients
Global assessments of pain are preferable to assessment of multiple components of pain
The effect on pain should be accompanied by presentation of treatment effects on other patient-important outcomes.
Individual trials should report the proportion of patients achieving a percentage reduction from baseline pain, a desirable pain state, and / or a combination of change and state.
Meta-analyses should convert all continuous measures for pain to a 10cm VAS for pain, report the pooled mean change and the pooled mean change divided by the minimal (1cm), appreciable (2cm) and substantial (5cm) difference in pain improvement.
To further increase interpretability the pooled estimate on the VAS for pain should be transformed to a binary outcome and expressed as relative risk and risk difference using these same thresholds.

A recently published systematic review (Halls et al., 2017) considered the measurement properties of stiffness assessment tools in people with rheumatoid arthritis (RA). 25 studies were identified for inclusion in the review and between them they evaluated 52 different PRO measures of stiffness. 51 of these focussed on morning stiffness and the majority were concerned with duration of stiffness with less than half asking about severity or intensity. The majority had a response item of duration in minutes but some used a VAS and some used a NRS. With regards to measurement properties, there was no evidence regarding face or content validity of stiffness items and limited and

inconsistent evidence regarding criterion and construct validity, reliability or responsiveness. Severity items performed better than duration items in relation to construct validity, discrimination between disease states and responsiveness.

Physical function

The measurement properties of the HAQ have been studied most heavily in RA where it has been shown to have high test-retest reliability (ICC 0.95) and internal consistency (Cronbach's alpha 0.9) (Maska et al., 2011). Good correlation has been demonstrated with physical capacity measures, observed functional performance and clinical and laboratory measures of inflamed joint counts and inflammatory markers. Minimally clinically important differences for HAQ scores have been suggested to be around 0.22 but this varies widely depending on population and construct used (Maska et al., 2011). The mHAQ and the HAQ have been shown to be highly correlated but the mHAQ has slightly lower average scores than the full HAQ (Anderson et al., 2010). The mHAQ has also been shown to have good test-retest reliability and correlation with other physical function measures. Its ability to detect change is similar to the HAQ but it has a higher floor effect of up to 25%, which is a significant limitation (Maska et al., 2011).

Laboratory markers of inflammation

ESR and CRP are used ubiquitously in 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 at baseline indicating that the relationship between these biomarkers and PMR disease activity is not straightforward (Cantini et al., 2000).

A review of the literature on the truth, discrimination and feasibility of ESR and CRP in ankylosing spondylitis (Ruof & Stucki, 1999) found some evidence that both were correlated with disease activity but inconclusive data on their ability to discriminate between active and inactive disease states. Neither measure was found to be clinically superior in terms of validity and the authors concluded that feasibility aspects are most relevant in terms of choice of measure.

Similarly, a study of correlation of ESR and CRP with joint inflammation in RA (as measured by synovial biopsy from an affected joint) found that although there was a positive relationship between levels of the inflammatory markers and joint inflammation (stronger with CRP than with ESR), a significant proportion (49%) of those with a normal CRP had evidence of inflammation on synovial biopsy of the painful joint (Orr et al., 2018).

5.5.2 Strengths and limitations of this review

The comprehensive search strategy used in this review, borne out of the two-step approach of initially identifying all studies in which outcomes of PMR were measured before focussing specifically on studies of measurement properties, means that it is a thorough assessment of the literature to date. Through following the OMERACT process

for evaluation of evidence on measurement properties, and working with the PMR working group review team, the process was systematic and explicit.

One potential source of criticism is the decision to include the two studies by McCarthy et al. (2013, 2014) in which one participant out of 60 had biopsy-proven GCA as well as PMR. 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.

5.6 Summary and conclusions

The aim of this review was to evaluate the literature on the measurement properties of selected instruments in PMR and determine if there is sufficient evidence to meet the OMERACT Filter 2.1 criteria.

Pain VAS, morning stiffness duration, stiffness severity VAS, the HAQ and mHAQ and the biomarkers ESR and CRP were evaluated.

Pain VAS has some evidence supporting its discriminative abilities in PMR. It is recommended by OMERACT as the measure of choice in reporting pain in clinical trials (Busse et al., 2015). Work to evaluate its face validity and feasibility in patients with PMR is underway and it is likely to be deemed satisfactory in relation to these considerations.

Duration of morning stiffness has been shown in this review to have poor discriminative properties in PMR, mirroring previous findings in RA. Qualitative studies about stiffness in RA and PMR support the assertion that this measure has limited value. Measures of stiffness severity have better supporting evidence in RA and from qualitative work but

there are no studies evaluating measurement properties of a stiffness severity VAS in PMR.

The HAQ and mHAQ have some weak evidence supporting their discriminative properties in PMR. There is better evidence for these measures in other conditions but whether this is transferable is questionable. Again, there is work underway to evaluate face validity and feasibility of the HAQ in people with PMR, which will help decide ultimately whether it satisfies the Filter 2.1 criteria.

ESR and CRP have some evidence supporting their responsiveness and thresholds of meaning in PMR and CRP appears to be more responsive than ESR. There is no information about their test-retest reliability in this condition. However, they are the cheapest and most widely available biomarkers and therefore best fulfil the Filter 2.1 requirements for the pathophysiological domain.

The OMERACT Filter 2.1 instrument selection process requires consideration of three pillars of evidence – truth, discrimination and feasibility. The literature identified in this review focussed on the discriminative abilities of the instruments being evaluated and found that none of the instruments have robust evidence derived from studies of people with PMR that supports their use. Evidence from other rheumatological conditions could pragmatically be used to help inform choice of instruments. This may be reasonable for domains such as pain and stiffness and for biomarkers but is less logical when considering physical function as functional limitations are likely to differ considerably between different conditions.

In conclusion, this review has found that the instruments used in clinical studies of PMR currently do not have good evidence to support their use. Further studies are needed to evaluate existing measures and to develop new measures to fill gaps in our ability to

assess the condition. Given the heterogenous nature of PMR, and the fact that the manifestations of the condition (pain, stiffness, fatigue, functional impairment) are best measured from the patient perspective, a multi-dimensional patient-reported-outcome-measure could be a useful tool to fill this gap.

Chapter 6: Methodology of PROM Development and Evaluation

6.1 Introduction

The preceding chapters have identified the gap in our ability to assess the impact of PMR on patients and the principles and applications of a type of measurement tool, a patient reported outcome measure (PROM), which has potential to fill this gap.

The subsequent chapters will describe the research undertaken to develop a new PROM for PMR.

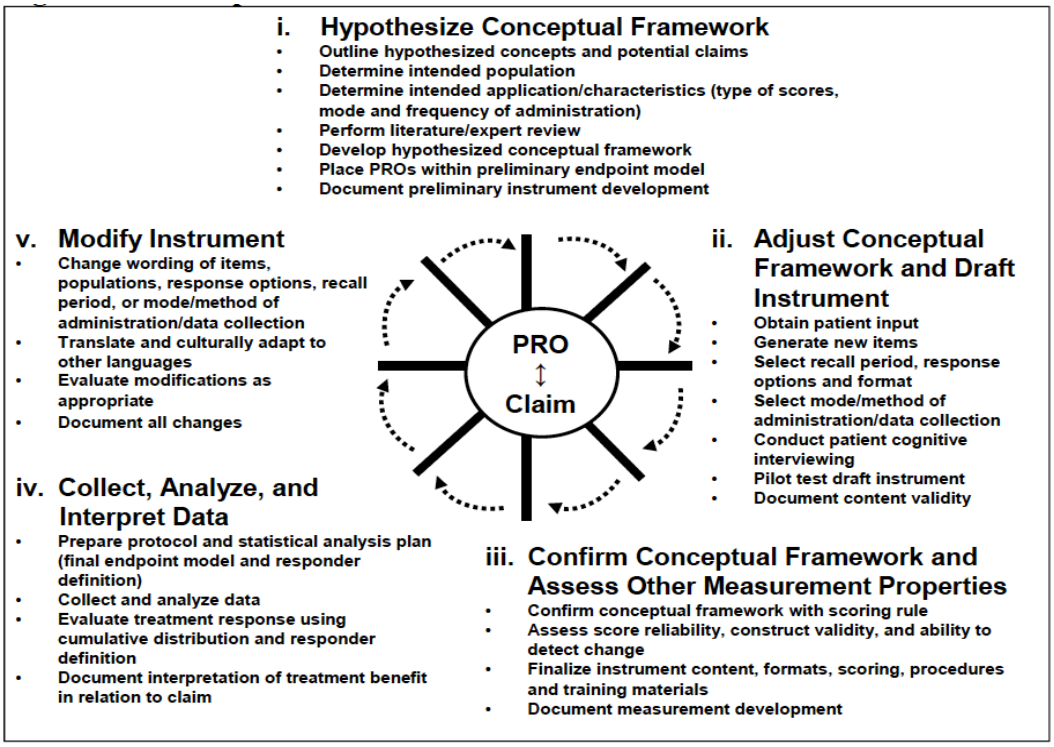
This chapter focuses on the methodology of PROM development, building on the principles introduced in Chapter 2. I shall present a broad overview of the process of PROM development and then discuss the methodology of each stage in turn, linking the principles to my particular aim of developing a PROM for PMR.

6.2 Overview of PROM development

There is no single standard process for PROM development and details will vary according to the scientific discipline and the intended use of the instrument. However, the broad methodological approach for development of any PROM is similar and illustrated by the examples in Figure 6.1 and Figure 6.2. A key feature of the development process, common to both models, is that it is iterative and circular – the steps are intertwined and may need to be repeated in a continuous process of evolution and adaptation.

Figure 6.1: FDA model of development of a PRO instrument

(taken from Guidance for Industry, PROMs: Use in Medical Product Development to Support Labeling Claims (U.S Department of Health and Human Services, 2009))



As discussed in section 2.6, much of the literature and guidance about PROMs focuses on their evaluation. Indeed, whilst the model in Figure 6.1 is a model of PRO 'development', it is taken from a guidance document on how the U.S. Food and Drug Administration (FDA) reviews and evaluates PRO instruments used to support claims in medical product labelling (U.S Department of Health and Human Services, 2009). Evaluative guidance, including the COSMIN taxonomy and checklist (Mokkink et al., 2010b), can however, be useful to inform new PROM development.

Figure 6.2: Overview of the steps in the development and evaluation of a measurement instrument

(adapted from (de Vet, Terwee, et al., 2011b)

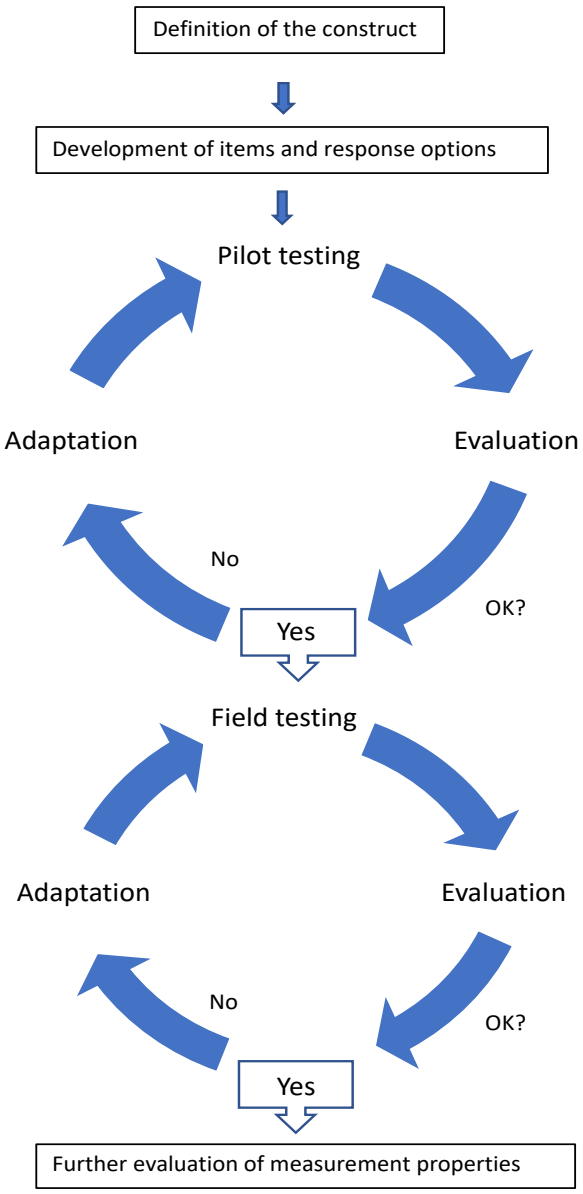
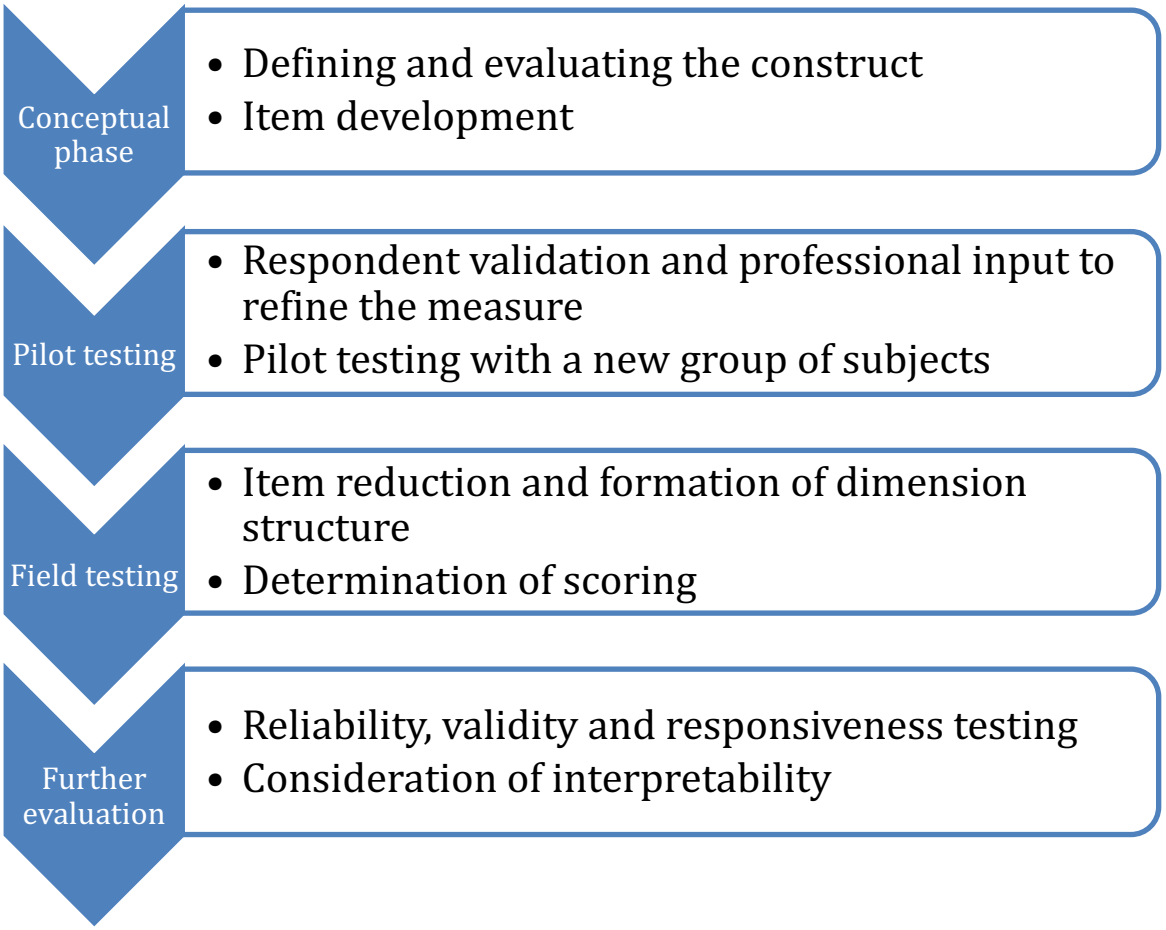


Figure 6.3 provides an overview of the methodological process I will follow. It aligns with the broad principles of the model in Figure 6.2 but contains additional details from the FDA model (Figure 6.1) and from wider reading of guidance on PROM development and

evaluation (de Vet, Terwee, et al., 2011b; Mokkink et al., 2018; OMERACT, 2019; Reeve et al., 2013).

Figure 6.3: Overview of the process of PROM development adopted in this thesis



6.2.1 COSMIN versus OMERACT

The COSMIN taxonomy and checklists for assessing the quality of studies on measurement instruments (Mokkink et al., 2010b, 2019; Prinsen et al., 2018) have been widely accepted and adopted internationally. OMERACT differs in in its use of

terminology and methodology to COSMIN in some areas, for example, where COSMIN use the term 'responsiveness', OMERACT use 'longitudinal construct validity'. Whilst I have used the OMERACT framework for my systematic review of outcome measures and instruments used in studies of PMR (Chapter 4 and Chapter 5), OMERACT is specific to rheumatological conditions and is less widely recognised than COSMIN. In developing a new measure, I therefore decided to use the COSMIN terminology in order to ensure the highest standards of rigour and make its development methods as easily understood and widely accepted as possible.

6.3 Defining the construct and developing a conceptual framework

6.3.1 Defining the construct

Being clear about 'what' is to be measured is essential before further consideration can be given to developing an instrument.

In the case of PMR, I have demonstrated that there is a gap in our ability to assess the effects of PMR on patients' lives from their own perspective. This view is supported by the conclusions of the 2017 Lancet seminar paper on the condition (González-Gay et al., 2017), which state that in future "management recommendations must focus on patient perspectives" and emphasised by the research agenda set out in 2015 EULAR / ACR guidelines in which "identifying which outcome measures, including patient-related outcomes….should be used in PMR" is the first item in the list of research priorities. (Dejaco, Singh, Perel, Hutchings, Camellino, Mackie, Abril, et al., 2015). I therefore want to develop an instrument to measure the impact of PMR on a person's life; a construct

that could be termed 'PMR-related quality of life'. This construct maps to the broader term, Health Related Quality of Life (HRQoL).

Quality of Life

Quality of life (QoL) is an intangible, amorphous concept that theoretically incorporates all aspects of a persons' life. Arguments about concepts such as happiness, human needs and desires and the 'good life' have been debated by social scientists and philosophers dating back to Aristotle (384-22 BC). Today, discussion of the nature of QoL and how to assess it spans many disciplines including philosophy, economics, geography, literature and the medical and social sciences. In one widely cited article on the nature of quality of life, Edlund and Tancredi (1985), state "*The general conclusion is that perception and achievement of quality of life is dependent on an individual's preferences and priorities in life. The meaning of the concept of quality of life is thus arguably dependent on the user of the term, their understanding of it, and their position and agenda in the social and political structure*".

Amongst the wide-ranging discussions of the meaning of QoL, one point that has consensus is that health is an important component dimension. From the earliest studies of indicators of QoL, health has repeatedly ranked amongst the most important areas of QoL and is consistently identified as one of the most valued states (Bowling, 2001, Chapter 1). The need to focus on health aspects of QoL in medical and social science research and be able to operationalise this in studies, has led to the development of the narrower concept of 'health-related quality of life'.

Health-related Quality of Life

The International Society for Quality of Life Research (ISOQOL) define HRQoL as:

"the functional effect of a medical condition and / or its consequent therapy upon a patient. It is subjective and multidimensional encompassing physical and occupational function, psychological state, social interaction and somatic sensation." (International Society for Quality of Life Research, 2018).

It is a narrower concept than overall QoL, which would include other considerations including housing, finances and a person's perceptions of their environment (Bowling, 2001). It is, however, a broader construct than 'health status' or 'functional status' in that it incorporates a patient's subjective experiences, perceptions and values in relation to their overall well-being (Crosby et al., 2003) . It also incorporates the patient's level of satisfaction with treatment, outcome and health status and with future prospects (Bowling, 2001, Chapter 1).

Instruments measuring HRQoL need to capture the degree to which a medical condition or its treatment impact the patient's life in a valid and reproducible way. They can then be used to measure changes in HRQoL over time (in clinical trials, observational studies or healthcare delivery settings) or to compare the HRQoL of patients with different conditions or who receive different treatments (clinical trials or comparative effectiveness research).

6.3.2 Exploring the construct

Having determined the construct, consideration needs to be given to its component dimensions and to the purpose of the instrument.

For a multidimensional instrument it may be possible to determine specific aspects to consider, such as physical and mental functioning, during the conceptual phase and these can then be explored further with patients as the conceptual framework is developed (see below).

The target population and the application of the instrument needs to be considered and specified. PROMs can have three broad measurement objectives – discriminative, evaluative or predictive (Kirshner & Guyatt, 1985). Instruments described as discriminative are used to discriminate between people with different levels of disease burden, to assist diagnosis or severity stratification, whereas evaluative instruments aim to evaluate the impact of a disease and effects of treatment over time. Predictive instruments are aimed at determining prognosis and are often called prediction models rather than measurement instruments, as they typically contain a number of different constructs and variables (de Vet, Terwee, et al., 2011b). An instrument may have more than one measurement objective, but the purpose needs to be clear as it influences the development of the instrument - items and scales need evaluating according to whether they meet the specified objectives (Kirshner & Guyatt, 1985). For example, evaluative instruments need to be responsive, while instruments used for discriminative purposes do not (Mokkink et al., 2010a)

The conceptual framework

HRQoL is not directly observable and therefore needs to be measured indirectly by measuring observable characteristics that relate to the construct. The 'conceptual framework' is the term used to describe the relationship between the measurable

characteristics and the non-observable construct (de Vet, Terwee, et al., 2011a) and it is often depicted visually as a diagram.

A conceptual framework can be hypothesised from reviewing the literature and consulting with experts, but it needs refining with input from people with lived experience of the construct being measured.

Conceptual models for PROMs can either be reflective or formative (Edwards & Bagozzi, 2000). In a reflective model, all of the items are manifestations of the same construct and are highly correlated (they are 'effect indicators'). For example, the construct 'anxiety' is reflected by items including worry, panic and restlessness. Any change in the construct is expected to affect all items. In a formative model, the items 'form' the construct and they are known as 'causal indicators'. For example, the construct 'stress' is formed by items including bereavement, divorce and job loss. In this type, a change in the construct does not necessarily mean a change in all the items. The distinction is important because concepts such as structural validity and internal consistency are only applicable to reflective models (Mokkink et al., 2019). Some complex constructs, such as QoL, can combine reflective and formative elements (de Vet, Terwee, et al., 2011a).

Measurement theories

In addition to describing the theoretical relationship between the items and the construct (the conceptual framework), it is also useful to be able to describe the statistical relationship between the items and the construct. For unobservable constructs measured by multi-item instruments, this statistical relationship is described by a measurement theory - a theory about how the scores generated by the items represent

the constructs to be measured (de Vet, Terwee, et al., 2011b). Two of the most well-known measurement theories are classical test theory and item response theory and these are discussed in detail below (6.6.1) and in Chapter 8 (Section 8.3).

Exploring and refining the construct for the PMR-PROM

My PROM, aimed at measuring the unobservable, complex construct PMR-related QoL (henceforth referred to as the PMR-PROM), needs to be applicable to people at all stages of the PMR disease process and it will have evaluative and discriminative applications.

The conceptual framework will be developed through qualitative work with people with the condition and discussion with relevant professionals. The multi-dimensional PMR-PROM will contain both reflective and formative elements and I will use classical test theory and item response theory in its development where appropriate.

6.4 Item development

There are many possible sources for items to potentially include in a new instrument. Existing questionnaires covering similar or related constructs can be reviewed for applicable items and literature about the condition and experts in the field can help derive and review items to ensure comprehensive coverage. Whilst seeking out expert opinion and peer reviewed literature is important to identify key disease characteristics and consequences, the best source of information for PROMs that seek to measure symptoms, functioning and perceived health, is often patients with the condition.

Once the subject of each item is determined, consideration needs to be given to the format of the questions, ensuring they are comprehensible, specific and clear, and to the response options e.g., whether to use Likert scales, visual analogue scales or free text.

Decisions about these issues will be determined by the intended population group, the nature of the construct and the intended use of the instrument. Again, reviewing existing questionnaires can be helpful at this stage, to identify established formats used in the relevant field.

Documenting the process of item formation is important in ensuring that the PROM meets criteria for content validity when researchers or clinicians are evaluating it for use in the future.

Item banks

In recent years 'item banks' have been developed for specific topics. These are large collections of questions about specified constructs, which have been developed by item response theory, meaning that they can be ordered hierarchically (de Vet, Terwee, et al., 2011b). One of the most well know of these is PROMIS (Patient Reported Outcomes Measurement Information System, https://www.healthmeasures.net/explore-measurement-systems/promis/intro-to-promis). The PROMIS item bank consists of sets of items that can be used to evaluate physical, mental and social health in adults and children, either in the general population or in those living with chronic conditions. There are different item sets for different domains - for example in the 'physical health' domain, there are sets of items for pain intensity, pain interference, physical function, fatigue and sleep disturbance.

The advantages and disadvantages of such item banks are given in Table 6.1. One of the major advantages is that they allow computer adaptive testing (CAT). In CAT, items are dynamically selected for administration based upon the respondent's previous answers so that each response helps to refine a person's score. In this way, respondents need to answer fewer questions and are not asked questions that are not relevant to them.

Table 6.1 Advantages and disadvantages of using item banks such as PROMIS

Advantages of using item banks	**Disadvantages of using item banks**
Create a common currency of items and allow use of a common scoring system	May not be valid for evaluation of a specific condition (items are developed from the general population and individuals with an amalgam of chronic diseases)
Allow CAT which increases precision (thus reducing sample size of studies) and reduces respondent burden	CAT requires computer infrastructure and requires respondents to have a minimum level of IT literacy.
Item bank measures have a larger range of measurement than conventional measures, decreasing floor and ceiling effects	

Item development for the PMR-PROM

In the case of PMR, I have already demonstrated that there are no condition-specific existing measures from which to import items. Generic measures assessing QoL do exist, but these are likely to miss the detail of effects specific to PMR.

There is very limited qualitative literature seeking to understand patient perspectives of the condition and as mentioned above, I therefore carried out in-depth exploratory interviews with patients about their experiences of the condition as a basis for exploring the construct of PMR-related QoL further. These data were also used for generating an initial list of items. These items were then refined and formatted with reference to existing literature, other related PROMs and relevant experts.

This process and decisions made with respect to item development for the PMR-PROM are detailed in Chapter 7 (Section 7.4).

6.5 Pilot testing

Pilot testing of an instrument allows its comprehensibility, comprehensiveness, relevance, acceptability and feasibility to the target population to be assessed (de Vet, Terwee, et al., 2011b). Participants in a pilot test study are asked to complete the questionnaire but also to give feedback on their experience. This can be done individually with participants using cognitive interviewing to talk through each item in turn (Willis et al., 1991) or it can be done using a standardised instrument developed for this purpose, such as the QQ-10 questionnaire, which is a questionnaire designed to test the face validity, feasibility and utility of healthcare questionnaires (Moores et al., 2012). Cognitive interviewing refers to a set of techniques (e.g. think aloud protocols, verbal probes) that enable a researcher to

analyse how respondents understand the survey questions they are to answer (Collins, 2003). It can reveal problems in participants' comprehension, judgement and response processes related to the items, which would not be apparent from other methods such as expert review of the questionnaire (Ryan et al., 2012). However, cognitive interviewing is time-consuming and expensive. A standardised measure, such as the QQ-10, provides less in-depth analysis but can be used with a larger sample at lower cost. It also provides quantitative data, which can be used to compare different versions of questionnaires at specific stages of development or the same questionnaire in different populations (Moores et al., 2012).

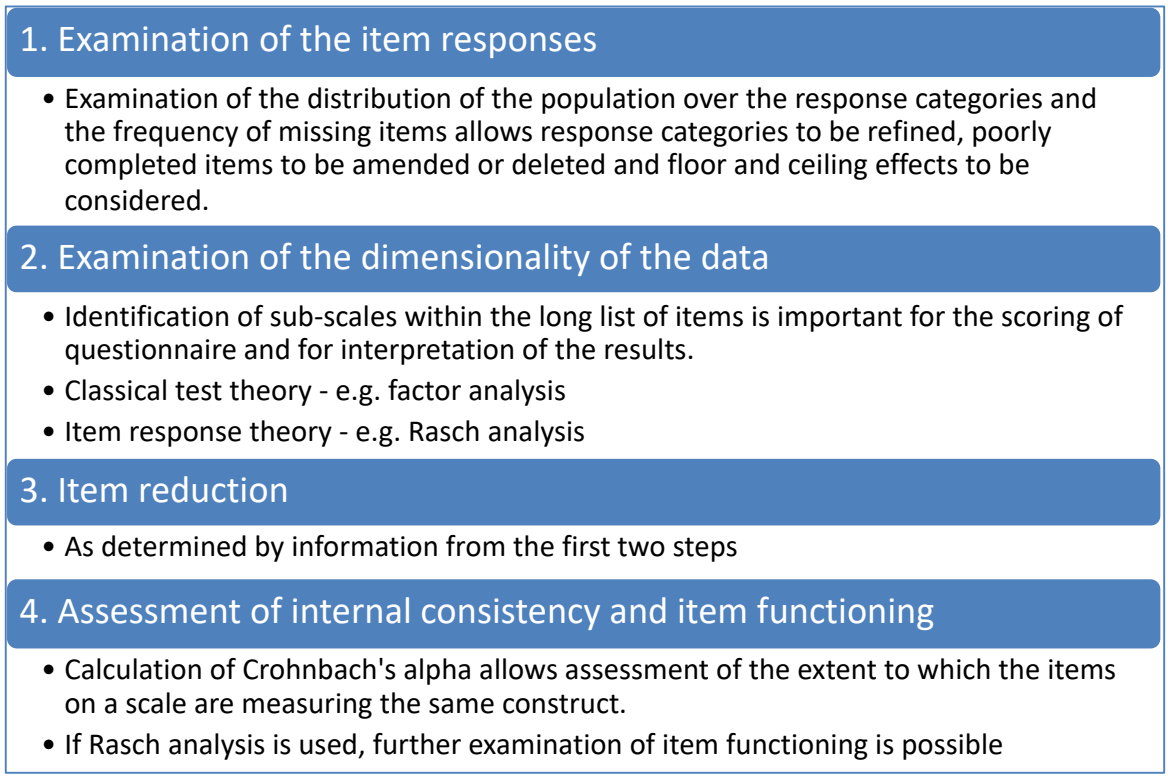
The pilot testing phase of the development of the PMR-PROM is described in Chapter 7 (Section 7.5).

6.6 Field testing

After an instrument is found to be performing satisfactorily during pilot testing, it needs to be subjected to field testing. This is a quantitative stage, which requires data from a large number of completed questionnaires. Analysis of this data allows consideration of dimensionality and the definitive selection of items per dimension.

Figure 6.4 shows an overview of the stages involved in field testing a PROM during its development. Chapter 8 describes the study in which the PMR-PROM was subjected to field testing and the methodology of each stage is described in depth in that chapter.

Figure 6.4: Overview of the process of field testing a new PROM



6.6.1 Classical versus modern test theory

As introduced above, a measurement theory is a theory describing the relationship between the scores generated by the items and the construct to be measured (de Vet, Terwee, et al., 2011a). Measurement theories provide a strategy by which constructs that are not directly observable can be measured using multiple observable items.

Classical test theory

Classical Test Theory (CTT) (Lord et al., 1968) is a measurement theory that can be applied where each item is an indicator of the construct to be measured (i.e. the model is reflective) and the construct is unidimensional. The theory states that the observed score

of an item is the sum of the true score of the unobservable construct plus the associated unobservable measurement error of the item. Assuming that the measurement error is not correlated with the true score (and therefore the average value of measurement errors will be close to 0), taking the average value of the observed scores for many items will therefore approach the true score for the construct (de Vet, Terwee, et al., 2011a).

Factor analysis is the most common method used to examine dimensionality of the data within the CTT paradigm. It is a statistical process to "identify the interrelationships among a large set of observed variables and then, through data reduction, to group a smaller set of these variables into dimensions or factors that have common characteristics" (Nunnally & Bernstein, 1994).

Item response theory

Item response theory (IRT) is an alternative measurement theory to classical test theory. It can be applied when the underlying model is reflective and when the items can, to some extent, be ordered according to difficulty. It was developed by psychologists in the 1960s (Birnbaum, 1968) and IRT models are typically used to measure 'ability'. Each individual has a place on the continuum of the 'ability' being measured (e.g. walking ability) and this is called the patient location. IRT models make it possible to estimate the locations of patients from their scores on a set of items. The items themselves also have a location on the continuum of ability and this is called the item location or item difficulty.

IRT theory is based on Guttman scales (Guttman, 1950). Guttman scales consist of multiple items measuring a unidimensional construct which have a hierarchical order of

difficulty. They are 'deterministic' such that if a person can do an item, they will be able to do all items that are easier providing there are no misclassifications. IRT is based on this principle but allows for some misclassification (as this inevitably occurs) and therefore incorporates probability.

IRT methods describe the association between a respondent's underlying level of ability or severity and the probability of a particular response to the item. Each item can be represented by an item characteristic curve which plots the ability of a person on the x-axis against the probability of reporting difficulty with the item. The higher the ability of the person, the more likely it is that they give a positive answer to any relevant item. The more difficult the item, the less likely that the item is answered positively by any relevant person (i.e., the probability of a successful outcome is governed by the combination of a person's ability and the item's difficulty). However, there will always be some unpredictability in the interactions between persons and items, hence the need for the model to be based on probability. Rasch models (Rasch, 1960) are a particular type of one-dimensional IRT model and are discussed in more depth in Section 8.3.6.

If data fit an IRT model, this theory can be used to examine the dimensionality of the data but also to carry out more sophisticated analysis of items in relation to their position on the scale and their functioning.

6.6.2 Determination of scoring

Individual items on a questionnaire are scored in a way that reflects their measurement level (nominal, ordinal, interval or ratio) and their response options. For example, intensity of pain measured on a 0-10 Visual Analogue Scale (VAS) is a continuous variable

and the score is simply the number in centimetres or millimetres marked by the respondent. Response to a Likert scale is an ordinal variable (there are a number of classes and they have an order) and in this case, response categories are assigned a number, which usually becomes 'the score'.

Individual scales within a multi-dimension instrument are usually scored by simply adding the scores of the items or taking the average of the scores within the scale. There are caveats to this however, and this issue is discussed further in Chapter 8.

The overall score from a multi-item instrument can be presented in several different ways. In some cases, it may make sense to combine the scale scores into one overall score (often called an index). This has the advantage of being simple and giving a result that is easy to work with. However, information about the underlying separate dimensions is inevitably lost which can reduce the usefulness of the result (Devlin et al., 2010). For example, a situation could arise where two individuals have the same index score on a PROM, but one has a lot of pain and little mood disturbance and one has low mood but minimal pain. If the aim was purely to assess the 'overall' degree to which the individual was affected by the condition, this might be appropriate but the required therapeutic response for these two individuals may be very different and purely looking at the index score could give a false impression.

One way to manage the relative importance of different dimensions within an index score is to weight the dimensions. Although empirical evidence can be used to support decisions, in practice this is usually a subjective judgement. In the case of PROMs or QoL measures, patients can be involved in the weighting process through a consensus process.

An alternative reporting method to the index score, is to present total scores for each scale separately, as a profile. This retains the information about individual scales but is less convenient to work with.

6.7 Reliability

COSMIN defines reliability as:

"the extent to which scores for patients who have not changed are the same for repeated measurements under different conditions." (Mokkink et al., 2010b)

In other words, it is the degree to which a measurement is free from measurement error.

For any measurement there are several potential sources of variation. These include the underlying properties of the instrument itself (e.g., its internal consistency), the persons performing the measurement, the patients undergoing the measurement and the circumstances under which the measurement is taken.

When assessing the reliability of any instrument it is therefore important to consider:

1) The internal consistency of the instrument.

2) Test-retest, intra-rater and inter-rater reliability – reliability over time, with different persons on the same occasion or by the same person on different occasions.

3) Measurement error.

6.7.1 Assessment of reliability of the PMR-PROM

The internal consistency of the PMR-PROM will be considered during its development phase (Chapter 8). As the instrument will be completed by patients themselves, inter-

rater and intra-rater reliability are not relevant. Test re-test reliability (with calculation of parameters of reliability and measurement error) will be assessed in the evaluation of the final PROM (Chapter 9).

6.8 Validity

Validity is defined as *"the degree to which an instrument truly reflects the constructs it purports to measure"* (Mokkink et al., 2010b).

There are three main types of validity:

1. Content validity – whether the content of the instrument adequately reflects the construct to be measured.
2. Criterion validity – whether the scores of a measurement instrument agree with scores of a gold standard.
3. Construct validity – whether the instrument provides the expected scores based on existing knowledge about the construct.

Validity is specific to the situation that the instrument is tested in. If an instrument is to be used in a new target population or for a new purpose, new validation studies should be done.

6.8.1 Content validity

For unobservable constructs measured by multi-item assessment tools, evaluation of content validity requires a clear understanding of the construct being evaluated and the purpose of the measurement (i.e., whether it is to discriminate between persons at a

point in time, to evaluate change over time or to predict future outcomes). A judgement then needs to be made as to whether the items are relevant and comprehensive in relation to the construct and purpose.

Face validity is a form of content validity describing a subjective overall assessment of whether the instrument 'looks' as though it reflects the construct being measured.

To ensure an instrument can be assessed for content validity, developers need to fully describe the construct, the conceptual model and the development process of the instrument. Researchers using the instrument can then determine if the content reflects what they are trying to measure in their particular target population. In most cases this is a subjective judgement with no formal statistical process involved.

6.8.2 Criterion validity

If a gold standard measurement instrument is available for assessing a particular construct, the new instrument can be compared to this one to assess its validity. A hypothesis about the relationship and the required level of agreement is decided *a priori* and the scores for the two instruments are obtained independently. Statistical parameters can then be used to compare the results from the two instruments as appropriate to the level of measurement (see Table 6.2).

If the new instrument is shown to be valid compared to the existing gold standard, information about costs, burdens, reliability etc. can then be incorporated into the decision about which instrument is superior overall.

Table 6.2 Parameters for evaluating criterion validity according to level of measurement

(adapted from (de Vet, Terwee, et al., 2011))

Gold standard	New measurement instrument		
	Dichotomous	**Ordinal**	**Continuous**
Dichotomous	Sensitivity and specificity	ROC	ROC
Ordinal	ROC	Weighted kappa or Spearman's r	ROCs / Spearman's r
Continuous	ROC	ROCs / Spearman's r	ICC / Bland and Altman limits of agreement

ROCs = receiver operating characteristic curves, ICC = intra-class correlation coefficient

6.8.3 Construct validity

In many cases where new instruments are developed, there is no gold standard for comparison. Assessment of validity is therefore made through a process of evidence gathering rather than a straightforward comparison.

Detailed knowledge of the construct and the conceptual model is essential to be able to evaluate it and this must be clearly specified by instrument developers.

Structural validity is one aspect of construct validity and is defined as "the degree to which scores of a measurement instrument are an adequate reflection of the dimensionality of the construct to be measured" (de Vet, Terwee et al., 2011). In the development phase this is assured by methods such as exploratory factor analysis to determine the structure of the instrument. If studies are then required to test structural validity, confirmatory factor analysis can be used.

Once structural validity is confirmed, hypothesis testing can then be used to further evaluate construct validity. The scores of the instrument are tested against hypotheses, specified *a priori*, about expected relationships with scores of other instruments or expected differences in scores between relevant groups. The hypotheses ideally need to describe the direction and expected magnitude of change and details of the comparator instruments also need to be clearly described.

When the construct is multidimensional, each scale or part of the instrument that measures a specific dimension needs to be evaluated by forming hypotheses for the different dimensions separately.

6.8.4 Assessment of validity of the PMR-PROM

Since there is no gold standard for evaluation of disease-specific quality of life in people with PMR, criterion validity is not relevant. Instead, I will consider content and construct validity. In the development process this will be assured through detailed exploration of the construct, documentation of item selection, face validity testing of the items and determination of dimension structure using exploratory factor analysis and item response theory. I will then evaluate construct validity in a study where hypotheses about relationships to comparator instruments and expected changes in response to treatment are tested (Chapter 9)

6.9 Responsiveness

Responsiveness is defined by COSMIN as:

"the ability of the instrument to detect change over time in the construct to be measured" (Mokkink et al., 2010a).

In other words, an instrument is responsive if scores on the instrument change correspondingly when patients change with respect to the construct in question.

COSMIN are specific about the change being evaluated needing to be a true change in the construct of interest because previous definitions of responsiveness encompassed any statistically significant change after treatment (Mokkink et al., 2010b). This meant that change caused by 'noise' or caused by change in a different construct could be taken as evidence of responsiveness. The COSMIN definition places responsiveness within the concept of validity as the instrument needs to truly measure the construct it purports to measure to capture change in that construct.

When assessing responsiveness, a study therefore needs to be set up in which the patient group are expected to change in the construct being evaluated and measurements need to be taken at two time points. There are two main methodological approaches to responsiveness studies, a criterion approach and a construct approach.

6.9.1 The criterion approach

This is used when a gold standard measurement instrument for the construct of interest is available. The required level of agreement between changes on the new measurement instrument and the gold standard is defined *a priori* and the change scores on the measurement instrument and the gold standard are obtained independently but over the same time period. The strength of the relationship between the change scores is then calculated.

The statistical test used for this comparison is determined by the level of measurement in the same way as for criterion validity testing (Table 6.2). If the scores on the new measure and the gold standard are continuous variables, correlations between change scores can be used. If the gold standard is a dichotomous value, e.g. if people are grouped into 'improved' or 'stable', a receiver operating characteristic curve (ROC) analysis can be carried out. In this, the area under the curve (the AUC) is considered to measure the ability of the new instrument to discriminate between people who have improved and those who have not, according to the gold standard.

6.9.2 The construct approach

This is used when there is no gold standard measurement instrument for the construct being evaluated. To assess responsiveness in this situation, hypotheses need to be formed *a priori* about expected change scores in certain situations. Hypotheses can be formed about:

a) expected mean differences in change scores between different groups e.g., treated versus untreated participants or groups determined by ratings on a global rating scale
b) expected correlations between change scores on the instrument under test and changes in scores in other instruments known to have adequate responsiveness.
c) expected relative correlations between change scores on different instruments e.g., the change on instrument X is expected to correlate more strongly with the change on instrument Y than the change on instrument Z because the constructs measured by X and Y are more closely aligned.

6.9.3 Assessment of responsiveness of the PMR-PROM

As for the evaluation of construct validity, responsiveness testing of the PMR-PROM will be based on testing hypotheses in an observational study where patients are treated with usual medical care and followed over a period of time (see Chapter 9). Hypotheses about expected change scores on the PMR-PROM will be formed and tested.

6.10 Interpretability

Interpretability can be defined as *"the degree to which one can assign qualitative meaning to quantitative scores"* (Mokkink et al., 2010a).

Whilst interpretability is not strictly a measurement property, COSMIN include it in their taxonomy because consideration of the meaning of scores from an instrument is vital for its appropriate use in research or clinical practice.

Issues to consider when assessing interpretability of scores from a measurement instrument include:

- the distribution of scores in the study sample
- floor and ceiling effects
- availability of scores for relevant groups and subgroups (e.g., for normative groups or the general population)
- the smallest detectable change
- the minimally important change

The distribution of scores (item responses) and the assessment of risk of floor and ceiling effects are concepts that require consideration in the development of an instrument as well as at the point when its interpretability is considered. These issues are discussed in more depth in the development context in Chapter 8 of this thesis, but they are mentioned here as they need to be evaluated again when an instrument is considered for use in a particular population.

6.10.1 Distribution of the scores of the instrument

In the development of an instrument, the distribution of scores in a study sample is used to consider whether the instrument 'fits' the population. This is discussed in relation to field testing of the PMR-PROM in Chapter 8 (Section 8.3.2). In the validation stage, the distribution of scores is also important because it aids interpretation of the scores and of the reported measurement properties.

The mean and standard deviations of the scores, or the proportional distribution over response categories, provide information about the location of the study sample on the measurement instrument e.g., whether the patients in the study sample are homogenous and clustered at a particular location on a scale.

With respect to measurement properties, the heterogeneity of the sample influences the reliability parameters and the strength of correlations between scores, so knowledge of the distribution of the study sample can aid interpretation of reliability and construct validity assessments.

6.10.2 Floor and ceiling effects

Floor and ceiling effects can occur when a high proportion (usually taken to be >15%) of the total population has a score at either the lower or upper end of the scale (de Vet, Terwee, et al., 2011d). If the score on a measure decreases as the construct 'improves' then a floor effect indicates that people in this range are 'better' than the measure can detect.

If present, they may indicate limited content validity (not enough items at either end of the scale) and reduced reliability (patients with very high or very low scores cannot be distinguished from others with similar scores). Responsiveness is also affected as patients who score at either end of the scale cannot show any improvement or deterioration so any change in that direction in the construct being measured cannot be detected by the instrument. Similarly, if a patient moves into, or out of, the ceiling / floor range on repeat testing, the size of the change cannot be calculated.

6.10.3 Interpretation of scores through known groups

To understand the meaning of the scores on an instrument, scores for certain populations can be calculated and used in comparison e.g., the distribution of scores in the healthy population or the distribution of scores in people with certain conditions. This can help users of the instrument get a feel for what different scores mean.

6.10.4 Smallest detectable change and minimally important change

The smallest detectable change, SDC, (sometimes referred to as minimal detectable change, MDC) is the smallest change beyond measurement error, which can be measured by an instrument (de Vet, Terwee, et al., 2011d). It is a concept closely related to the

reliability of an instrument – if the repeated scores in stable patients vary greatly, the changes on the measurement instrument have to be large before they can be said to be represent real change rather than measurement error.

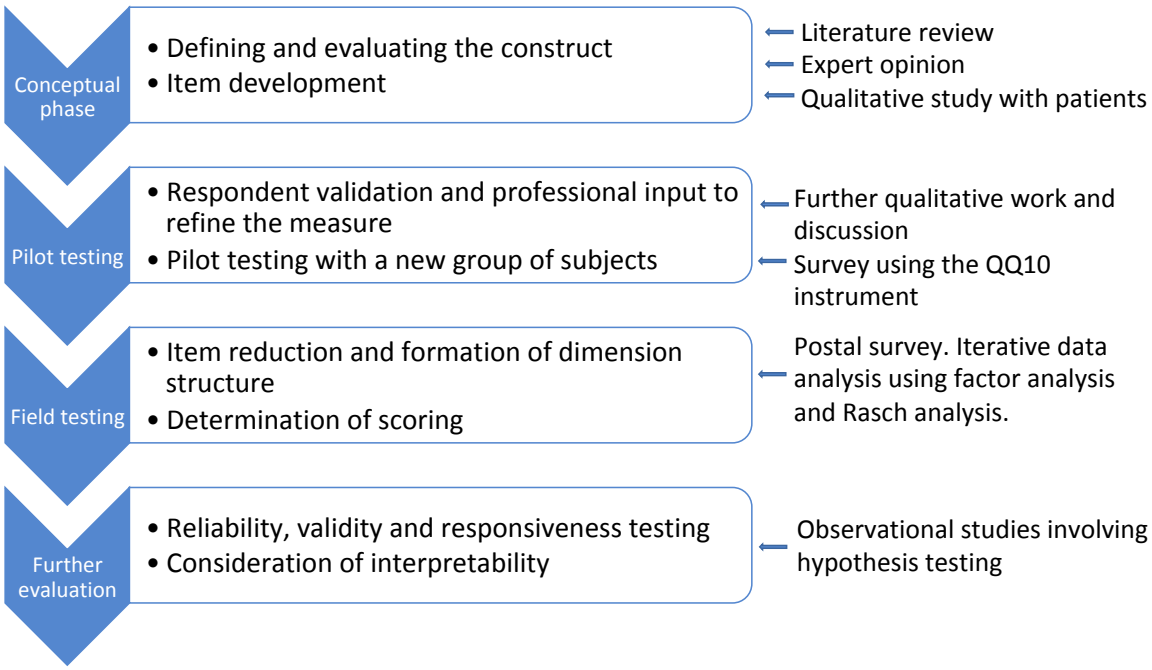
The minimally important change (MIC) is the smallest change in score in the construct being measured that is perceived as clinically important (de Vet, Terwee, et al., 2011d). For PROMs, this is best determined from the perspective of the patient (Dworkin et al., 2008). The concept of the MIC helps answer the question of whether changes that may be statistically significant, are actually of relevance for patients.

There are several different methods to determine the SDC and the MIC and these are discussed in more detail in Chapter 9 (Section 9.3.5).

6.11 Summary

In this chapter I have outlined the key stages of PROM development and discussed the relevant methodological issues and choices for each step of the process. I have described the way in which this theory will be applied to the development of a new PROM, the PMR-PROM, and in subsequent chapters the studies relevant to each stage will be described. Figure 6.5 shows an overview of the development process of the PMR-PROM. It is a repeat of Figure 6.3, from the introduction, but with the addition of the methods to be used for each stage.

Figure 6.5: Overview of the process and methods for development of the PMR-PROM



Chapter 7: Development work

7.1 Introduction

This chapter describes the initial development work for the PMR-PROM, which was carried out prior to starting my PhD. This is included in this thesis to provide context and to describe steps previously taken. The studies contributing to the development have been published (Twohig et al., 2015, 2018) and the full papers are included in Appendix 7.1: Published papers on development work carried out prior to this PhD. I will provide an overview of the work here for completeness, so that the full process of instrument development is described.

7.2 Patient and public involvement

Throughout the process of developing the PROM, I have been in contact with the charity PMRGCAuk (www.pmrgca.co.uk), which represents and supports people with PMR and GCA. I first contacted them in 2012, at the inception of the process, and attended a meeting of the PMRGCAuk North East Support Group to discuss participants' thoughts on the idea of developing a PROM. I maintained contact with the group following this, seeking their input and feeding back on progress at each stage of the study. Trustees of the national group were consulted on study design and funding applications for the distinct research studies which have contributed to the overall development of the PMR-PROM and informal discussions with members at group meetings helped to refine the process at various stages along the way.

As discussed in Chapter 4, I am also a member of the OMERACT Special-Interest Group for PMR. This group has PMR patient partners amongst its members and they have also given feedback at specific points of the process. The results of the qualitative study (Twohig et al., 2015) and the conceptual framework were discussed within the group and then presented at the international OMERACT meeting in 2016 to a mixed audience of professionals and patient partners. Ideas generated from these discussions fed into the content and format of successive versions of the PMR-PROM.

7.3 Development of the conceptual framework

7.3.1 Qualitative study of patient experiences of PMR

In-depth semi-structured interviews with 22 people with PMR, recruited from primary care practices in South Yorkshire, were carried out in 2013-14.

Participants were eligible for inclusion in this study if they were aged over 50 years and had a Read coded PMR diagnosis and classical PMR symptoms (documented in the electronic medical record as having bilateral shoulder and / or pelvic girdle pain and stiffness for at least 2 weeks, and evidence of an acute phase response (raised ESR / CRP)). They were also eligible if they had a diagnosis of PMR but had atypical features (e.g. normal ESR / CRP), providing their diagnosis had been made by a rheumatologist. Exclusion criteria included significant dementia or memory impairment, a primary diagnosis of giant cell arteritis, a concomitant inflammatory arthropathy, active cancer or if the screening GP decided that participation wasn't appropriate (e.g. other terminal illness).

The topic guide for the interviews was based on existing relevant literature and input from the multidisciplinary advisory group. Topics included in the initial guide were onset of the condition, symptoms and functional effects, diagnosis, flares and relapses, starting and stopping treatment, resolution of the condition and information provision. An open questioning style was used with minimal prompts to allow themes to emerge naturally. Thematic analysis, using a constant-comparative method, was used. Analytic codes and categories were developed through an iterative, thematic and self-conscious process, beginning in parallel with the data collection and informing subsequent interviews as concepts and themes emerged. The process of constant comparison continued until theoretical saturation was reached and no new themes were emerging.

7.3.2 Emergent themes

Five key themes were identified: 1) pain, stiffness and weakness, 2) disability, 3) experience of care, 4) treatment and course of the condition and 5) psychological impact. The themes and sub-themes with illustrative quotes are detailed in Table 7.1.

7.3.3 Development of the framework

As described in Chapter 6 (Section 6.3), a conceptual framework is the term used to describe the relationship between component measurable characteristics and the non-observable construct. Data from the qualitative study was used to inform development of a framework for the construct PMR-related quality of life. The framework was discussed with the multi-disciplinary study team, patient partners and members of the OMERACT PMR-SIG and modified through several iterations to the final version shown in Figure 7.1.

Table 7.1 Summary of themes from interviews on people's experiences of PMR

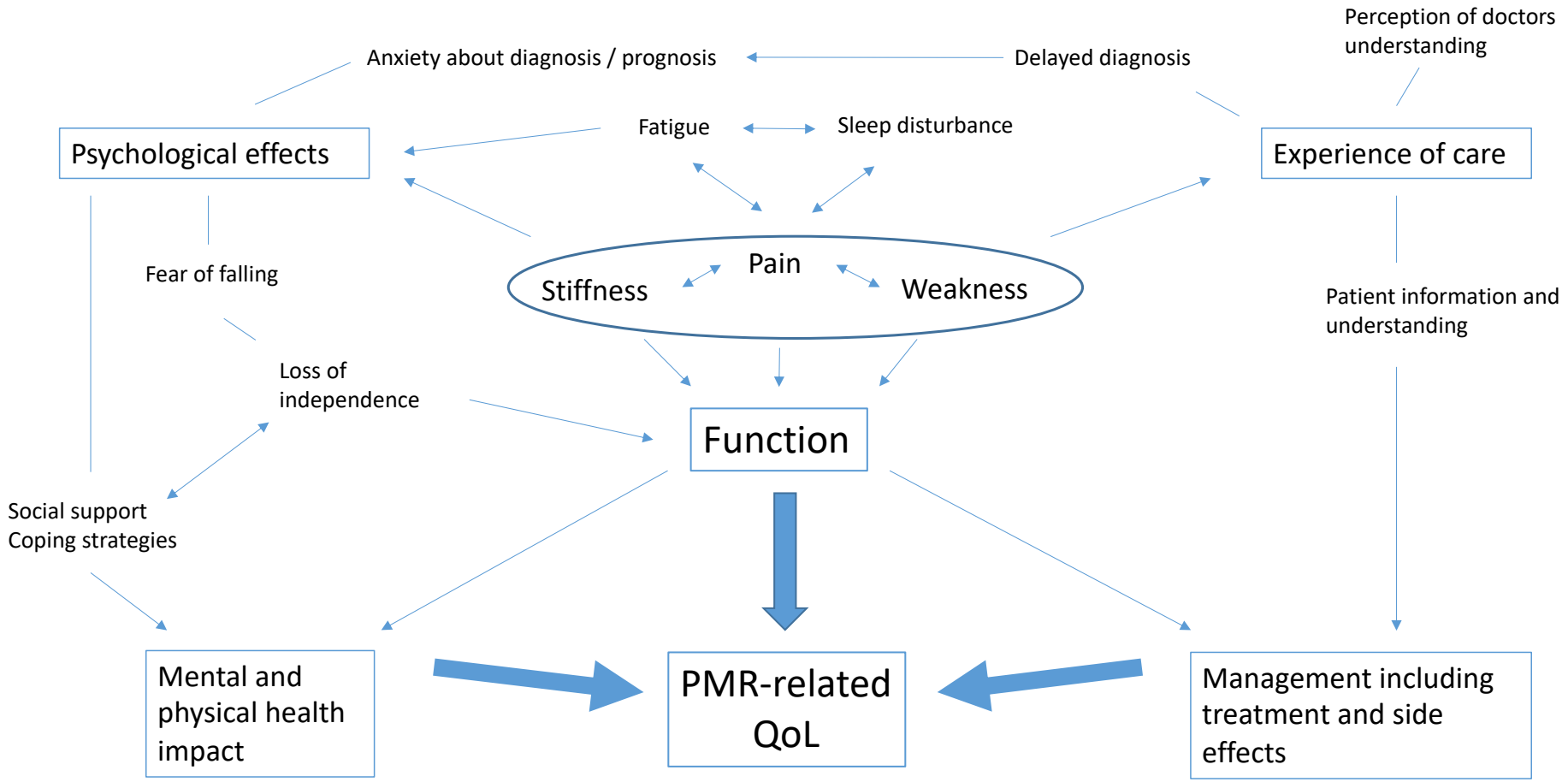
Pain, stiffness and weakness		
Heterogeneity of symptoms	Difficulty distinguishing and describing pain / stiffness / weakness	*"I could hardly move in bed, it was aching all down my back and I just felt, I suddenly felt I'd aged from like I was about 80 year old, that's what it felt like. And very stiff, very achy like when you turned over in bed it was painful." P16* *"And I really screamed in pain. You know, to get dressed. Or even to lift my arms up. The pain was terrible." P18* *"Well it's not pain, it were more of a bad ache and I couldn't do much, you know." P13* *"..the shoulders and the biceps... they felt very weak… they weren't painful, just wouldn't work" P 5* *"When you first have PMR, it used to take me til about tea-time to actually come round. And even when I started on the prednisolone, I didn't sort of come round straight away as I've told you. But that's when I noticed the prednisolone was working, that the pain was -, I was freer much earlier in the day." P2*
Variation throughout the day	Classic 'morning stiffness' not often described. Worse in the mornings and evenings and after rest.	
Effect of activity	Many felt that symptoms were worse for a few days after activity	
Disability		
Sub-acute onset of significant disability	Often in people who were previously fit and active. Described what they couldn't do rather than 'symptoms'	*"I didn't know how to get in the car because my legs wouldn't bend, my arms wouldn't bend, she had to put one of her little one's booster seats on the front seat so I didn't have to lower myself quite so low and it had got to the stage where I couldn't lift my arms to comb my*

	Links with the significant psychological impact.	*hair… really struggling with everything, walking upstairs and everything." P14* *"I woke up and I was on my front, I couldn't get over in bed. And I developed strange pains across the top of my shoulders. I came back from that holiday….and I just went down within about a week of not being able to get out of bed, not being able to turn over in bed." P2*
Fatigue	Normal ADLs became a huge effort Sleep disturbance	
Experience of care		
Delayed diagnosis	Many had multiple visits to the GP prior to a diagnosis being made. Some had trials of analgesia, physio, stopping statins etc. Retrospective frustration.	*"I think I'd been up to see her when it first started, 'cause I could hardly walk... She kept sending me for these blood tests and the last time they wanted another blood test off me, my husband went up; he says, 'Look, my wife can't get out of bed this morning.' And they sent a doctor down to take it. And then he says to her, 'I think you ought to send her in hospital. She needs treating. She's not getting anywhere.' And that's what she did then, you see – she sent me to hospital." P18* *"I did actually have a month on a, what do they call it, you know the antidepressants, because I was going with all these pains and I wasn't getting anywhere at all. But I knew as soon as I started on the antidep-, it wasn't for me and that was it, after the month I came off them and I thought well, you know, I'm just going to see this through and I'm just going to have to see what's going to happen. And I'm going to have to create eventually and ask to see a specialist or something, because when you get to my age and you've been fit, you do know your own body, you know if there's something right or wrong." P2*
Perception of doctors understanding	Patients felt they had a condition which was poorly understood by the medical profession.	
Patient information sources	Frequently sought out information on-line. Many finding support from 'peers'.	

		"When they fetched me back in and told me what I'd got and she printed so many sheets out and she said, the doctor, 'this is exactly you' and it was, that you can't get out of bed and you can't do this and you can't do the other….I mean it was all about it and it was me, definitely me." P17 *"When they said what I'd got, I was very pleased when they gave me the medication and it started to work so well. I was very happy about that but when they said that there's no cure for it because we don't know what it is, that was a bit upsetting." P5*
Treatment and course of the condition		
Effect of prednisolone	Rapid resolution of symptoms – 'miraculous' Multiple side effects Difficulty distinguishing side effects from the PMR	*"He put me on these Prednisolone and it was like magic, it was just so good, you know, that I had no pain and I went back again to let him know how I was going on and I says 'thank you, you know, I can't say to you what a difference that's made to me'" P12* *"Well I can put up with it, I can live with it, it's affected me all these aches and pains, aching and that, it's not as much of a sharp pain, it's just, you know, like, nagging ache. I can put up with that, but it's just, I think it's these side effects what I'm getting with the tablets what's worse. I feel as though this is worse now than the actual bad aching." P13* *"Every time he dropped the dose for a week, I could tell that it had dropped dose and I weren't very well, but I carried on and it like worked itself off, I worked through it sort of thing" P11*
Flares and relapse	Resurgence of original symptoms but less intense Slight worsening with each dose reduction	
Long-lasting effects	Never quite recovering to level of health they had had before	
Psychological impact		
Impact of the disability	Often devastating impact on quality of life	

Fear about diagnosis / prognosis	Anxiety about muscle wasting disease / terminal illness Fear of never recovering function Disease name unfamiliar and frightening to some	*"But, well, I thought worst, you know, I thought I were like, what these illnesses where you just finish paralysed, I don't know what they call them but I felt it were going to be something like that, because I were getting worse." P11* *"But as I say, it were just – it got to a stage as I say when I went to the doctors – it got to a stage when I were literally struggling to turn over in bed – that were quite frightening because you'd lay on your back and all of a sudden you're thinking 'well, it's almost like being locked in your body in a way'. Yeah – you hear about – and I forget what the name of these – some of these things – but these wasting away diseases…" P11* *"Because I hadn't heard of it at all. I really did think oh thank God somebody's listening to me. I thought I was imagining it." P3*
Loss of confidence	Fear of falling Loss of independence	
Relief at diagnosis	Importance of having a diagnostic label to validate symptoms. Belief that diagnosis means effective treatment.	

Figure 7.1 Conceptual framework for PMR-related Quality of Life developed from qualitative study of patient experiences of PMR



7.4 Item development

An initial list of candidate items for the PROM was derived from the interview data. Each functional activity that participants mentioned as being affected was listed along with symptoms experienced, psychological and social effects and steroid side effects (see Appendix 7.2: List of items derived from the interview data).

Questions were formulated to encompass each of these items and this long-list of potential questions was sent to all the participants in the original interview study who were invited to send it back with comments. Ten of these participants agreed to a structured telephone interview to discuss the proposed items, question wording and response options in more depth and these comments and discussions were used to refine the format of the questions. The resultant questionnaire (Appendix 7.3: PMR-PROM Version 1) was circulated to the study Steering Group comprised of GPs, consultant rheumatologists and an expert in PROM development methodology, and modified according to their feedback to form Version 2 (Appendix 7.4: PMR-PROM Version 2).

Key decisions made by the end of this process were:

1. To divide the PROM into four domains
2. To include 'weakness' in the key symptoms (as this was discussed by patients in the interviews despite not being a classically described symptom of PMR)
3. To use a look-back period of two weeks for the questions
4. To use a numeric rating scales rather than visual analogue scales in the symptoms domain
5. The long-list of items and their response options for the function, emotional and psychological well-being and steroid side effects domains had been agreed.

7.5 Pilot testing

The next step in the process was to pilot test the PMR-PROM to establish its face validity.

A new group of 28 people with PMR were recruited via a patient-led support group, PMRGCAuk North-East Support, and via rheumatology clinics at Leeds General Hospital NHS Trust. Participants were asked to complete the PMR-PROM and then complete the QQ-10 questionnaire, which is a measure designed to assess the face validity, utility and feasibility of healthcare questionnaires (Moores et al., 2012).

The PMR-PROM used in this study had been updated with the addition of a demographic details sheet and modification to the layout to make it suitable for use in this postal survey and this version was labelled Version 3 (Appendix 7.5: PMR-PROM Version 3).

The QQ-10 questionnaire is a self-completed questionnaire containing 10 items with a Likert response scale (scored 0-4). The first six questions contribute to a 'value' score (higher scores equate to greater value) and the next four questions contribute to a 'burden' score (higher scores indicate greater burden) for the questionnaire under assessment. There are also three questions requiring a free text response. A copy of the questionnaire is given in Appendix 7.6: QQ-10 Questionnaire.

The findings from this study have been published and the paper is included in Appendix 7.1: Published papers on development work carried out prior to this PhD (Twohig et al., 2018). In summary, the overall mean value score in this study for this version of the PMR-PROM was 79% (SD 12) and the mean burden score was 21% (SD 18). A chart showing the distribution of responses to each question is shown in Figure 7.2. The component contributing most strongly to the burden of the questionnaire at this stage was its length,

but this was something that would be addressed through the subsequent item reduction process.

The free text responses to the final three questions of the QQ-10 were analysed thematically and results from this are summarised in Table 7.2.

Figure 7.2 Chart showing the distribution of responses to the questions on the QQ-10 when used to assess face validity and feasibility of Version 3 of the PMR-PROM

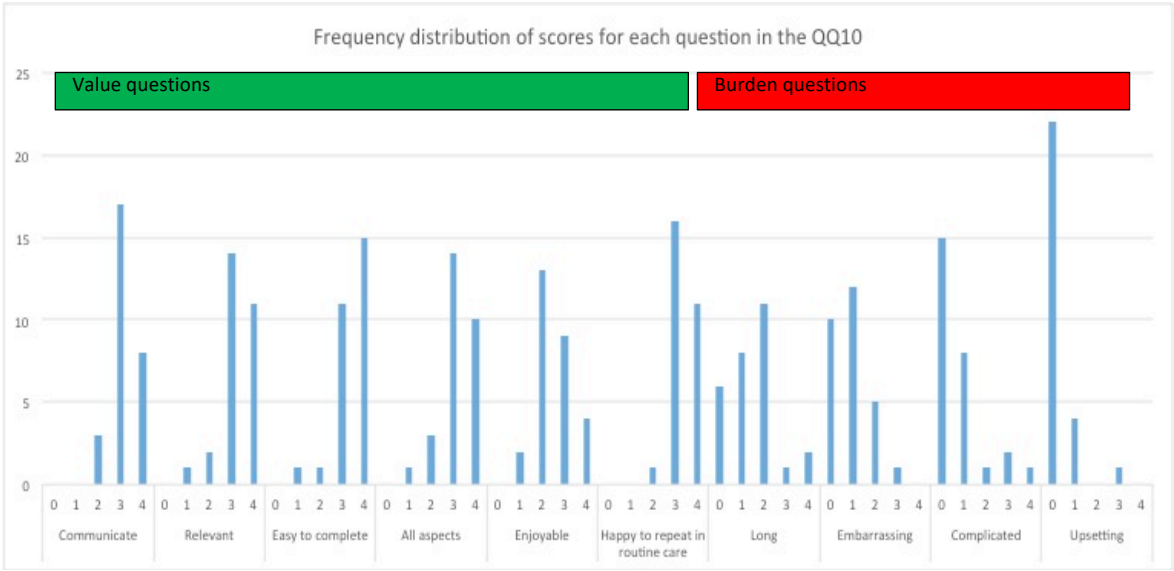


Table 7.2 Content analysis of the free-text responses to the QQ-10 questionnaire

Theme	Sub-theme	Quote
Layout		I think it's well set out in a fairly simple and effective format
		Where the table goes over the page, it would be helpful to repeat the headers. Page numbering would be helpful.
		Sheets should be numbered and column headings should be repeated where they go onto 2 pages.
		Questions about age etc. should be at the start.
		Figure diagrams are a more direct indication of the type and location of pain than the written word.

Content	Depth and detail	Too much detail, cut down on the number of boxes, they overlap too much. Q5 has too many choices.
		Seems quite straightforward
		Q9 (Do you feel back to the level of health you were at before you had PMR?) - if answer is no, ask why?
	Specificity to PMR	PMR is not that specific. For me I sometimes ache and sometimes feel a bit down. Muscle power has diminished but that may be age.
		Many things you ask could be for other reasons such as depression, arthritis, cancer etc. it's not all PMR.
		Some info on other conditions should be included e.g. I had had a stroke previously and PMR imposed symptoms on top of those resulting from that.
		Some symptoms (like weakness and difficulty reaching things in cupboards) I already had from a shoulder injury it's difficult to tell whether the PMR made it worse or not.
		As everyone is different there should be room for an individual's particular symptoms and concerns.
		Q9 is irrelevant, no-one gets back to feeling as well as before.
	Individual variance	Should pain / ache be quantified? Different people will have different meanings.
		Some questions didn't seem to fit my symptoms but I don’t think I'm very severely affected - time will tell.
		PMR affects all of us in different ways and the questionnaire covers all aspects and does no harm even if some questions overlap.
Benefits and when it would be of use		I think it's an excellent questionnaire from the outset of a PMR diagnosis to a record of the PMR journey.
		I think the questionnaire will be very helpful, especially to people at the beginning of their treatment when they probably have all of the problems listed. I would have been reassured to think the doctors knew how I was feeling.
Specific items missing		The side effects of prednisolone
		Diabetes, fluid retention leading to lymphoedema
		The area of pain
		More questions about fatigue could be included. Skin / hair condition missed out.
		Swelling of the joints is not mentioned (hands, wrists, feet and ankles)
		No real questions about where the pain was or what I couldn't do.

		One of my main concerns at the start was inability to fasten my own bra - had to ask for help.
Other concerns that could be included		Concerns about recurrence are important.
		You might like to know how the steroid treatment has helped / been successful.
		It may be helpful to ask patients to put down aspects of their health that may not be PMR related but which they wish to discuss

7.6 Further development of the PMR-PROM

The results of the face validity testing study informed further amendments to the PMR-PROM. The updated questionnaire was again circulated to members of the study Steering group (two rheumatologists and a GP) for comment and modified further in response. The section on steroid side effects was also amended at this stage in response to a paper published on patient perspectives on glucocorticoid adverse effects (Black et al., 2017).

The key changes made through this process were:

1. Alteration to the layout to include page numbers, repeating headings where tables continue over a page and amendment to the numbering system for the questions.
2. Alterations to the layout and questions included in the 'personal details' section.
3. Revising the look-back time to three days as two weeks was deemed too long, particularly in the early stages of the condition.

4. Changing the response options to the symptom duration questions to include the category of '<30 minutes' as this is the usual cut off for classifying morning stiffness and having five response options makes scoring simpler.
5. Addition of questions about dose of prednisolone and overall severity of steroid side effects.
6. Addition of items to the steroid side effects list and alteration of the response options for this section.

Versions of the PMR-PROM which reflect these changes are included as Appendix 7.7: PMR-PROM Version 4, and Appendix 7.8: PMR-PROM Version 5.

7.7 Summary

In this chapter, I have documented the stepwise, iterative early development of the PMR-PROM - from its inception, through the qualitative work to define and explore the conceptual framework, the development and modification of the items and finally, the face validity testing of the long-form version. Each stage was carried out in accordance with the methodology outlined in Chapter 6, with close patient and professional input at appropriate points. This background sets the context for the studies which follow in this thesis.

Chapter 8: Field Testing the PMR-PROM – Item reduction and Scale Generation

8.1 Introduction

In Chapter 7, I described the early stages of the development of a PROM for PMR. Following PROM development guidance, as outlined in Chapter 6 (6.2), the next stage is to carry out field testing with a view to developing and analysing scale structure and reducing the number of items. This process will also enable the scoring system to be developed.

In order to do this, responses to the questionnaire are needed from a large number of participants. This chapter describes the postal survey used to achieve this, the subsequent analysis and the PROM development process.

8.2 Aims and objectives

8.2.1 Aim

To field-test the draft PMR-PROM to enable item reduction, scale development and development of the scoring system.

8.2.2 Objectives

1. To carry out a postal survey to collect responses to the PMR-PROM from a sample of people with PMR.

2. To analyse this data using descriptive statistics, Classical Test Theory and Rasch analysis to reduce items and form scales.
3. To develop the scoring system for the PROM.

8.3 Methodology

8.3.1 Methodology relevant to data collection

In order to collect a large number of questionnaire responses from people with PMR, I decided to recruit participants through general practices and carry out a postal survey.

Identifying participants through primary care ensured that they had an established diagnosis of PMR, as opposed to recruiting through support groups or by general advertisement where diagnosis would rely on self-report. Using primary care rather than secondary care ensured access to people with the full spectrum of the condition (as only a small subset of people with PMR are referred to secondary care) and increased the numbers of potential participants.

Postal survey methods have several advantages over face-to-face interviews or telephone interviews to collect questionnaire data (Bowling, 2005; Rea & Parker, 1997). They allow data collection from large numbers of participants over a wide geographical area, are relatively low cost, quick to do and avoid interviewer effects. From a participant perspective, they allow ample time to complete questions at a time that is convenient for them and allow anonymity, increasing willingness to answer questions on sensitive topics. Disadvantages of postal surveys however, include the administrative burden, relatively low response rates with associated non-response bias, reduced ability to control the

context of the response, problems of missing data and difficulty gauging the salience of responses.

Using on-line platforms to collect questionnaire data is becoming increasingly common. However, data from 2018 from the Office of National Statistics showed that 56% of over 75-year-olds reported using the internet 'more than 3 months ago or never' (Office for National Statistics, 2018) suggesting that people in the age group affected by PMR are less likely to be comfortable with this approach currently. Rates of internet usage in this age group have increased each year since the data was first collected in 2011 however, and web-based surveys may become more viable for this demographic in the future. A recent study in Sweden compared response rates, demographics of respondents and the outcomes measured by the questionnaires themselves from a group of older adults in responding to a web-based survey to those who only responded to the paper version (Kelfve et al., 2020). They found that the respondents who did not answer until they got a paper-questionnaire were more likely to be female, retired, single, and to report a lower level of education, higher levels of depression and lower self-reported health, compared to web-respondents. This suggests that in addition to the risk of overall low response rates to web-surveys for this age group, there is a risk of bias as particular sub-groups of the population are even less likely to engage.

For this study therefore, due to the numbers of participants required, their age demographic and the nature of the questionnaire (types of questions and response options), I felt it was most appropriate to use a postal survey to collect self-completed questionnaires.

8.3.2 Examining the item responses

The first step in field testing a PROM is to examine the distribution of the population over the response categories and the frequencies of missing items.

Examining the distribution of scores at an item level allows assessment of whether all response options are informative and to check whether there are items for which a large part of the population has the same score. If there are response categories which are rarely used, these might be able to be combined. Items for which a large part of the population has a similar score are not able to discriminate between patients and are therefore less useful.

For items scored on a continuous scale e.g., a Visual Analogue Scale, the mean and standard deviation (SD) of the score provide information about the distribution. Very high or low mean item scores represent items on which nearly everyone agrees. Items with a small SD are poor at discriminating between groups within the population.

The percentage of 'missing' responses can also be helpful in evaluating an item. In general, <3% missing responses is ideal whereas >15% is not acceptable (de Vet, Terwee, et al., 2011c). However, whether an item is modified or excluded on this basis will also depend on the reasons why responses might be missing and the importance of the item to the construct. For example, if an item is felt to be central to the construct but is positioned at the end of the questionnaire and has a high number of missing responses, it might be appropriate to re-test the questionnaire with this item moved up to see if this affects whether it is answered.

Examining the distribution of item responses also allows consideration of risk of floor and ceiling effects. Floor and ceiling effects can occur when a high proportion of subjects

score the lowest or highest score, meaning that the measure cannot discriminate between subjects at either extreme of the scale (Lim et al., 2015). The generally accepted proportion for risk of this occurring is if >15% score either the lowest or highest score (McHorney & Tarlov, 1995). Floor and ceiling effects have consequences for the responsiveness and interpretability of the instrument because if a person has a score above that of the most difficult item and they improve, this will not be detected (ceiling effect) and similarly, if a patient has a score below that of the easiest item and they get worse, this will not be detected (floor effect). However, having a high proportion of subjects scoring at either extreme only causes a true floor or ceiling effect if it is actually important to try to discriminate these people further. For example, if the aim of an intervention is to produce improvement up to a certain point and further improvement beyond this threshold does not need to be measured, then having a high proportion of people score above this threshold is not a true ceiling effect because there is no need to try to discriminate further between these individuals (de Vet, Terwee, et al., 2011d).

The distribution of item responses will be examined for all sections of the PMR-PROM. For constructs where the underlying model is reflective (i.e., the items are indicators of the unobservable construct), specific statistical techniques can be applied to identify sub-scales (also known as dimensions or factors) within the long list of items and identify items that do not contribute to the construct and are therefore able to be deleted. This is applicable to the functional and psychological sections of the PMR-PROM and the process for this is explained in the next sections.

8.3.3 Examining the dimensionality of the data

Identification of different dimensions within the long list of items is important for the scoring of questionnaire and for interpretation of the results.

Factor analysis

As introduced in Chapter 6 (6.6.1), factor analysis is the most commonly used statistical process to examine the dimensionality of data within the Classical Test Theory paradigm. It is used to distinguish meaningful dimensions in data by identifying the inter-relationships within a large set of variables and allowing these variables to be grouped into smaller sets that have common characteristics (Nunnally & Bernstein, 1994).

The main principle of factor analysis is 'between item correlation' i.e. items with high correlation cluster within one factor and have low correlation with items belonging to other factors. It determines how many meaningful dimensions can be distinguished within one construct and allows removal or modification of items that do not cluster within one of the factors.

Exploratory factor analysis (EFA) is used to determine the number of dimensions within an instrument or scale and can help decisions on item reduction. It is usually applied within the development phase of an instrument. Confirmatory factor analysis (CFA) tests whether the data fit a predetermined structure and is more appropriate if there are hypothesised dimensions from theory or previous analyses (Pett et al., 2011a).

The most commonly used statistical method for exploratory factor analysis is principal component analysis (PCA). There are theoretical differences between the approaches of true factor analysis and PCA in that FA tries to explain the maximum amount of common variance (the extent to which the individual scores for a particular item deviate from the

item's mean) in the matrix using the smallest number of explanatory factors whereas PCA tries to explain the maximum amount of total variance in the matrix by transforming the original variables into linear components. However, the results are often similar and PCA is the simplest method to carry out (Field, 2009, Chapter 17).

The process involves creating a correlation matrix of the variables (Pett et al., 2011c). The diagonal elements are all '1' because each variable correlates perfectly with itself. The other elements are the correlation coefficients between the pairs of variables or items.

Factors / components can be visualised as the axis of a high-dimensional graph along which the variables are plotted. The coordinates of variables along each axis represent the strength of the relationship between that variable and the factor. Ideally the variable will have a large coordinate for one of the axes and small coordinates for the others which would suggest that it is related to only one factor. The coordinate of a variable along a classification axis is known as a factor loading and this gives information about the contribution of that variable to the factor.

The variance for each variable will have two components - some will be shared with other variables (common variance) and some will be unique to that variable (unique variance). The proportion of common variance present in a variable is known as the communality. A variable that has no unique variance will have a communality of 1 and a variable that shares none of its variance with any other variable will have a communality of 0. PCA assumes all variance is common and sets the communality of every variable to 1 – an assumption that can be criticised, but which makes the statistics simpler (Field, 2009, Chapter 17).

The stepwise process of factor analysis is outlined below:

- Step 1: correlation of items. An inter-item correlation matrix presenting the correlation of all items with each other is produced. Those items that do not correlate strongly with any others (<0.3) can be discarded. Those that show very high correlation (>0.9) need careful consideration, as they can cause problems with interpretation (Pett et al., 2011c). Variables negatively correlated with the others may need a reverse score to facilitate interpretation at a later stage. Sampling adequacy is also checked at this stage using the Kaiser-Mayer-Olkin (KMO) measure (Kaiser, 1970). This is the ratio of the squared correlation between variables to the squared partial correlation between variables. A KMO value of 0 indicates that the pattern of correlations is such that factor analysis is not appropriate whereas a value of 1 indicates that the patterns of correlations are compact and factor analysis should result in distinct and reliable factors.
- Step 2: extraction of the factors. A table of eigenvalues (the eigenvalue of a factor is the sum of squared factor loadings representing the total amount of variance in a data set explained by this factor (Field, 2009, Chapter 17), percentage variance and cumulative percentage variance is produced. In principle, factors with large eigenvalues are retained and those with small ones are discarded. However, there are different ways of determining whether an eigenvalue is large enough to represent a meaningful factor. One method is to keep those with an eigenvalue >1 (the Kaiser-Guttman rule) (Pett et al., 2011c). Another method is to create a scree plot (Cattell, 1966), which is a plot of the extracted factors against their eigenvalues, and look for where there are distinct breaks in the slope of the plot.

It is also important to look at the cumulative variance as if this is low, more factors might be retained to give a better account of the variance.

- Step 3: rotation of the factors. Once factors are extracted it is possible to calculate the degree to which variables load onto each of these factors. Most variables will have high loadings onto the most important factor and smaller loadings onto other factors. This makes interpretation difficult. Rotating the axes of a plot of the variables on the factors facilitates interpretation as it results in factor loadings that are close to 1 or 0 and thus helps to discriminate them. Varimax rotation (a type of orthogonal rotation) is the most often used method and aims to maximize the dispersion of loadings within the factors, trying to load a smaller number of variables highly onto each factor resulting in more interpretable clusters of factors.
- Step 4: interpretation of the factors. A subjective choice is made as to what the common 'thing' is that items loading on the same factor are measuring and the factor is given a name that reflects the meaning of these items.

8.3.4 Item reduction

Studying the distribution of item responses and missing items may allow deletion of some items but further reduction can be achieved by looking for items that have low factor loadings on any factor and deleting them. A minimum loading of 0.5 is usually taken as a threshold (de Vet, Terwee, et al., 2011c). Items that load substantially (>0.3) onto more than one factor also need consideration as these may make interpretation difficult. After each deletion, factor analysis has to be carried out again as deletion of one item may change the loadings of the others.

Inter-item and item-total correlations can also be used to aid item reduction. After factor analysis the inter-item correlations within one dimension should be between 0.2 – 0.5. If correlation is >0.7, the two items are measuring the same thing and one can be deleted. Item-total correlation gives an indication of whether the items discriminate between participants in the construct under study. An item with an item-total correlation of <0.3 does not contribute much to the distinction between mildly and highly affected patients and is a candidate for deletion.

8.3.5 Internal consistency testing

Once factor analysis has been used to show which items cluster into one dimension, the functioning of the items within each unidimensional scale needs to be examined.

The relationships between items are assessed by evaluating their internal consistency – the degree of the interrelatedness among the items. In a unidimensional subscale of a multi-item instrument, the internal consistency is a measure of the extent to which items assess the same construct.

This can be assessed using Crohnbach's alpha (Crohnbach, 1951) , which works on the principle of splitting the list of items in half and testing if the scores of the two half-scales correlate. A scale can be split in half many ways and Cronbach's alpha represents a mean-value of each calculated correlation, adjusted for test length. The accepted value for a good Cronbach's alpha is 0.7-0.9 (de Vet, Terwee, et al., 2011c). A value of >0.9 indicates that there is redundancy of the items. It can be calculated with each item omitted in turn to see which items can be deleted before Cronbach's alpha reduces to within acceptable levels.

It is important to note that Cronbach's alpha is not a measure of validity as it only assesses whether the items measure the same construct, not whether they measure the construct they claim to measure. Instead, it is a dimension of reliability and is included as such in the COSMIN taxonomy (Mokkink et al., 2010b). Another important feature of Cronbach's alpha is that it is dependent on the number of items in the set. Increasing the number of items will increase alpha, even when the correlations among the items are small (Pett et al., 2011b).

8.3.6 Rasch models

IRT was introduced in Chapter 6 (6.6.1) and is a measurement theory applicable when items can be ordered hierarchically (Lord et al., 1968). It is an alternative to CTT with the advantages of allowing a more detailed examination of the distribution of items on the scale and assessment of characteristics such as item difficulty and item discrimination.

Rasch models (Rasch, 1960) are algebraically, a type of IRT model - a one-parameter logistic model in which all the curves have the same 'S' shape. Rasch analysis compares the response patterns of individuals to the entire sample to estimate person ability and item difficulty. The measures produced are equal interval scales, common to both persons and items (Duncan et al., 2003).

Despite being mathematically similar, Rasch and other IRT approaches have philosophical differences and some view them as coming from fundamentally different paradigms (Andrich, 2003). In the traditional statistical paradigms and in IRT, data is examined to see which model it fits best and in general, more complex models that better account for the data are retained. In the Rasch paradigm, the model is taken as the starting point and

data is examined to identify and eliminate misfitting items or people until a good fit is achieved. In a Rasch model the only factor affecting the intra-class correlation of the various items in the test is the item difficulty so the slopes and intercepts of the curves are the same for each item. By fitting data to this, an interval scale (like a ruler) can be created which allows fundamental measurement (i.e., magnitudes can be observed directly, rather than being derived from other measurements, and addition or subtraction of scores is empirically meaningful).

The construct (referred to as a 'trait' or 'latent ability') is considered to lie upon a liner ruler with 'maximum disability' at one end and 'no disability' (unaffected) at the other. A person's ability is expressed in logits, the natural logarithm of the odds of a person being able to perform a particular task (Duncan et al., 2003).

Item difficulty is also expressed in logits on the same linear scale. When person and item locations are equal, the probability of a person reporting difficulty with that item is 0.5.

If the items have multiple response options (i.e., the model is polytomous), there will be a level of difficulty for each response option of each item. The 'threshold' of each of these is the point where the probability of being in that response category as opposed to the next one is 0.5. In a perfect scale, each of these thresholds lie equidistant from each other on the continuum of the construct. This is rare in practice but if this is the case, a Rating Scale Model framework (Andrich, 1978) can be used. If the distance between the thresholds varies, a Partial Credit Scale Model framework (Masters, 1982) has to be used, which is more complex.

8.3.7 The process of testing fit to a Rasch model

When developing a new instrument, the items can be selected in such a way that the data fits a Rasch model, with the justification that the instrument will be stronger if it conforms to this model.

Computer software packages

Several software packages exist which can estimate the Rasch model and provide tests of fit of a data set to the model. For this study, the program RUMM2020 (Andrich et al., 2003) will be used.

Model estimation procedures

For polytomous models, a likelihood ratio test can be run in RUMM2020 to determine whether a Rating Scale Model can be used or whether the unrestricted Partial Credit Model should be applied.

Class intervals

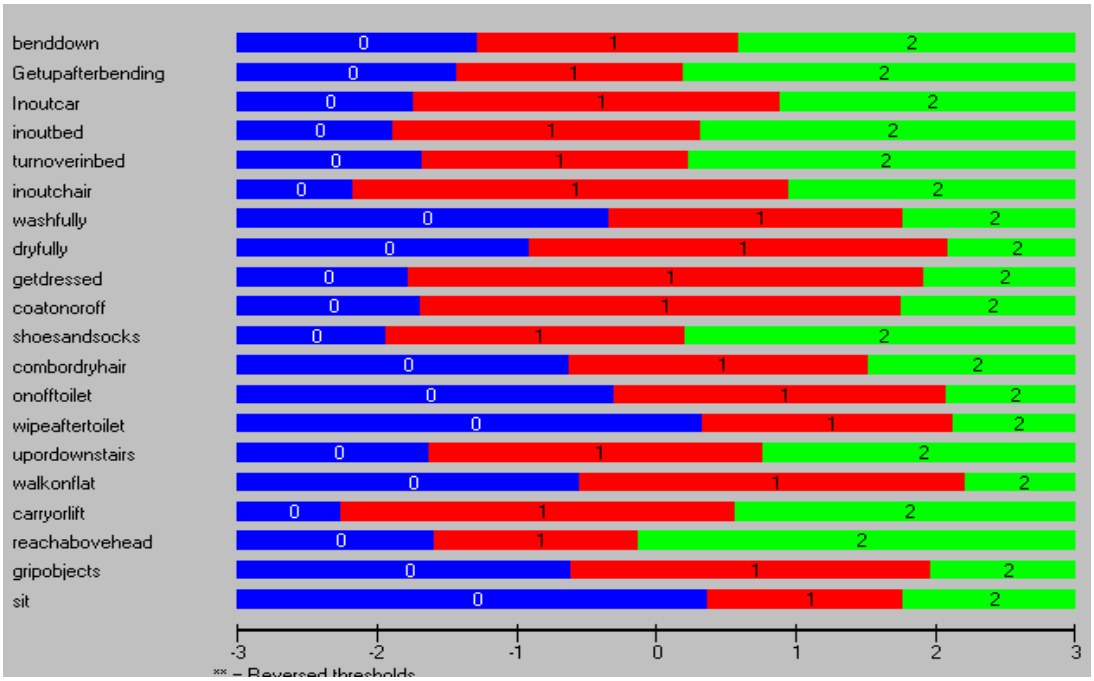
Class intervals are groups of individuals defined by rank order of person locations (abilities) according to the logit score produced by the Rasch estimation procedure (Andrich, 2003). The only purpose of forming class intervals is to test the fit of the data to the Rasch model in the RUMM package. The number of class intervals is determined by the sample size but there should be approximately equal numbers in each class interval. It is recommended to have around 50 individuals in each (Tennant & Conaghan, 2007)

Threshold ordering

The threshold point is the point on the scale between two scores for a specific item where it is equally likely that an individual will obtain either score e.g. the point at which the probability of scoring a '1' or a '2' is 50/50.

Whilst it is assumed that respondents will use the response options in the order that they were intended, this might not be the case particularly if there are many response options or ambiguity in the wording. A threshold map can be created which shows the ordering of the categories as interpreted by the respondents (see Figure 8.1). Disordered thresholds mean that an item is not working properly and its scoring categories do not progress in a logical order.

Figure 8.1: Example of an appropriately ordered threshold map

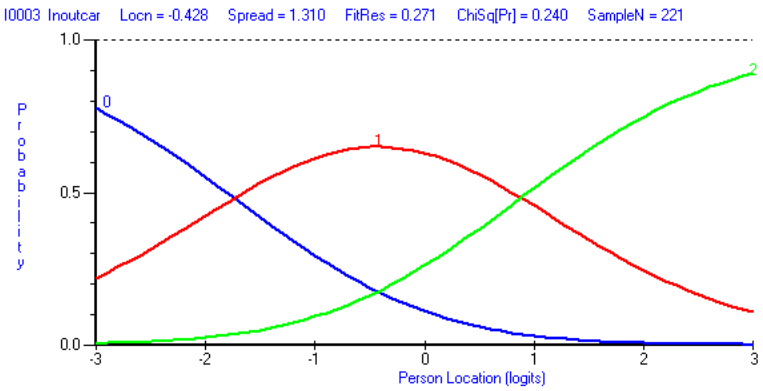


Another way of looking at this is to create a plot of the latent trait on the x-axis against the probability of response on the y-axis, for a single item. If the thresholds are ordered

correctly, there will be a point at which each response option is most likely. Figure 8.2 is a representation of one of the items from Figure 8.1 and confirms that the ordering of the thresholds is correct.

Disordered thresholds can be addressed by combining response categories to re-score the items.

Figure 8.2: Example of category probability curves showing appropriate threshold ordering



Unidimensionality

The Rasch model assumes that the items are measuring a single construct and this needs to be formally tested during the analysis process. In RUMM2020 this is done by carrying out Principal Component Analysis (PCA) of the two most different groups of items within the scale thus calculating two different person locations for each individual. An independent t-test comparing these person locations is carried out to see if they are significantly different. If the scale is unidimensional, the responses to any subset of items within one scale should give the same estimate of person ability.

It is suggested that for a scale to be considered unidimensional, no more than 5% of people should have person locations from the two sets of items that are significantly different at the 5% level (Tennant & Conaghan, 2007) .

Local response dependency

The Rasch model assumes that the responses to items are statistically independent. If a person's response to one item is determined by their response to another item, misfit can occur. This can be assessed using the residual correlations between items i.e., correlations between items after the 'Rasch factor' has been removed. If there is no local dependency, there should not be any pattern of correlation in this residual data. Residual correlations of >0.3 are usually considered to be indicative of local dependency (Andrich, 2003)

Targeting

In Rasch software the scale is always centred on zero logits representing the item of average difficulty for the scale. Comparison of the mean location score for persons with the value of 0 set for the items, provides an indication of how well targeted the items are for people in the sample (Tennant & Conaghan, 2007). If the measure is well-targeted, the mean location for persons will be close to 0. A positive mean value for persons indicates that the sample as a whole is located at a higher level of ability than the average of the scale whilst a negative value suggests the opposite.

Overall fit

Overall fit to the Rasch model can be assessed using a chi-square test to examine whether the data fit the model for each class interval (the overall chi-square value is called the item-trait interaction statistic). If this is significant it indicates that the ordering of the

items is not the same at all levels of the latent trait and the data are a poor fit to the Rasch model. Misfit suggests that the items are assessing something in addition to, or other than, the construct of interest.

Mean item and person fit residuals also give an indication of overall fit to the model. These residuals represent the divergence between the observed data and expected values and can be calculated in RUMM2020. They are transformed to approximate a Z-score which represents a standard normal distribution, so the distribution of the mean fit residuals ought to have a mean of approximately 0 and a SD of 1. Mean fit residuals of -0.4 to 0.4 with SD of <1.4 are generally considered acceptable (Andrich et al., 2003).

A power of test-of-fit can also be calculated, which gives an idea of the reliability of the fit statistics and is categorised as excellent, good, reasonable, low or too low. It is based on the person-separation index, which reflects the power of the construct to discriminate amongst the respondents – the higher the value, the greater number of groups that can be discriminated. If the person-separation index is low (as is would be in a homogenous group of people), the fit statistics are less reliable.

Individual Item fit

There is no single statistic that assesses the fit of individual items to a Rasch model, but a series of statistics can be used to build a picture of an item's fit (Hendriks et al., 2012).

Firstly, individual item fit residual statistics can be considered. As described above, these are standardized to a z-score and in this case, values outside the range -2.5 to 2.5 indicate that the item is not a good fit to the model at the 99% confidence level (Andrich et al., 2003). Large negative residuals indicate item redundancy and large positive residuals

indicate that the item is not able to distinguish between people at different levels of the latent trait.

Secondly, chi-square statistics for each item can be used. These are calculated from the difference between the expected values for each individual in the sample and the actual values observed. These are then summed across the class intervals to get an overall chi-square statistic for each item. If this is significant, it suggests the item does not fit the Rasch model well.

Thirdly, item characteristic curves can be plotted to visually represent each item's fit. The person location is on the x-axis and the observed scores for the probability of reporting difficulty with an item for each class interval are shown as dots against the curved line representing the expected scores for this item (see Figure 8.3).

Figure 8.3: Example of a Guttman Curve and a Rasch Item Characteristic Curve for a single dichotomous (yes/no) item

(taken from (Hendriks et al., 2012))

Individual Person fit

Individual person fit residuals indicate how closely a person's responses fit a Guttman pattern. Values outside of the range -2.5 to 2.5 suggest misfit. Very negative values indicate a purer Guttman pattern than expected, suggesting a person may be responding in a fixed way, whilst very positive values indicate that a person is responding in an unexpectedly disordered way (Hendriks et al., 2012) . These individuals may need to be

removed as they can skew the whole analysis but removing people on this basis can have implications for the generalisability of the scale.

Differential item functioning

Rasch analysis also allows assessment of differential item functioning (DIF). DIF describes the situation where people from certain sub-populations (e.g., males versus females) with the same severity of disease, do not score the same on the relevant item (Teresi et al., 2000). In other words, the item is measuring different things in the different populations. DIF can either be uniform (where the item is easier or more difficult for one population at all levels of the construct) or non-uniform (where the item is easier for one population at a certain level of the construct but more difficult for that population at another). DIF can be assessed for using CTT or logistic regression techniques but IRT, including the Rasch model) is a particularly powerful way of detecting it. The item characteristic curves from the populations of interest are compared to see if the people from the different populations who have an equal score on the latent trait (construct) have a same probability of endorsing the item. Where there is non-uniform DIF present, the item characteristic curves will cross. If DIF is identified, the reasons for it need to be considered. Even if it is statistically significant, it may not be of any clinical relevance but if it is felt to be important, adjusted scores may be able to be calculated. Significant DIF can also be a reason to remove an item if there is item redundancy.

Examining item and person distribution

A plot showing an overview of items and the population depicted at the same trait level (level of the construct) can help visualise where the patients are in relation to where the

items are. For polytomous data, the person-threshold location distribution is used and each threshold has a location plotted on the scale.

If there are a lot of patients on locations of the scale with only a few items, more items may need to be developed in this range of the scale otherwise the ability to discriminate between these patients will be poor. Sparseness of items at either end of a scale causes risk of floor and ceiling effects as discussed in Section 8.3.2.

Where there are items measuring the same level of difficulty, these are candidates for deletion.

8.4 Methods

8.4.1 Protocol development

The protocol for this study was developed in March 2018. Copies of the practice invitation letter, participant invitation letter and participant information sheet are given in Appendix 8.1: Practice invitation letter, Appendix 8.2: Participant invitation letter and Appendix 8.3: Participant information sheet.

8.4.2 Ethics and governance

NHS Health Research Authority (HRA) approval was given on April 18th, 2018 and a favourable opinion was given by the Proportionate Review Sub-committee of the North East - York Research Ethics Committee on April 20th, 2018 (REC reference 18/NE/0140). The approval letter is included as Appendix 8.4: Confirmation of ethical approval. No amendments were made during the course of the study.

8.4.3 Version of the PROM and method of use

Version 5 of the PMR-PROM was used in this study (see Appendix 7.8: PMR-PROM Version 5). Development of the PROM to this point has been described in previous chapters. To be used in this study, I formatted it into a booklet with the Keele University logo and instructions for use on the front page.

To ensure that the final PROM is valid for use across the whole PMR disease course it was important to gather responses from people at all stages of the condition. However, the relatively low incidence of the condition and the usual rapid response to treatment mean that recruiting people at the point of diagnosis, when they are maximally affected, is challenging. I therefore decided to ask people to complete the questionnaire twice; once reflecting how they felt at that point in time and once remembering back to how they felt at diagnosis.

Collecting PROM data retrospectively may not be reliable for two reasons; recall bias and response shift. Recall bias occurs when memories area affected by time and is influenced by the interval between the event and the time of its assessment but also by the impact of the event being recalled (Schmier & Halpern, 2004). Response shift describes the changes in perception that can occur when circumstances change due to reconceptualisation (e.g., an individual redefining what a good quality of life is for them), reprioritization and recalibration (e.g., an individual experiencing new levels of a symptom may change how they rated it previously) (de Vet, Terwee, et al., 2011d). However, a systematic review of six studies comparing scores of contemporaneously and retrospectively collected PROMs found moderate to strong agreement for most, although the agreement was stronger when the time interval was shorter (Kwong & Black, 2017).

Therefore, acknowledging that there are some flaws in using the method of asking people to remember back to the point of diagnosis (and that it places extra burden on individual participants) I felt it was the best way to try to capture the impact of PMR across the disease course. The significant burden of symptoms at diagnosis found in previous qualitative work meant that it seemed likely that people would remember how they felt at that time.

The questionnaires for the different time points were printed on different coloured paper with different instructions on the front and as a header on each page as a prompt to fill in one thinking back to how they would have answered when they were first diagnosed.

8.4.4 Sample size

For factor analysis it is recommended that 3-5 times the number of respondents than the number of items on the questionnaire are needed (Norman & Streiner, 2008). Version 5 of the PMR-PROM has 61 possible candidate items though only 38 of these are in sections that will be subject to factor analysis.

For Rasch modelling, the sample size needed to obtain stable item or person calibrations depends on its modelled standard error. For high stakes situations where it is desirable to have 99% confidence that no item calibration is more than one logit away from its stable value, the sample size for most purposes is recommended to be 250 (Linacre, 1994).

For this study therefore, the target was 250 respondents. Assuming a non-response rate of 30% (based on past experience of working with this patient population), the questionnaire needed to be sent to 400 people to achieve sufficient respondents.

8.4.5 Recruitment of practices

Practices were recruited through the NIHR West Midlands Clinical Research Network, which comprises 320 practices of whom 280 participated in research during the previous year.

The CRN advertised the study to practices and if they were willing to participate, they were sent the full study information. Practices were reimbursed for their time following CRN guidance.

8.4.6 Identification of potential participants

Practices were asked to run a search of their patient databases to identify people diagnosed with PMR within the preceding two years. Either a GP or nurse from the practice or a member of the CRN team with appropriate access rights, then screened this list of patients against the inclusion and exclusion criteria.

Inclusion criteria:

To be included, people needed a diagnosis of PMR made within the previous 2 years and not subsequently changed.

The diagnosis should be supported by the following features, which are based on the British Society for Rheumatology / British Society for Health Professionals in Rheumatology guidelines (Dasgupta, Borg, & Hassan, 2010):

- Age > 50 years.
- Bilateral shoulder or pelvic girdle aching or both for at least 2 weeks.
- Morning stiffness.
- Evidence of an acute phase response (raised ESR / CRP).
- Diagnosis made by a rheumatologist despite the presence of atypical features (e.g. normal ESR / CRP).

Exclusion criteria:

- Diagnosis of GCA
- Inability to read / write English well enough to understand the instructions and complete the questionnaire.
- Comorbidities that made an invitation to participate in the study inappropriate in the view of the participant's GP (dementia, significant anxiety / depression, receiving end of life care etc.).

Invitation letters and study documents were sent from the practice to potential participants with a reply envelope addressed to me at the School of Primary, Community and Social Care, Keele University.

As no personally identifiable data was being collected, return of the questionnaires was taken as implicit consent to participate. The patient information sheet made this clear and explained how the data was going be used.

8.4.7 Data entry

Data was entered into SPSS (IBM, 2014) Two databases were created, one for responses to the contemporaneously completed PROM ('now' dataset) and one for the one completed retrospectively ('at diagnosis' dataset).

8.4.8 Data analysis

Calculation of descriptive statistics for each dataset, analysis of distribution of item responses and exploratory factor analysis was carried out in SPSS (IBM, 2014).

Data was then imported into RUMM2020 (Andrich et al., 2003) to carry out assessment of fit to a Rasch model.

8.5 Results

8.5.1 Response rate

449 questionnaire packs were sent out in a single mailout. 271 responses were received giving a response rate of 60.4%. Eight were excluded as they were returned either

entirely blank or partially completed such that there was insufficient data for them to be used. Seven were excluded as the individual was diagnosed more than five years ago. No data was available on non-responders due to it being an anonymous survey.

A Health Technology Assessment on design and use of questionnaires (McColl et al., 2001) and reported that although published standards for response rates for postal surveys suggest rates of 60-69% are 'acceptable' with only higher rates considered good or excellent, rates of response actually published in the medical literature are generally low and typically around 60%. However, the reason for striving for a high response rate is to ensure that the sample size is satisfactory to minimise risk of bias. Given that the purpose for which the data is being collected in this study is not to draw any inferences about the sample itself and that the sample size achieved was adequate for the statistics to be applied, my response rate of 60.4% is satisfactory.

The inclusion criteria stated that participants had to be diagnosed within the preceding two years. However, there were 21 responses from people who were diagnosed longer ago than this, 13 of whom reported they were still on prednisolone treatment and can therefore reasonably be classed as still having 'active' PMR. After discussion with the study advisory group, I decided to extend the cut off to include those diagnosed up to five years ago. Excluding those diagnosed more than five years ago excluded seven respondents (of whom four were still taking prednisolone). This seemed an appropriate cut off based on the typical natural history of the condition and reliability of the data due to memory effects, to maximise the use of the data I had obtained without introducing significant bias.

There were five respondents for whom the question on duration since diagnosis was not answered and these were also included in the analysis.

It is not known why study packs were sent to people outside of our stated inclusion criteria, but it is likely to be due to coding anomalies in primary care records (e.g., polymyalgia rheumatica may have been entered as a 'new' diagnosis more than once) or it may have simply been error on the part of the person screening potential participants.

8.5.2 Descriptive statistics

256 paired questionnaires were suitable for analysis. Demographic details of the 256 participants are given in Table 8.1.

Table 8.1: Demographic details of participants

Category	Responses	Missing responses (%)
Mean age (range)	73.9 (52-98)	0.39
Gender (% female)	67.1	0.39
Mean duration since diagnosis (range)	17.5 months (0.5-60)	1.95
Referred to rheumatology (%)	34.4	1.56
Mean number of medications in addition to prednisolone (range)	3.6 (0-13)	9.77

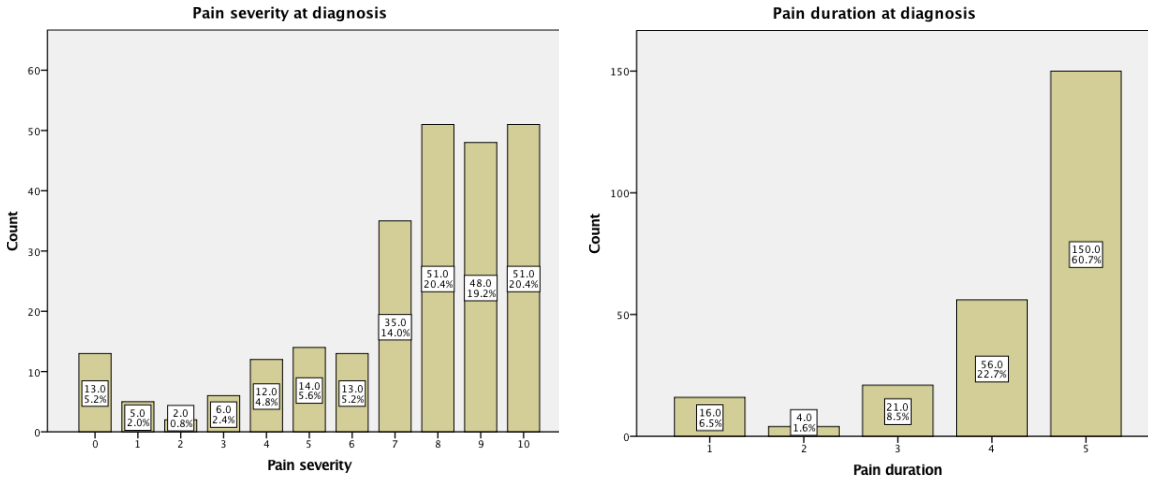
The largest UK primary care cohort study of PMR (Muller et al., 2016), which included 652 patients, reported a mean age of 72.4 years with 62.4% being female. The results of the sample in this study are similar, suggesting that the sample demographic fairly reflects the UK PMR population.

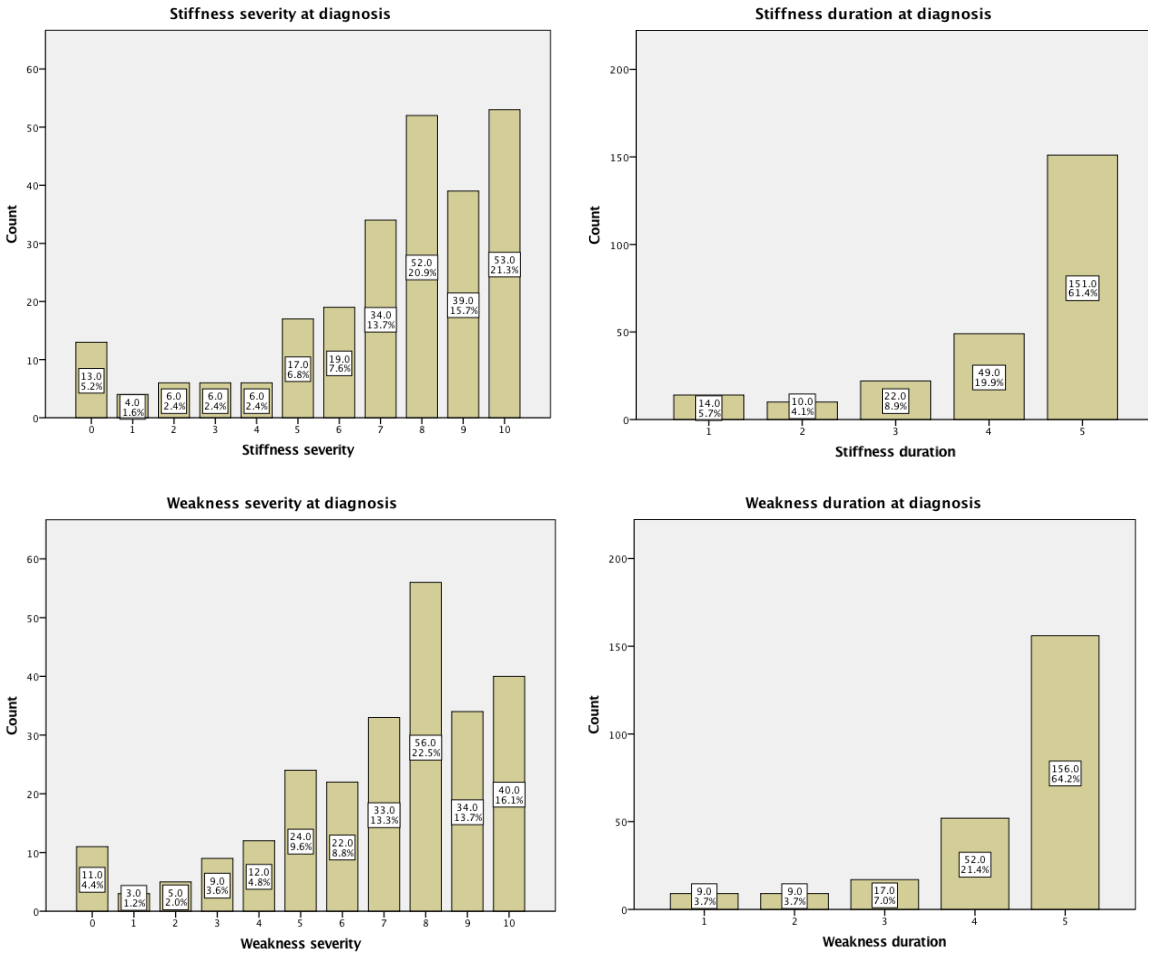
8.5.3 Distribution of responses for pain, stiffness and weakness questions

Tables of results of the distribution of responses to the questions on severity and duration of symptoms at both time points are given in Appendix 8.5: Results tables of distribution of item responses.

Bar charts visually depicting the data are shown in Figure 8.4 and Figure 8.5.

Figure 8.4: Bar charts depicting distribution of responses to questions on pain, stiffness and weakness at diagnosis





Question asked on severity: how severe has the (pain / stiffness / weakness) from your PMR been during the last 3 days?

Responses options: visual analogue scale (VAS), scored from 0-10 where 0 = no pain and 10 = the worst pain you have ever felt.

Question asked on duration**:** on average much of each day has the (pain / stiffness / weakness) from your PMR been present for during the last 3 days?

Response options: 1 = less than 30 mins, 2 = less than 1 hour, 3 = around 1-3 hours, 4 = about half the day, 5 = all day

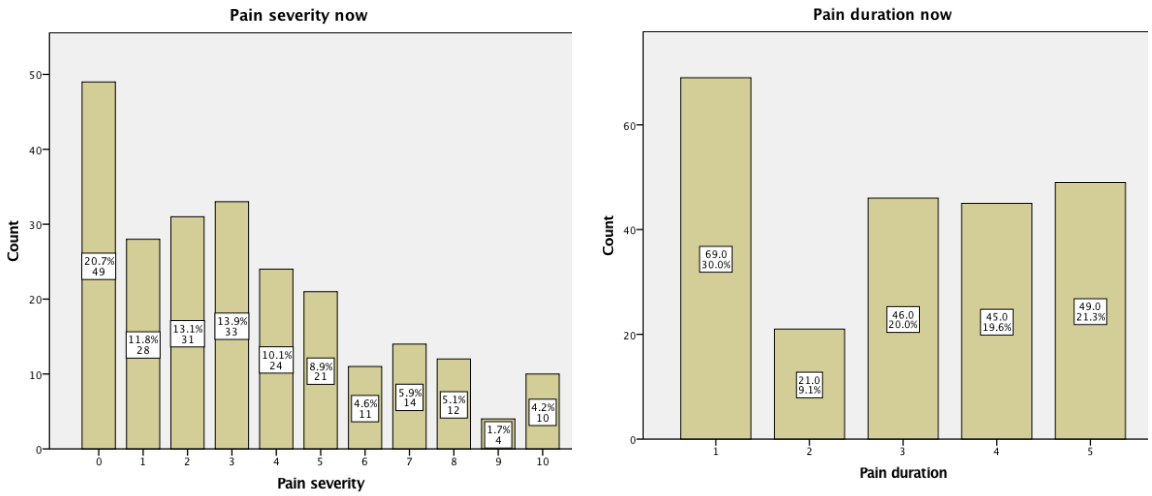
These results show that most respondents report high scores for pain, stiffness and weakness at diagnosis and in a high proportion, these symptoms last either half the day or all day. There are some respondents who report no pain, stiffness or weakness at diagnosis which is unexpected but could be due to misinterpretation of the direction of

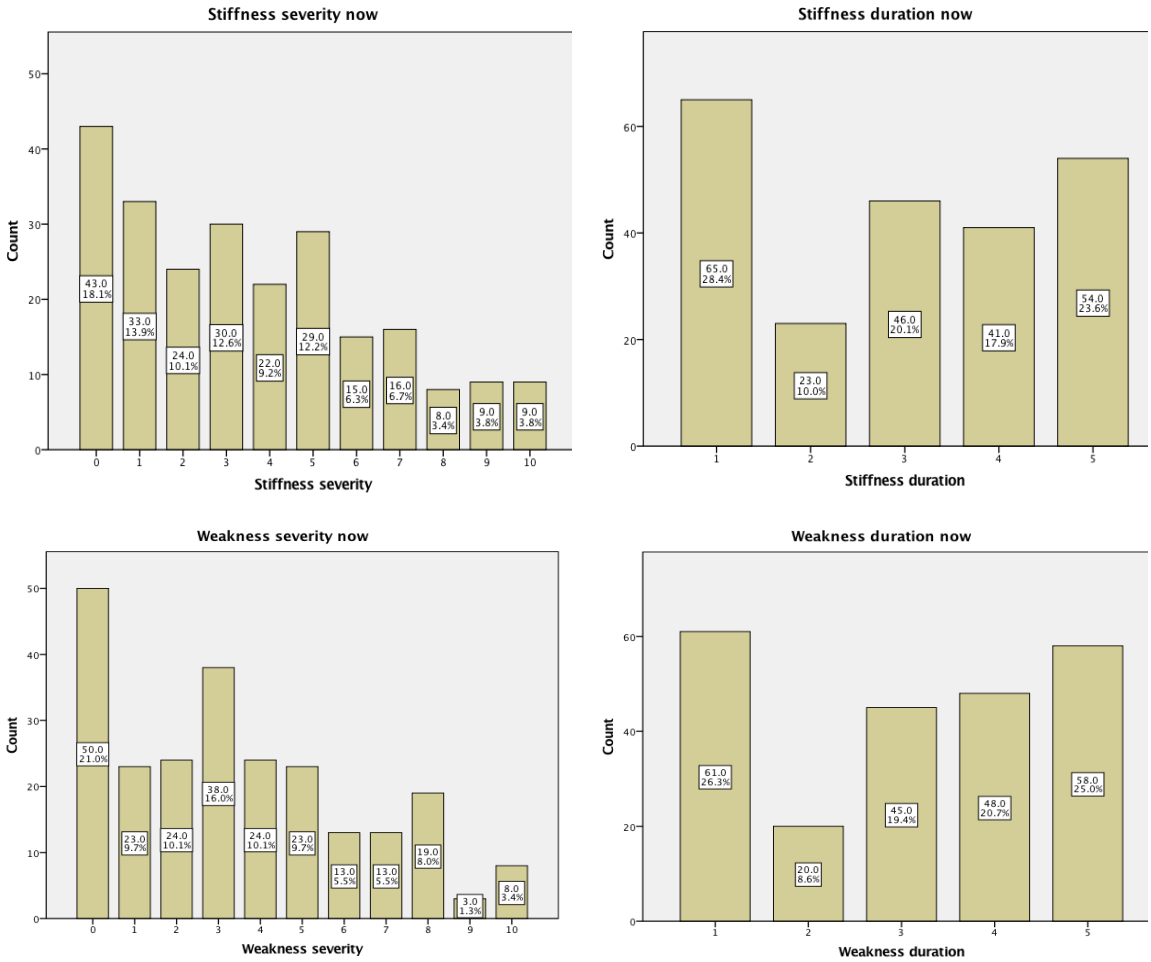
the scale. In the duration question, the response option of 'less than one hour' was least often chosen and might not be adding to the ability to discriminate between groups.

Missing responses were <6% for all questions which is acceptable (as per the guidance discussed in section 8.3.2).

The fact that more than 15% scored each of either 8, 9 or 10 for symptom severity at diagnosis and more than 15% reported the top two categories for symptom duration suggests a ceiling effect in the ability of the PROM to discriminate between people with severe and long-lasting symptoms in the early stages of the disease course.

Figure 8.5: Bar charts depicting distribution of responses to questions on pain, stiffness and weakness now





Question asked on severity: how severe has the (pain / stiffness / weakness) from your PMR been during the last 3 days?

Responses options: visual analogue scale (VAS), scored from 0-10 where 0 = no pain and 10 = the worst pain you have ever felt.

Question asked on duration**:** on average much of each day has the (pain / stiffness / weakness) from your PMR been present for during the last 3 days?

Response options: 1 = less than 30 mins, 2 = less than 1 hour, 3 = around 1-3 hours, 4 = about half the day, 5 = all day

These results show that severity of pain, stiffness and weakness later on in the disease course is more variable than at diagnosis, with a skew towards the lower end of the severity scale. Nearly 20% report no pain, stiffness or weakness at all (it is not clear from this analysis whether the same people are reporting no pain as are reporting no stiffness or weakness or whether the different symptoms are experienced to differing degrees).

The duration of these three symptoms is also skewed towards the lower end of the scale but there are still significant numbers reporting symptoms lasting much of the day. Again, the response option of 'less than one hour' was least often used.

Missing responses were <11% for all questions.

Having >15% score 0 for pain, stiffness and weakness severity and >15% report the lowest category for pain and stiffness duration (though not for weakness duration), suggests a floor effect. This is less marked than the ceiling effect observed in the 'at diagnosis' data but potentially more significant in terms of utility of the measure (this is considered further in the discussion (Section 8.7.3)).

Resulting amendments to the questions on pain, stiffness and weakness

Based on these results, I decided to alter the response options to the question about duration of pain, stiffness and weakness so that

0 = none

1 = less than 1 hour

2 = around 1-3 hours

3 = half the day

4 = all day

8.5.4 Addition of fatigue to the symptoms questions

During the course of my PhD research, I continued to work with the PMR special interest group of OMERACT towards developing a core outcome measurement set for the condition. It became apparent through patient and public involvement work and a further face validity and feasibility study in which I was involved (Yates et al., 2020) that

fatigue was increasingly being recognized as an important symptom in PMR. This symptom had been acknowledged in both my earlier qualitative work (Twohig et al., 2015) and qualitative / survey research by others (Helliwell et al., 2016; Mackie et al., 2014) but the importance of it as a central symptom had not been fully appreciated. Up until this point, fatigue had appeared in my PROM in the emotional and psychological well-being section but based on emerging evidence, I decided to add it to the symptoms section.

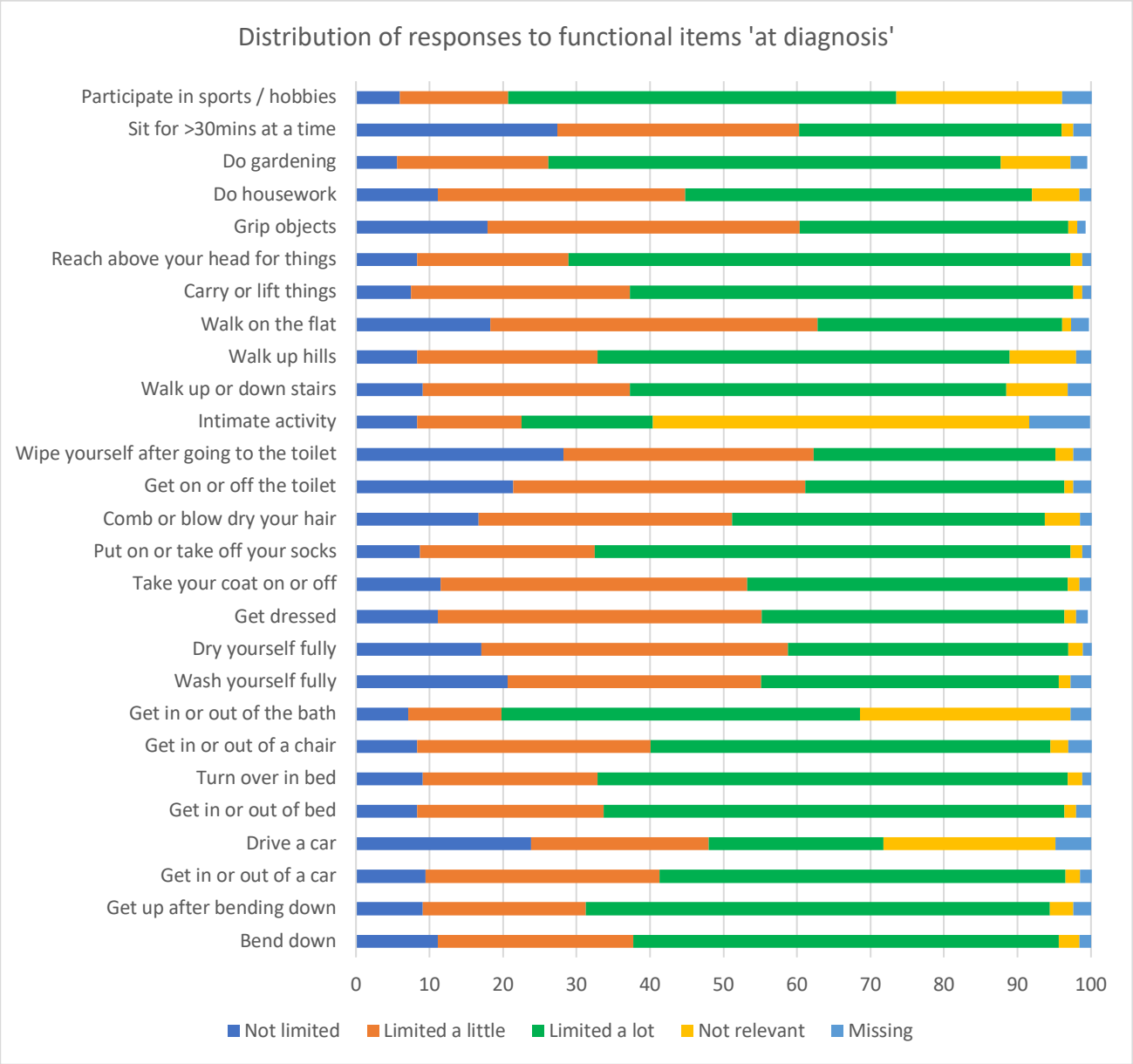
I consulted with a small Research User Group linked to the Research Institute for Primary Care and Health Sciences at Keele University about whether the term 'fatigue', 'reduced energy' or 'tiredness' was preferred, and they felt that 'fatigue' most accurately represented their experience.

I therefore added questions about fatigue severity and duration (worded in the same way as the pain, stiffness and weakness questions) to the symptoms section.

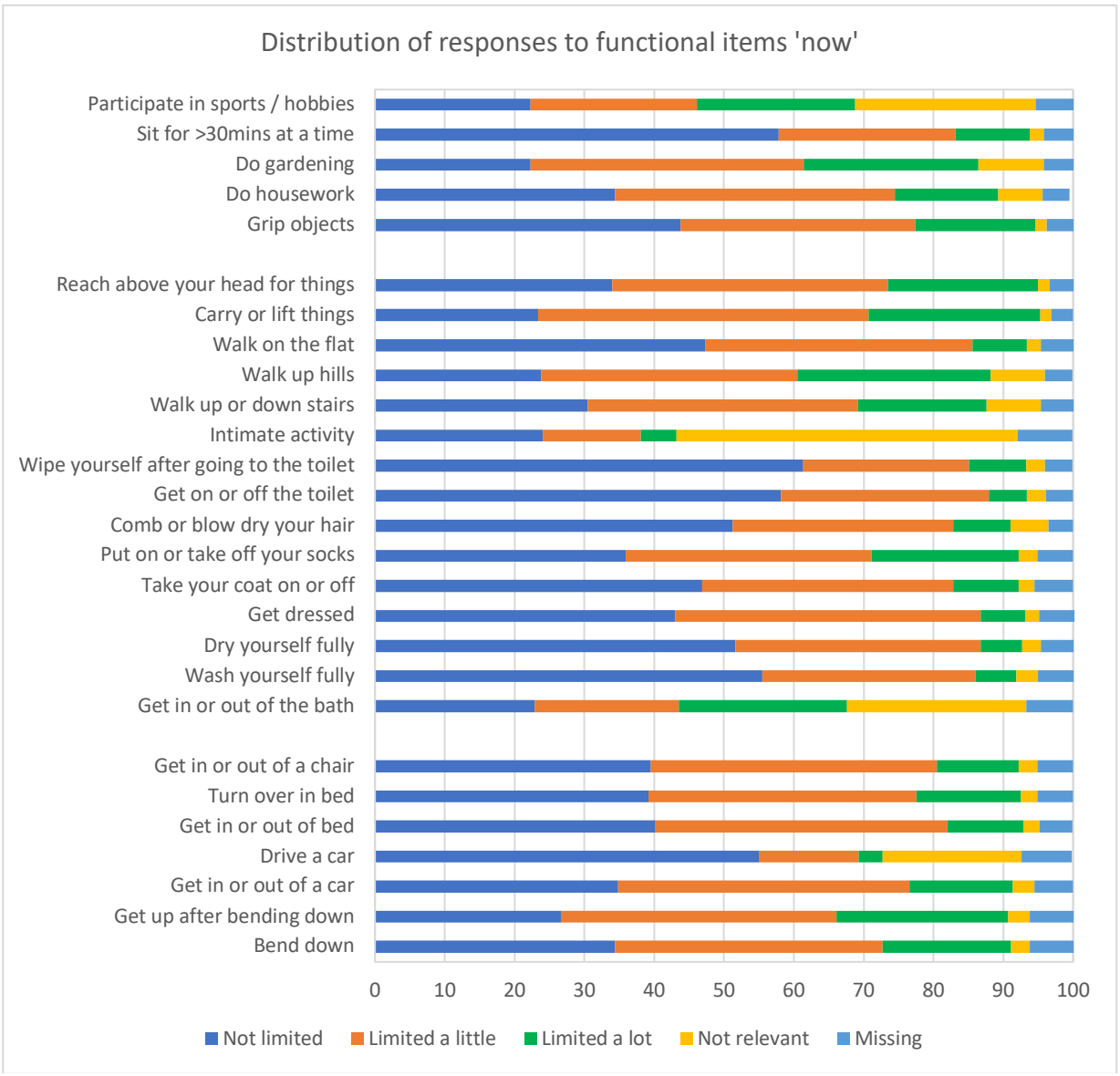
8.5.5 Distribution of responses to the functional activity items

Tables of results of the percentage distribution of responses to the functional activity questions at the two time points are in Appendix 8.5: Results tables of distribution of item responses. Bar charts visually depicting the same data are in Figure 8.6.

Figure 8.6: Bar charts depicting distribution of responses to functional activity items



Question asked**:** Over the last 3 days, compared to what you can normally do, has PMR limited your ability to do the following activities?



Question asked: Over the last 3 days, compared to what you can normally do, has PMR limited your ability to do the following activities?

Missing responses for all functional items were <8%. The item with the highest number of missing responses in both the 'at diagnosis' and 'now' datasets was the question about intimate activity.

The items with the highest percentage of responses in the 'not relevant' category were those about 'driving a car', 'getting in or out of a bath', 'intimate activity' and 'participating in sports or hobbies'.

5 items had >20% responses in the 'not limited' category at diagnosis (driving, sitting for >30 mins, wash yourself fully, wipe yourself after going to the toilet, get on and off the toilet) which suggests that they might be less good items to assess functional limitations caused by PMR.

Resulting amendments to the functional item scale

Based on consideration of the distribution of responses to these items, I decided to apply two principles for exclusion of items:

1. Exclude items that had >20% responses in 'not limited' category at diagnosis (results in exclusion of driving, sitting for >30 mins, wash yourself fully, wipe yourself after going to the toilet, get on and off the toilet)
2. Exclude items with >10% either missing or not relevant in either data set (results in additional exclusion of getting in / out bath, intimate activities, gardening, sports and hobbies, walk up hills, walk up or down stairs, do housework).

The only item which differed between data sets for principle 2 was 'do housework' which had combined missing and not relevant responses of 8% in the 'at diagnosis' dataset and 11.2% in the 'now' dataset.

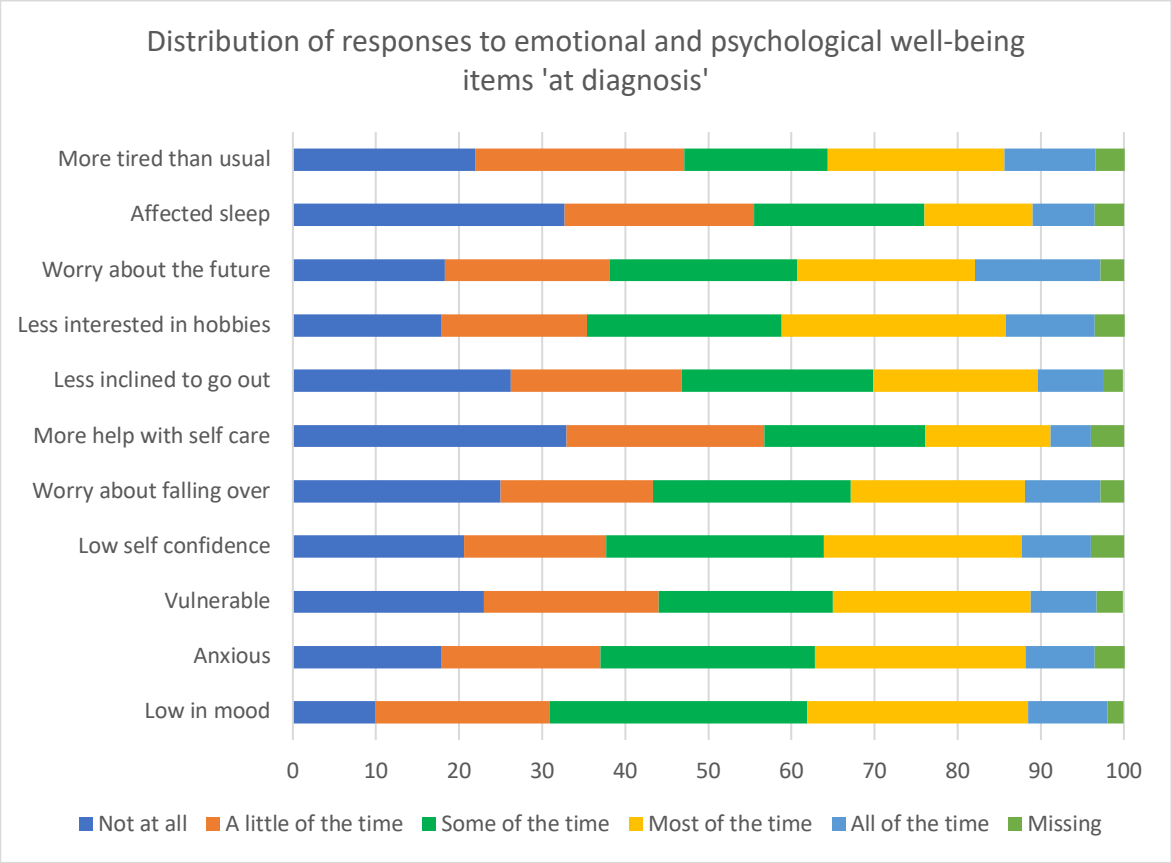
Although the item 'walk up or down stairs' met the criteria for exclusion (most likely because a significant proportion of respondents lived in flats or bungalows and therefore

answered ‘not relevant’), I decided not to exclude it because I felt that it was such an important functional ability in situations where it was relevant to an individual.

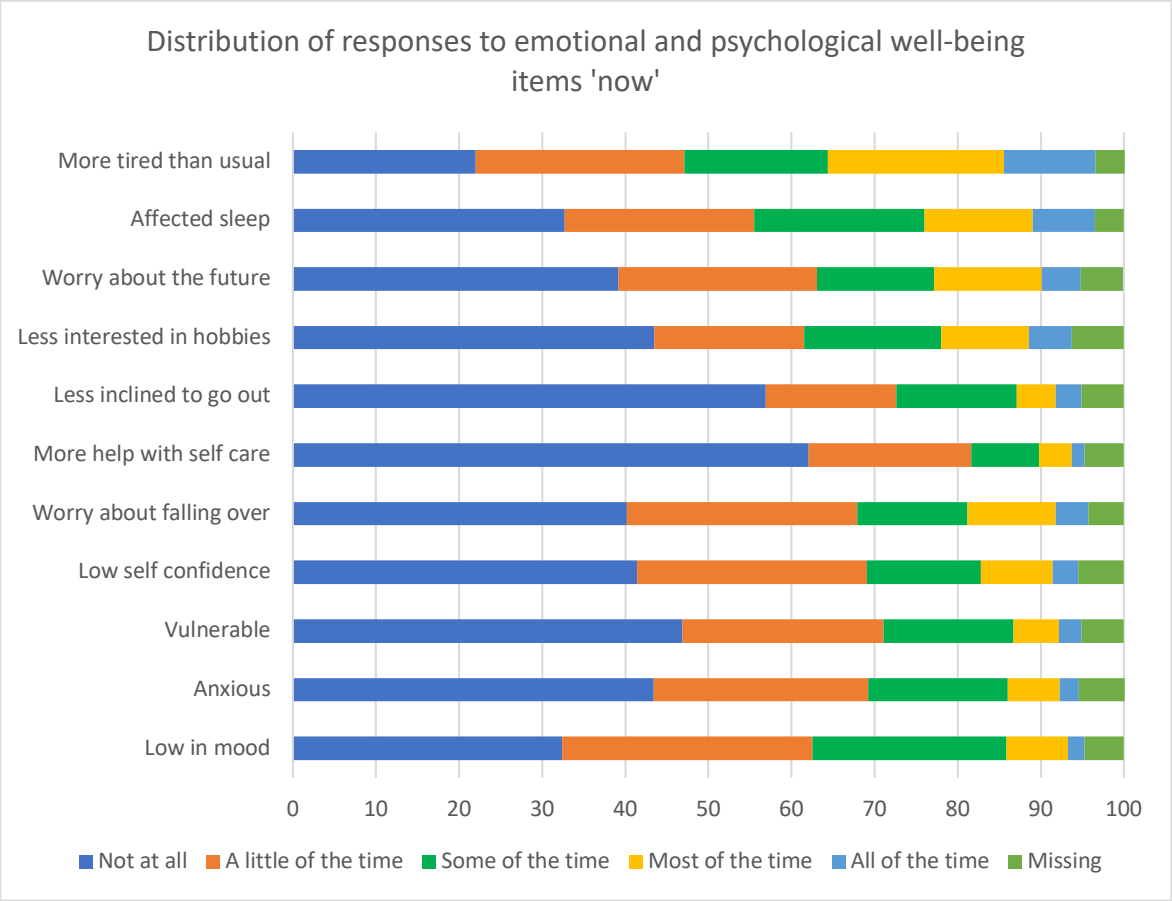
8.5.6 Distribution of responses to the emotional and psychological well-being items

Tables of results of the percentage distribution of responses to the emotional and psychological well-being questions at the two time points are in Appendix 8.5: Results tables of distribution of item responses. Bar charts visually depicting the same data are in Figure 8.7.

Figure 8.7: Bar charts depicting distribution of responses for emotional and psychological well-being items



Question asked: Over the last 3 days, have your PMR symptoms….?



Question asked: Over the last 3 days, have your PMR symptoms….?

Missing responses for all emotional and psychological well-being items were <6%. There are no items that stand out as having much greater numbers of missing responses than others.

In general, the responses to these items were more evenly spread over the categories than responses the functional items although there was still the expected skew towards participants being less affected 'now' that 'at diagnosis'.

Resulting amendments to the emotional and psychological well-being scale

No items were excluded based on the distribution of responses and the response categories were left unchanged.

8.5.7 Analysis of the functional scale using Classical Test Theory

The next step was to perform exploratory factor analysis (EFA) of the reduced functional item scale. This was done using the principal component analysis (PCA) method outlined in Section 8.3.3. The analysis was carried out in SPSS (IBM, 2014).

I used an iterative process, applying the principle that items loading with factor loading <0.5 should be excluded (as per (de Vet, Terwee, et al., 2011c)).

The 16 items subjected to PCA were:

Bend down
Get up after bending down
Get in or out of a car
Get in or out of bed
Turn over in bed
Get in or out of a chair
Dry yourself fully
Get dressed
Take your coat on or off
Take your shoes and socks on or off
Comb or blow-dry your hair
Walk up or down stairs
Walk on the flat
Carry / lift things
Reach above your head
Grip objects

Principal component analysis of functional items at diagnosis

A PCA was conducted on the 16 items with varimax rotation. The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis, KMO = 0.92 ('marvellous' by the Kaiser and Rice guidelines (Field, 2009, Chapter 17)).

The correlation matrix showed there were no correlations <0.3 or >0.9. The diagonal elements of the anti-image correlation matrix were all >0.5.

An initial analysis was run to obtain eigenvalues for each factor in the data. 3 factors had eigenvalues over Kaiser's criterion of 1 and in combination explained 68.4% of the variance (see Table 8.2).

The scree plot (Figure 8.8) shows a point of inflexion at component 3 suggesting 2 factors should be retained.

I decided to retain 3 factors at this stage to keep the analysis as rigorous as possible. The factor loadings after rotation are shown in Table 8.3.

Table 8.2: Eigenvalues associated with each factor before extraction, after extraction and after rotation (PCA 1 of functional items at diagnosis)

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8.634	53.963	53.963	8.634	53.963	53.963	4.954	30.961	30.961
2	1.260	7.876	61.839	1.260	7.876	61.839	3.365	21.033	51.994
3	1.052	6.577	68.416	1.052	6.577	68.416	2.627	16.422	68.416
4	.793	4.954	73.369						
5	.684	4.272	77.641						
6	.602	3.760	81.401						
7	.576	3.599	85.000						
8	.435	2.722	87.722						
9	.412	2.576	90.298						
10	.328	2.049	92.347						
11	.308	1.927	94.274						
12	.252	1.574	95.848						
13	.210	1.312	97.160						
14	.183	1.142	98.302						
15	.154	.963	99.265						
16	.118	.735	100.000						

Extraction Method: Principal Component Analysis.

Figure 8.8: Scree plot for PCA 1 of functional items at diagnosis

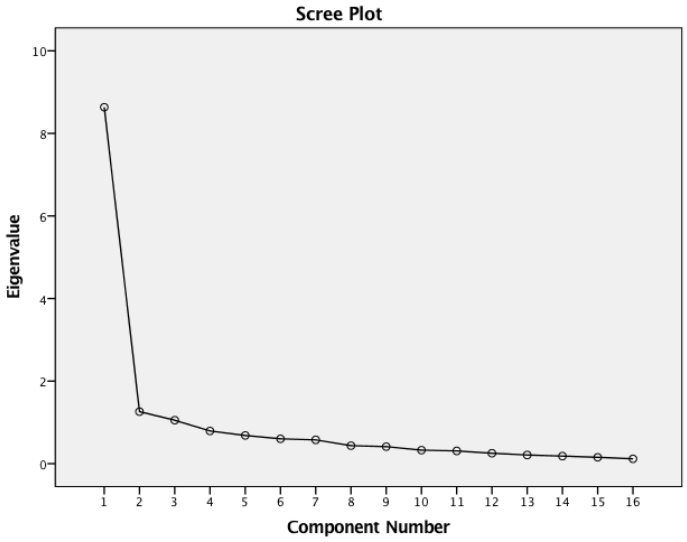


Table 8.3: Factor loadings after rotation (PCA 1 of functional items at diagnosis)

Rotated Component Matrix[a]			
	Component		
	1	2	3
Take your coat on or off	.790		
Get in or out of bed	.784		
Get dressed	.780		
Turn over in bed	.760		
Dry yourself fully after shower or bath	.688		.376
Put on or take off your socks and shoes	.686	.418	
Get in or out of a chair	.634	.446	
Comb or blow dry your hair	.557		.435
Get in or out of a car	.553	.461	
Get up after bending down		.845	
Bend down		.830	
Walk on the flat		.677	.363
Walk up or down stairs	.333	.618	
Grip objects			.780
Reach above your head for things	.387		.762
Carry or lift things	.310	.336	.733

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.[a]
a. Rotation converged in 6 iterations.

These tables show that component 1 explains 54% variance, component 2 explains 7.9% variance and component 3 explains 6.6% variance.

Items loading with factor loading >0.5 onto component 1 are: *coat on / off, in / out bed, get dressed, turn over in bed, dry fully, shoes and socks, in / out chair, comb hair, in / out car*

Items loading with factor loading >0.5 onto component 2 are: *get up after bending down, bend down, walk on the flat, walk up or down stairs.*

Items loading with factor loading >0.5 onto component 3 are: *grip objects, reach above head, carry / lift things.*

Those loading with factor loadings of >0.4 onto more than one factor are: *shoes / socks, in / out chair, comb hair, in / out car.* These 4 items were therefore deleted.

A second PCA was conducted on the remaining 12 items with varimax rotation. The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis, KMO = 0.887.

The correlation matrix showed there were no correlations <0.3 or >0.9. The diagonal elements of the anti-image correlation matrix were all >0.5.

An initial analysis was run to obtain eigenvalues for each factor in the data. 2 factors had eigenvalues over Kaiser's criterion of 1 and in combination explained 65.3% of the variance (see Table 8.4).

The scree plot (Figure 8.9) shows a point of inflexion at component 3 suggesting 2 factors should be retained.

I therefore decided to retain 2 factors. The factor loadings after rotation are shown in Table 8.5.

Table 8.4: Eigenvalues associated with each factor before extraction, after extraction and after rotation (PCA 2 of functional items at diagnosis)

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.649	55.407	55.407	6.649	55.407	55.407	4.628	38.570	38.570
2	1.182	9.847	65.254	1.182	9.847	65.254	3.202	26.683	65.254
3	.975	8.123	73.377						
4	.673	5.609	78.986						
5	.658	5.482	84.467						
6	.420	3.498	87.965						
7	.403	3.358	91.324						
8	.306	2.552	93.876						
9	.247	2.062	95.938						
10	.196	1.637	97.575						
11	.173	1.442	99.017						
12	.118	.983	100.000						

Extraction Method: Principal Component Analysis.

Figure 8.9: Scree plot for PCA 2 of functional items at diagnosis

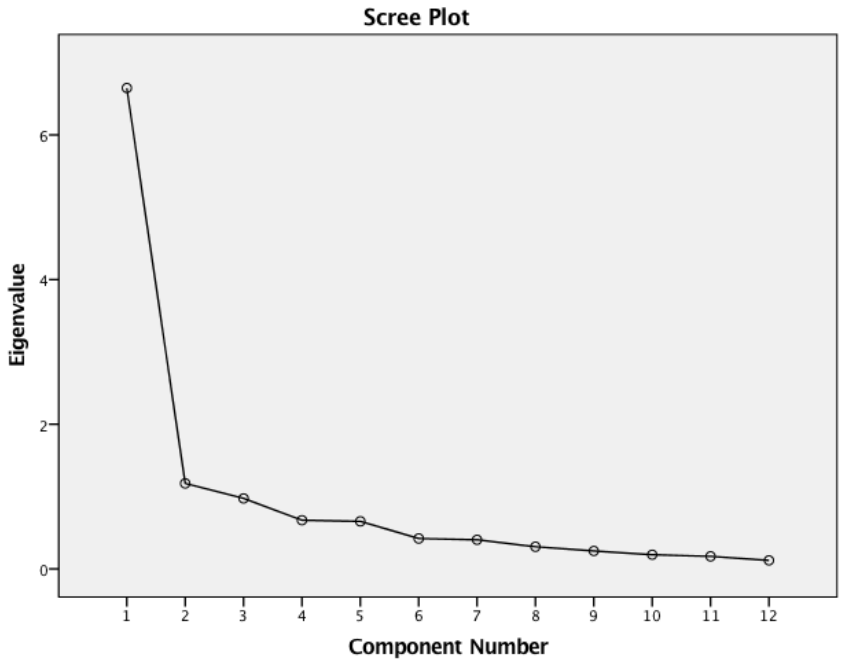


Table 8.5: Factor loadings after rotation (PCA 2 of functional items at diagnosis)

Rotated Component Matrix[a]		
	Component	
	1	2
Get dressed	.799	
Turn over in bed	.793	
Take your coat on or off	.791	
Dry yourself fully after shower or bath	.785	.317
Get in or out of bed	.770	
Reach above your head for things	.698	
Carry or lift things	.623	.432
Grip objects	.537	.383
Get up after bending down		.865
Bend down		.856
Walk on the flat		.719
Walk up or down stairs	.380	.648

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.[a]
a. Rotation converged in 3 iterations.

These tables show that component 1 explains 55.4% variance and component 2 explains 98% variance.

Items loading with factor loading >0.5 onto component 1 are: *get dressed, turn over in bed, coat on / off, in / out bed, dry fully, grip objects, reach above head, carry / lift things.*

Items loading with factor loading >0.5 onto component 2 are: *get up after bending down, bend down, walk on the flat, walk up or down stairs.*

Carry or lift things loads with factor loadings of >0.4 onto both components. This item was therefore deleted.

A third PCA was conducted on the remaining 11 items with varimax rotation. The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis, KMO = 0.87.

The correlation matrix showed there were no correlations <0.3 or >0.9. The diagonal elements of the anti-image correlation matrix were all >0.5.

An initial analysis was run to obtain eigenvalues for each factor in the data. 2 factors had eigenvalues over Kaiser's criterion of 1 and in combination explained 66.3% of the variance (see Table 8.6).

The scree plot (Figure 8.10) shows a point of inflexion at component 3 suggesting 2 factors should be retained.

I therefore decided to retain 2 factors. The factor loadings after rotation are shown in Table 8.7.

Table 8.6: Eigenvalues associated with each factor before extraction, after extraction and after rotation (PCA 3 of functional items at diagnosis)

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.117	55.605	55.605	6.117	55.605	55.605	4.236	38.509	38.509
2	1.181	10.737	66.342	1.181	10.737	66.342	3.062	27.832	66.342
3	.840	7.636	73.978						
4	.663	6.025	80.003						
5	.633	5.754	85.757						
6	.419	3.807	89.564						
7	.400	3.638	93.202						
8	.253	2.302	95.503						
9	.203	1.847	97.351						
10	.173	1.576	98.927						
11	.118	1.073	100.000						

Extraction Method: Principal Component Analysis.

Figure 8.10: Scree plot for PCA 3 of functional items at diagnosis

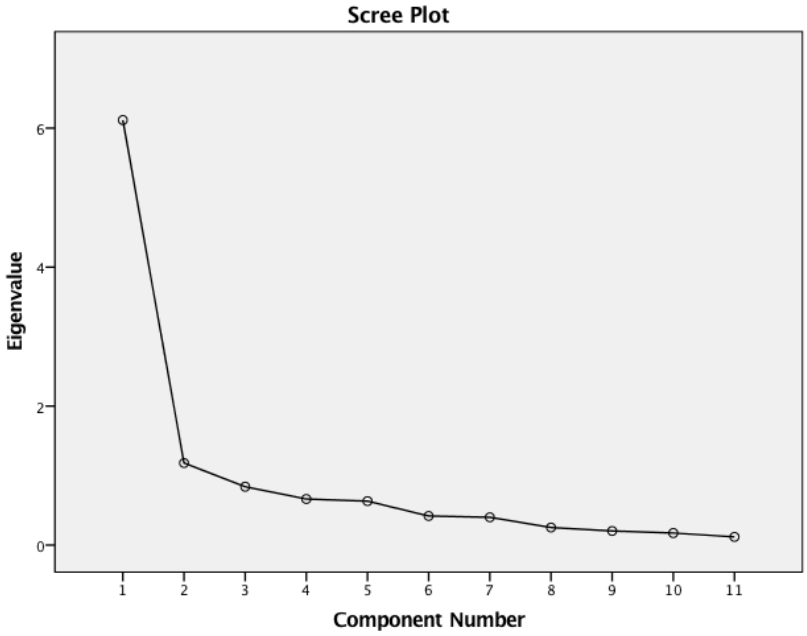


Table 8.7: Factor loadings after rotation (PCA 3 of functional items at diagnosis)

Rotated Component Matrix[a]		
	Component	
	1	2
Get dressed	.809	
Take your coat on or off	.804	
Turn over in bed	.800	
Dry yourself fully after shower or bath	.790	.326
Get in or out of bed	.782	
Reach above your head for things	.668	
Grip objects	.510	.389
Get up after bending down		.867
Bend down		.860
Walk on the flat		.722
Walk up or down stairs	.374	.651

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.[a]
a. Rotation converged in 3 iterations.

This analysis results in component 1 explaining 55.6% variance and component 2 explaining 10.7% variance.

Items loading with factor loading >0.5 onto component 1 are: *get dressed, coat on / off, turn over in bed, dry yourself fully, in / out bed, reach above head, grip objects.*

Items loading with factor loading >0.5 onto component 2 are: *get up after bending down, bend down, walk on the flat, walk up or down stairs.*

There are now two distinct components with no variables loading with >0.4 onto both. However, the separation into these components is not clinically meaningful.

Principal component analysis of functional items now

The same process was then applied to the functional item responses from the 'now' dataset.

A PCA was conducted on the 16 items with varimax rotation. The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis, KMO = 0.95.

The correlation matrix showed there were no correlations <0.3 or >0.9. The diagonal elements of the anti-image correlation matrix were all >0.5.

An initial analysis was run to obtain eigenvalues for each factor in the data. Only one factor had an eigenvalue over Kaiser's criterion of 1 and this accounted for 63.0% of the variance (see Table 8.8).

The scree plot (Figure 8.11) shows a point of inflexion at component 2 confirming that one factor should be retained.

As only one component was retained, no rotation was required. All variables loaded onto this one component with factor loadings of >0.5 (see Table 8.9).

Table 8.8: Eigenvalues associated with each factor before extraction, after extraction and after rotation (PCA of functional items now)

Total Variance Explained						
	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	10.081	63.005	63.005	10.081	63.005	63.005
2	.937	5.855	68.860			
3	.862	5.388	74.248			
4	.569	3.554	77.802			
5	.541	3.381	81.183			
6	.467	2.919	84.102			
7	.404	2.528	86.630			
8	.359	2.245	88.874			
9	.347	2.167	91.042			
10	.299	1.870	92.912			
11	.265	1.655	94.567			
12	.247	1.542	96.109			
13	.185	1.154	97.262			
14	.173	1.081	98.343			
15	.141	.879	99.222			
16	.124	.778	100.000			

Extraction Method: Principal Component Analysis

Figure 8.11: Scree plot for PCA of functional items now

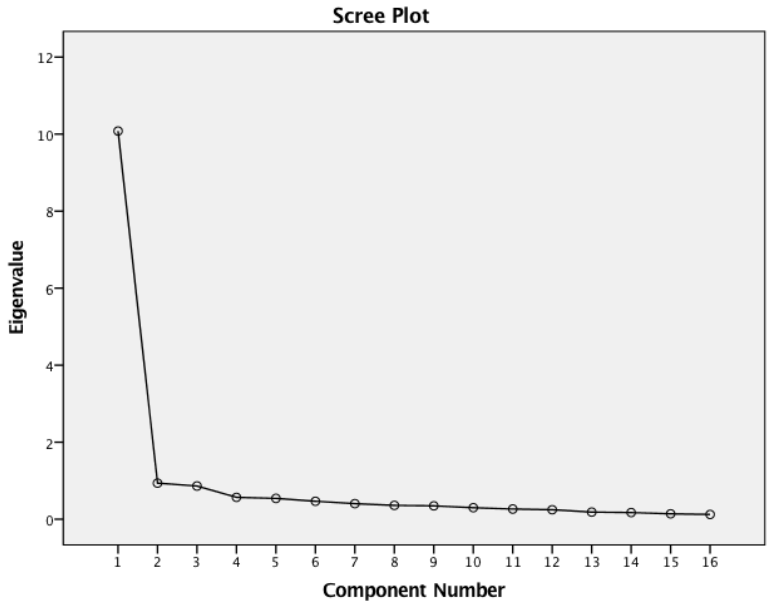


Table 8.9: Component matrix (PCA of functional items now)

Component Matrix[a]	
	Component 1
Get dressed	.851
Put on or take off your socks and shoes	.847
Take your coat on or off	.846
Get in or out of bed	.833
Bend down	.833
Get in or out of a car	.822
Turn over in bed	.816
Get up after bending down	.811
Dry yourself fully after shower or bath	.809
Get in or out of a chair	.808
Reach above your head for things	.795
Carry or lift things	.759
Walk up or down stairs	.740
Grip objects	.717
Walk on the flat	.709
Comb or blow dry your hair	.672
Extraction Method: Principal Component Analysis. a. 1 components extracted.	

Internal consistency reliability testing of functional items now

As the functional item scale has been shown to be unidimensional in the 'now' dataset, its internal consistency can be assessed using Cronbach's alpha. This was calculated in SPSS IBM statistics (IBM, 2014) and was found to be 0.96, indicating high internal consistency but suggesting that there might be some redundancy of items.

8.5.8 Analysis of the psychological and emotional well-being scale using Classical Test Theory

The same PCA method was applied to this scale.

Principle component analysis of emotional and psychological well-being items at diagnosis

A PCA was conducted on the 11 items with varimax rotation. The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis, KMO = 0.93.

The correlation matrix showed there were no correlations <0.3 or >0.9. The diagonal elements of the anti-image correlation matrix were all >0.5.

An initial analysis was run to obtain eigenvalues for each factor in the data. Only one factor had an eigenvalue over Kaiser's criterion of 1 and this accounted for 65.6% of the variance (see Table 8.10).

The scree plot (Figure 8.12) shows a point of inflexion at component 2 confirming that one factor should be retained.

As only one component was retained, no rotation was required. All variables loaded onto this one component with factor loadings of >0.5 (see Table 8.11).

Table 8.10: Eigenvalues associated with each factor before extraction, after extraction and after rotation (PCA of psychological items at diagnosis)

Total Variance Explained						
	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.221	65.643	65.643	7.221	65.643	65.643
2	.890	8.088	73.731			
3	.728	6.617	80.348			
4	.473	4.300	84.648			
5	.378	3.437	88.085			
6	.315	2.867	90.951			
7	.285	2.591	93.543			
8	.241	2.195	95.737			
9	.187	1.701	97.439			
10	.162	1.475	98.913			
11	.120	1.087	100.000			
Extraction Method: Principal Component Analysis.						

Figure 8.12: Scree plot for PCA of psychological items at diagnosis

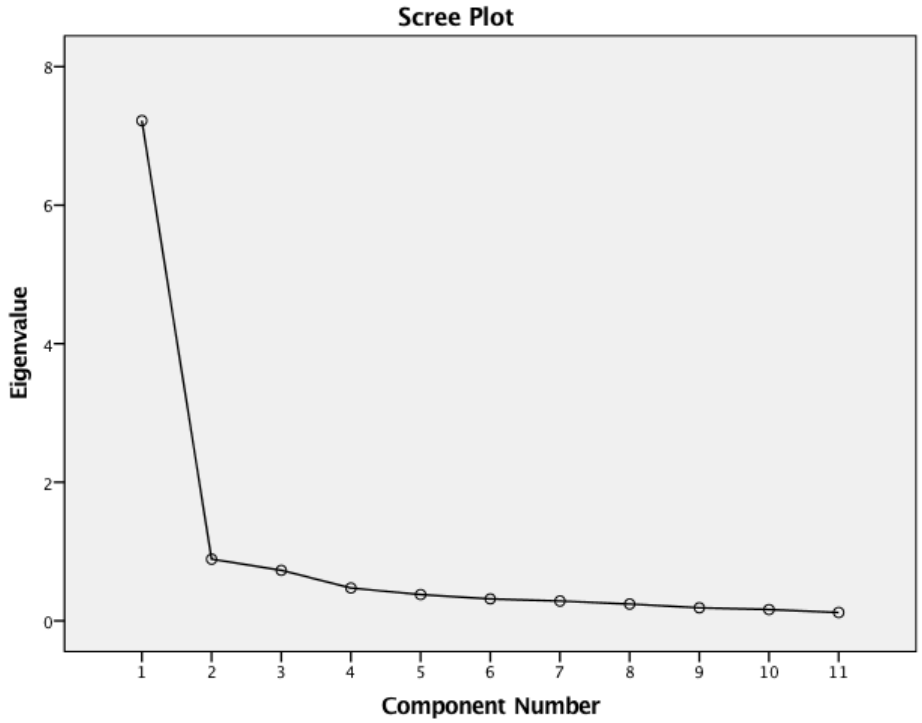


Table 8.11: Component matrix (PCA of psychological items at diagnosis)

Component Matrix[a]	
	Component 1
Low self confidence	.896
Vulnerable	.867
Anxious	.859
Low in mood	.845
Worry about the future	.839
Less inclined to go out	.829
Feel you need help looking after yourself	.809
Less interested in hobbies	.804
Worry about falling	.753
Affected sleep	.738
Feel more tired	.643

Extraction Method: Principal Component Analysis.
a. 1 components extracted.

Internal consistency reliability testing of emotional and psychological well-being at diagnosis scale

Crohnbach's alpha of this scale was 0.95 indicating high internal consistency but suggesting that there might be some redundancy of items.

There were no items for which deletion would improve alpha.

Principle component analysis of emotional and psychological well-being items now

A PCA was conducted on the 11 items with varimax rotation. The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis, KMO = 0.94.

The correlation matrix showed there were no correlations <0.3 or >0.9. The diagonal elements of the anti-image correlation matrix were all >0.5.

An initial analysis was run to obtain eigenvalues for each factor in the data. Only one factor had an eigenvalue over Kaiser's criterion of 1 and this accounted for 68.8% of the variance (see Table 8.12).

The scree plot (Figure 8.13) shows a point of inflexion at component 2 confirming that one factor should be retained.

As only one component was retained, no rotation was required. All variables loaded onto this one component with factor loadings of >0.5 (see Table 8.13).

Table 8.12: Eigenvalues associated with each factor before extraction, after extraction and after rotation (PCA of psychological items now)

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.569	68.811	68.811	7.569	68.811	68.811
2	.763	6.939	75.750			
3	.644	5.856	81.606			
4	.411	3.734	85.340			
5	.389	3.538	88.878			
6	.282	2.563	91.441			
7	.236	2.145	93.586			
8	.225	2.048	95.634			
9	.192	1.748	97.383			
10	.146	1.330	98.713			
11	.142	1.287	100.000			

Extraction Method: Principal Component Analysis.

Figure 8.13: Scree plot for PCA of psychological items now

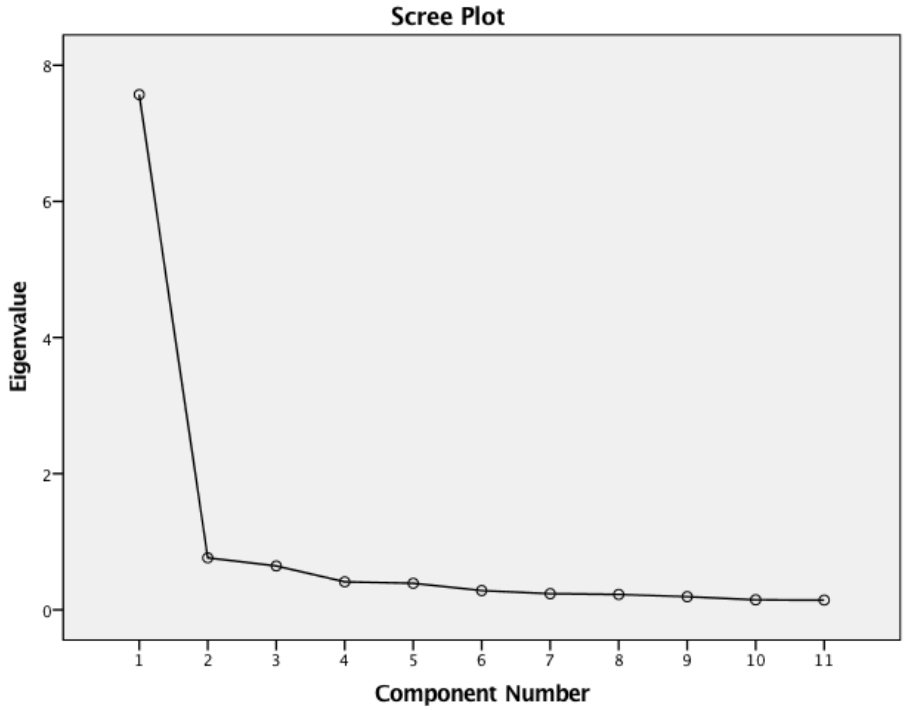


Table 8.13: Component matrix (PCA of psychological items now)

Component Matrix[a]	Component 1
Low self confidence	.885
Vulnerable	.882
Anxious	.879
Low in mood	.848
Less interested in hobbies	.840
Less inclined to go out	.833
Worry about the future	.830
Feel you need help looking after yourself	.811
Worry about falling	.774
Affected sleep	.768
Feel more tired	.762

Extraction Method: Principal Component Analysis.
a. 1 components extracted.

Internal consistency reliability testing of emotional and psychological well-being now scale

Crohnbach's alpha of this scale was 0.95 indicating high internal consistency but suggesting that there might be some redundancy of items.

There were no items for which deletion would improve alpha.

8.5.9 Summary of results of applying Classical Test Theory to the two scales

Using PCA has demonstrated that the scales for functional items 'now' and for emotional and psychological well-being both 'at diagnosis' and 'now', are unidimensional. The functional item scale 'at diagnosis' however, is still composed of two factors for which the separation is not clinically meaningful.

The unidimensional scales have high internal consistency. Although there may still be some redundancy, there is nothing in this analysis to further aid decisions on item reduction.

To aid further item reduction, through more detailed study of the performance of the items in measuring the construct I attempted to fit the full functional and emotional and psychological well-being scales to a Rasch measurement model.

8.5.10 Fitting the functional scale to a Rasch model

Data formatting

The data was formatted into a file that could be imported into RUMM2020 (Andrich et al., 2003).

This involved transforming 'age' and 'duration since diagnosis' into categorical data. Age was divided into <75 years and 75 years plus as 75 years was the median age of the respondents and resulted in similar numbers in the two groups, which is helpful for analysis purposes. Duration since diagnosis was divided into <18 months and 18 months plus as this was approximately the median duration.

The item response category 'not relevant' had to be removed as this did not fit into a hierarchy. As outlined in section 8.5.3, one of the early decisions was to delete items that had very high combined 'not relevant' or 'missing' responses so some of the most problematic items in this regard had already been deleted. I decided to assign the remaining 'not relevant' responses as 'missing' for the purposes of further analysis.

For this analysis, those items where >20% of people selected 'not limited' in the at diagnosis data set were left in as after discussion with my supervisors, I felt that items where up to 80% of people were affected were important to include at this stage.

In addition to assessing the fit of the 'at diagnosis' and 'now' datasets to a Rasch model, I created a combined dataset with an added person factor of 'time'. This enabled me to assess whether any items behaved significantly differently at diagnosis compared to later in the disease course by testing for DIF by time.

Results of the iterative process

I attempted to fit the data to a Rasch model, iteratively deleting items to improve the fit. In each case the likelihood ratio test did not support the use of a rating scale model so a partial credit analysis model was used.

Class interval structure was assessed for each analysis and in each case, four class intervals were formed with approximately 50 individuals in each.

There were no disordered thresholds at any iteration.

The stepwise process is presented below (Table 8.14) with summary tables of the results following (Table 8.15 and Table 8.16).

Table 8.14: Iterative process of fitting the functional scale to the Rasch model

Dataset	Findings	Resultant action
Step 1		
At diagnosis	Not unidimensional and does not fit the model. Two items (sitting for more than 30 minutes at a time and gripping objects) had significant Chi squared fit statistics (suggesting poor fit) and high positive fit residuals (suggesting under-discrimination)	Sitting for more than 30 minutes and gripping objects were deleted in the at diagnosis dataset and it was tested again.
Now	Not unidimensional and does not fit the model. Three items (wiping after going to the toilet, sitting for more than 30 minutes, grip objects) had significant Chi squared fit statistics and sitting and gripping also had high positive fit residuals.	
Combined	Tested for DIF by time. Three items show significant uniform DIF for time – getting in and out of bed, getting in and out of a chair and gripping objects.	
Step 2		
At diagnosis	The reduced scale was not unidimensional and not a good fit. The Chi square fit statistic was significant for three items (walk on flat, in / out chair, dress yourself fully) and walk on the flat had a high positive fit residual.	Three further items (get in or out of a chair, walk on the flat, dress yourself fully) were deleted.

Step 3		
At diagnosis	This scale was still not unidimensional though was statistically a better fit to the Rasch model. No items had significant Chi squared fit statistics or high positive or negative fit residuals but studying the item characteristic curves identified 'take your coat on or off' and 'get up after bending' to be visually the least well fitting.	Two further items (take your coat on or off and get up after bending) were deleted.
Step 4		
At diagnosis	The reduced scale was close to unidimensional with 5.8% of paired t-tests significant at the 5% level and was a good fit to the model. No items had significant Chi squared fit statistics but 'bend down' had a high positive fit residual. Studying the item characteristic curves identified 'bend down' and 'dry yourself fully' to be visually the least well fitting.	Two further items (bend down and dry yourself fully) were deleted.
Step 5		
At diagnosis	The resultant eleven-item scale was unidimensional and a good fit to the Rasch model.	Two further items (comb / blow dry your hair and wipe yourself after going to the toilet) were deleted.
Now	This same eleven-item scale was then created in the 'now' dataset and tested against the model. In this case it was not unidimensional (5.5% of paired t-tests were significant at the 5% level) and it was not a good fit to the model (the item trait interaction statistic was significant). The Chi square fit statistic was not significant for any item and no item had a high negative or positive fit residual but studying the	

	item characteristic curves, showed the least well-fitting items visually were comb or blow dry your hair and wipe after going to the toilet.	
Step 6		
Now	This nine-item scale was unidimensional and a reasonable fit to the Rasch model. The item trait interaction statistic was just significant at 0.04 but by all other measures, the data was a good fit. Studying the individual item fit did not show any misfitting items.	This was retained as the final scale.
At diagnosis	The same nine-item scale was tested in this dataset and also found to be unidimensional and fit the model well.	
Combined	The nine-item scale was tested for DIF by time. Two items showed uniform DIF – get in or out of bed and turn over in bed.	

Table 8.15: Summary of results for (at diagnosis) functional scale fit to a Rasch model

Scale iteration	Number of items	Targeting	Unidimensionality		Item fit residuals		Person fit residuals		Item trait interaction statistic	Power of test of fit
			% significant t-tests	Unidimensional?	Mean	SD	Mean	SD		
1	20	1.31	13.6	No	-0.26	1.52	-0.28	1.35	Significant	Excellent
2	18	1.45	11.3	No	-0.24	1.38	-0.33	1.41	Significant	Excellent
3	15	1.45	9.6	No	-0.18	1.19	-0.28	1.26	Not significant	Excellent
4	13	1.40	5.8	No	-0.15	1.37	-0.27	1.17	Not significant	Excellent
5	11	1.41	4.0	Yes	-0.12	1.24	-0.29	1.13	Not significant	Excellent
6	9	1.56	2.6	Yes	-0.14	0.68	-0.31	1.12	Not significant	Excellent

Table 8.16: Summary of results for (now) functional scale fit to a Rasch model

Scale iteration	Number of items	Targeting	Unidimensionality		Item fit residuals		Person fit residuals		Item trait interaction statistic	Power of test of fit
			% significant t-tests	Unidimensional?	Mean	SD	Mean	SD		
1	20	-1.44	12.0	No	-0.37	1.76	-0.36	1.37	Significant	Excellent
2	11	-1.56	5.5	No	-0.50	0.93	-041	1.14	Significant	Excellent
3	9	-1.15	3.5	Yes	-0.43	1.06	-0.38	1.07	Significant (0.04)	Excellent

NB. The highlighted rows show the scale iterations where a good fit was achieved

In addition to the considerations detailed in the tables, local dependency, individual person fit and DIF were considered at each stage.

Local dependency was assessed by studying the residual correlations between items. In the earlier iterations there were a few residual correlations of >0.3 but in the final nine-item scale, there was only one residual correlation of this magnitude in each dataset (get in or out of bed with turn over in bed).

Individual person fit was considered by looking at numbers of 'extreme' individuals (those scoring maximum or minimum scores) and people with high (>2.5) or low (<-2.5) fit residuals. In each case there were individuals who responded in an unexpected way (e.g. scoring 'not limited at all' to every question in the early stages of the disease) but at this stage in the development process I decided not to remove these people and to focus on item fit instead to keep the scale as generalizable as possible (Hendriks et al., 2012).

DIF by age and gender was considered in the 'at diagnosis' dataset and by age, gender and duration since diagnosis in the 'now' dataset. No items showed DIF for age or duration since diagnosis. 'Comb or blow dry your hair' did show DIF for gender but this item was later deleted. In the final nine-item scale, 'take your shoes or socks on or off' showed uniform DIF for gender in the 'now' dataset but not 'at diagnosis'.

The person-item threshold distribution was also studied to look at the distribution of items and people. Graphs representing this for the final nine-item scale are shown in Figure 8.14, Figure 8.15, Figure 8.16 and Figure 8.17. These show that even with the omission of participants with the most extreme responses, the items do not cover the full breadth of disability and therefore there is likely to be floor and ceiling effects. This is more apparent in the 'at diagnosis' data than in the 'now' data.

Figure 8.14: Person-item threshold distribution for (at diagnosis) functional scale for all individuals

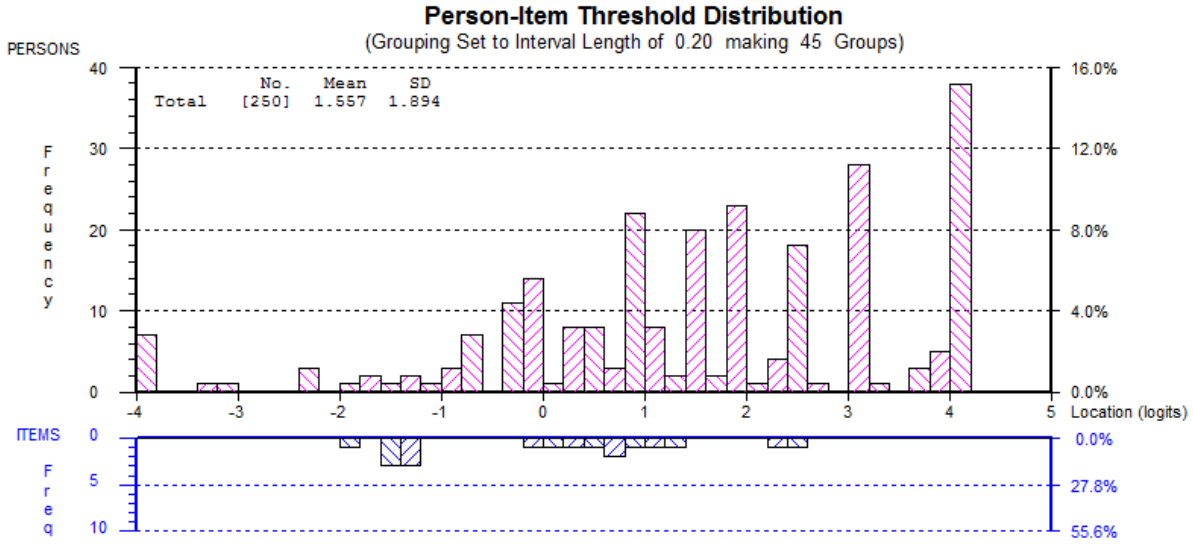


Figure 8.15: Person-item threshold distribution for (at diagnosis) functional scale with extreme individuals omitted

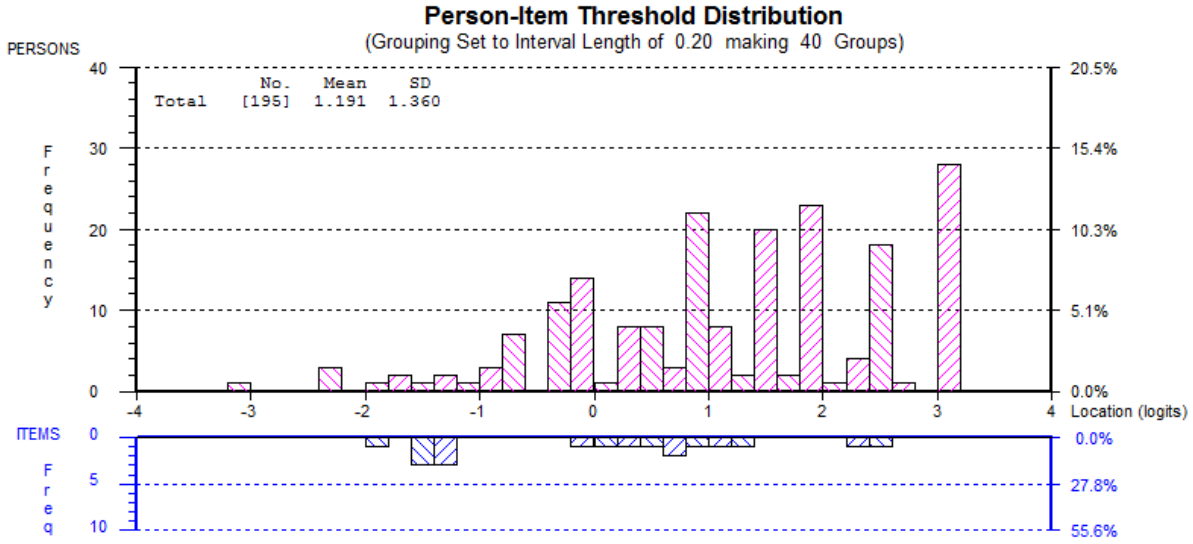


Figure 8.16: Person-item threshold distribution for (now) functional scale for all individuals

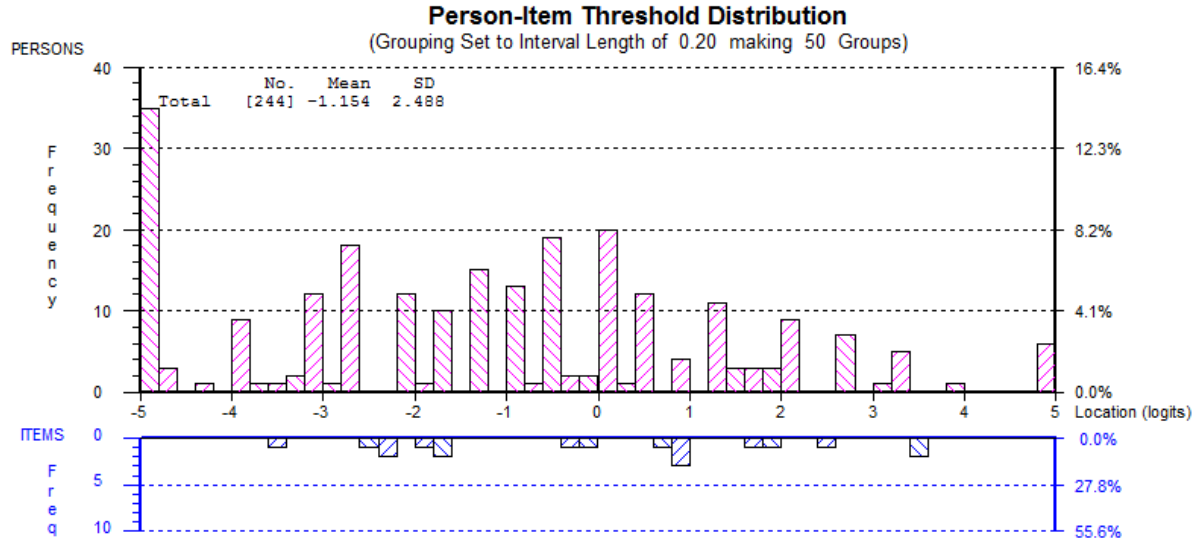
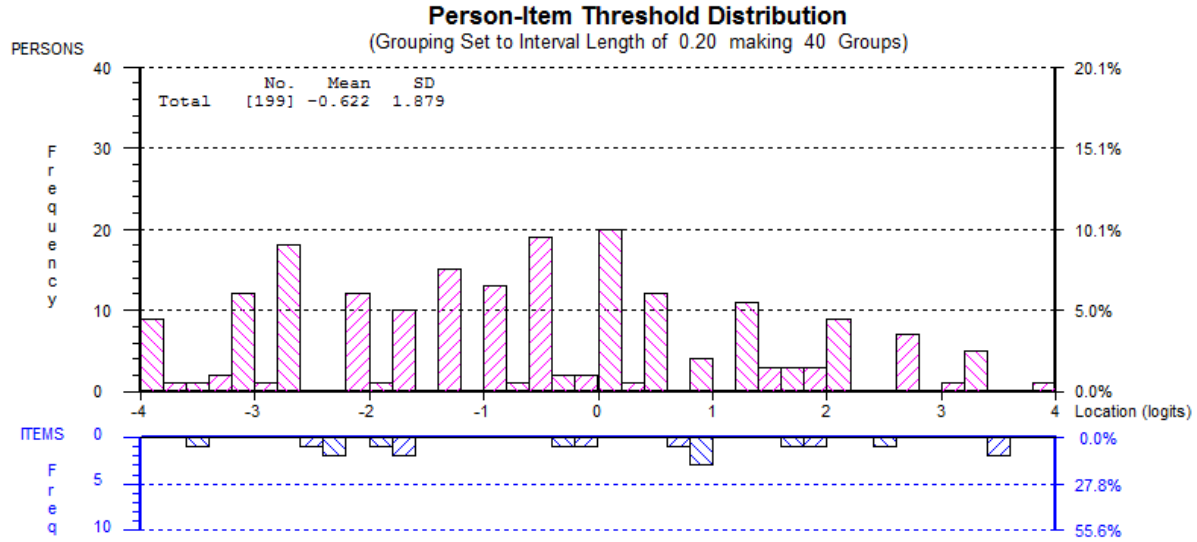


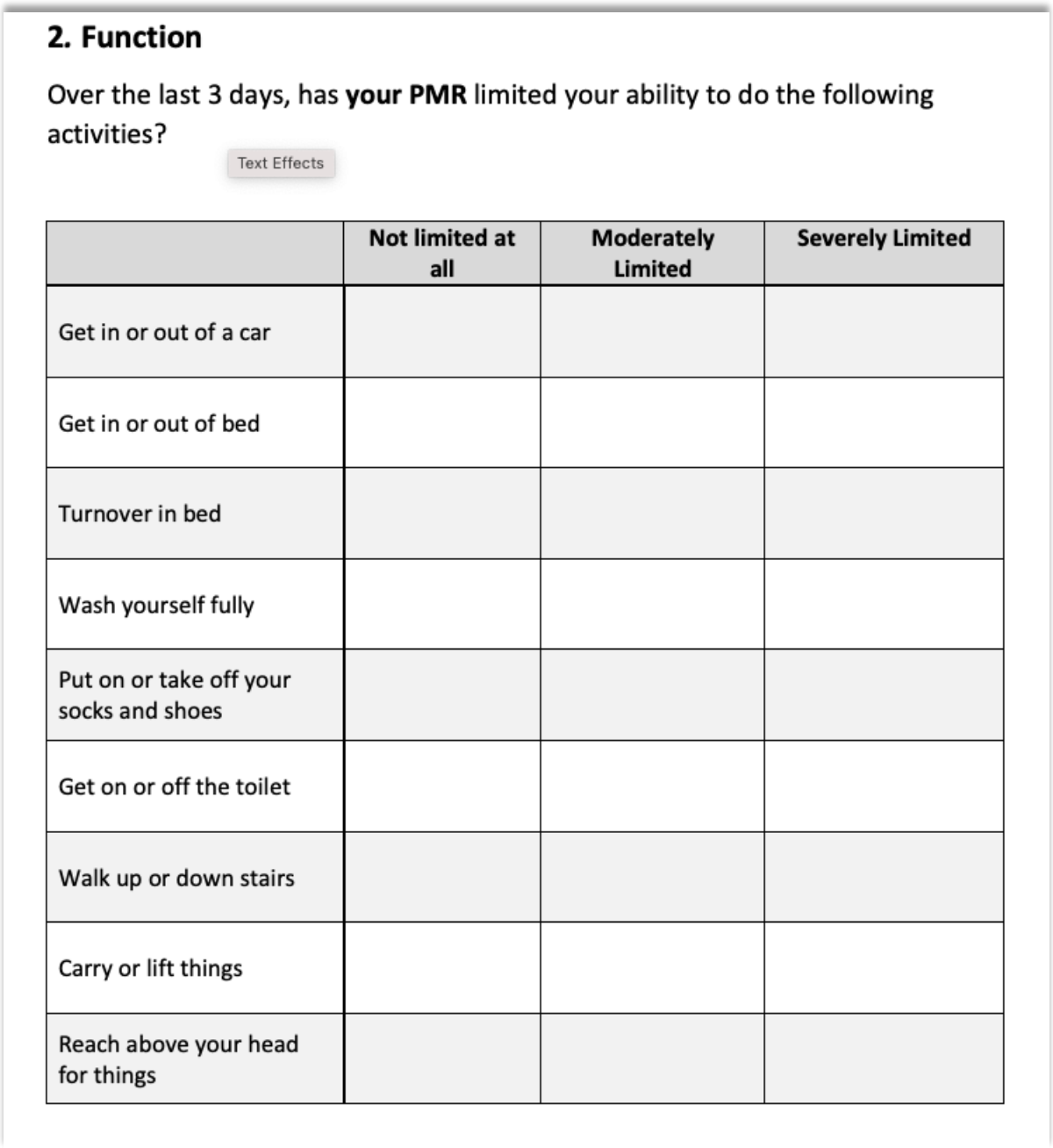
Figure 8.17: Person-item threshold distribution for (now) functional scale with extreme individuals omitted



Final functional scale

A nine-item functional scale has been created that is unidimensional and is a reasonable fit to a Rasch model in both datasets (Figure 8.18).

Figure 8.18: Final functional scale



2. Function

Over the last 3 days, has **your PMR** limited your ability to do the following activities?

Text Effects

	Not limited at all	Moderately Limited	Severely Limited
Get in or out of a car			
Get in or out of bed			
Turnover in bed			
Wash yourself fully			
Put on or take off your socks and shoes			
Get on or off the toilet			
Walk up or down stairs			
Carry or lift things			
Reach above your head for things			

8.5.11 Fitting the emotional and psychological well-being scale to a Rasch model

Data formatting

Data was formatted in the same way as for the functional scale but without the complication of there being a 'not relevant' response category. The data was then imported into RUMM2020 (Andrich et al., 2003).

Results of the iterative process

I attempted to fit the data to a Rasch model, iteratively deleting items to improve the fit. In the first iteration, testing the full scale in the at diagnosis dataset, the likelihood ratio test was significant, so a rating scale model was used. In each subsequent iteration however, the likelihood ratio test did not support the use of a rating scale model so a partial credit model was used.

Class interval structure was assessed at the start of each analysis and in each case, four class intervals were formed.

There were no disordered thresholds in any iteration though some were close as described in the detailed results below.

The stepwise process is presented in Table 8.17 with summary tables of the results following (Table 8.18 and Table 8.19).

Table 8.17: Iterative process of fitting the emotional and psychological well-being scale to the Rasch model

Dataset	Findings	Resultant action
Step 1		
At diagnosis	Not unidimensional and does not fit the model. One item (feeling more tired) had a significant Chi squared fit statistic (suggesting poor fit). Three items had high negative fit residuals (anxious, low mood, low self-confidence) suggesting redundancy. Studying the item characteristic curves showed 'feeling more tired' and 'worrying about the future' were visually the least well fitting.	As the scale was not unidimensional in either dataset, I decided to delete 'more tired' (poor fit to the model in both datasets), 'worry about the future' (poor fit in the at diagnosis dataset) and 'worry about falling' (poor fit in the now dataset and shows DIF for age and time).
Now	Not unidimensional and does not fit the model. The thresholds were ordered but the category probability curves, showed that the item 'less inclined to go out' had a very small section where response option one was the most likely. No items had significant Chi square fit statistics but studying the item characteristic curves showed 'more tired' and 'worry about the future' to be the least well-fitting. No items had high negative fit residuals but 'feeling more tired' had a high positive fit residual.	
Combined	Tested for DIF by time. Two items show significant uniform DIF – 'low mood' and 'worry about falling'.	

Step 2		
At diagnosis	The reduced scale was not unidimensional and not a good fit. The Chi square fit statistic was significant for 'affected sleep' and three items had high negative fit residuals (anxious, mood and low self-confidence).	One further item (affected sleep) was deleted.
Step 3		
At diagnosis	This 7-item scale was still not unidimensional though was a better fit to the Rasch model by all other parameters.	The analyses so far show that this scale is not unidimensional and therefore cannot fit a Rasch model. The deletions made are not improving this. Studying the items loading onto different factors during the PCA carried out to test unidimensionality shows that at each stage, there are four items consistently loading onto one factor with the other items loading onto the second factor ('low in mood', 'anxious', 'vulnerable' and 'low self-confidence'). Conceptually, these can be separated from the other items which are more functional / participatory in nature. I therefore decided to test these as two separate scales.
Step 4		

At diagnosis	The 4-item scale was found to be unidimensional and a good fit to the Rasch model. No items had significant Chi square fit statistics or high or low fit residuals and all the item characteristic curves look reasonable.	This was retained as the final scale.
Now	The same 4-item scale was unidimensional and a good fit. No items had significant Chi square fit statistics or high or low fit residuals and all the item characteristic curves look reasonable	
Combined	The same 4-item scale was tested for DIF by time in the combined dataset and no items showed this.	
Step 5		
At diagnosis	I then tested a scale made up of the remaining seven items. This was shown to not be unidimensional and not to be a good overall fit to the model. Although the Chi square fit statistic was not significant for any items and none had high negative or positive fit residuals, the item characteristic curves showed that the least well-fitting items were 'feeling more tired' and 'less inclined to go out'.	I deleted 'more tired' and 'less inclined to go out' and re-ran the analysis.
Step 6		
At diagnosis	The resultant 5-item scale was unidimensional and a good fit to the model. There were no disordered thresholds but 'worry about	I deleted the least well-fitting item ('need more help looking after yourself') and ran the analysis again.

	falling' had a small window where response option one was the most likely. The Chi square fit statistic was not significant for any items and none had high negative or positive fit residuals.	
Now	This same 5-item scale was unidimensional but was not a good overall fit (the item trait interaction statistic was significant). In addition, although the thresholds were ordered, studying the category threshold plots showed that some of the categories were very small. The Chi square fit statistic was not significant for any items and none had high negative or positive fit residuals but the item characteristic curves showed the least well-fitting item was 'need more help looking after yourself'.	
Step 7		
Now	This second 4-item scale was unidimensional and an overall good fit to the Rasch model. The thresholds were ordered but not well spaced with some of the categories being very small. The Chi square fit statistic was not significant for any items and none had high negative or positive fit residuals	This scale was not retained.

Table 8.18: Summary of results for (at diagnosis) emotional and psychological well-being scale fit to a Rasch model

Scale iteration	Number of items	Targeting	Unidimensionality		Item fit residuals		Person fit residuals		Item trait interaction statistic	Power of test of fit
			% significant t-tests	Unidimensional?	Mean	SD	Mean	SD		
1	11	-0.24	9.7	No	-0.50	2.72	-0.59	1.67	Significant	Excellent
2	8	-0.4	10.7	No	-0.42	2.49	-0.56	1.48	Significant	Excellent
3	7	-0.54	12.6	No	-0.11	1.93	-0.6	1.51	Not significant	Excellent
4	4	-0.53	3.2	Yes	-0.59	1.11	-0.81	1.45	Not significant	Excellent
5	7	-0.19	7.1	No	-0.13	1.76	-0.49	1.35	Significant	Excellent
6	5	-0.28	3.0	Yes	-0.02	1.42	-0.44	1.16	Not significant	Excellent

Table 8.19: Summary of results for (now) emotional and psychological well-being scale fit to a Rasch model

Scale iteration	Number of items	Targeting	Unidimensionality		Item fit residuals		Person fit residuals		Item trait interaction statistic	Power of test of fit
			% significant t-tests	Unidimensional?	Mean	SD	Mean	SD		
1	11	-1.77	9.4	No	-0.12	1.78	-0.30	1.18	Significant	Excellent
2	4	-3.53	2.3	Yes	-0.35	1.17	-0.60	1.20	Not significant	Excellent
3	5	-1.44	3.0	Yes	-0.21	1.27	-0.30	0.91	Significant	Excellent
4	4	-1.08	0.0	Yes	-0.08	0.97	-0.42	1.16	Not significant	Good

NB. The highlighted rows show the scale iterations where a good fit was achieved

In addition to the factors already discussed, local dependency, individual person fit and DIF were considered at each stage.

Local dependency was assessed by studying the residual correlations between items. In the final four-item scale in the 'at diagnosis' dataset there were there were two residual correlations >0.3 (low mood with feeling vulnerable and anxiety with low self-confidence). In the same scale in the 'now' dataset there were three residual correlations >0.3 (low mood with vulnerability and with low self-confidence and anxiety with low self-confidence).

Individual person fit was considered by looking at numbers of 'extreme' individuals (those scoring maximum or minimum scores) and people with high (>2.5) or low (<-2.5) fit residuals. In each case there were individuals who responded in an unexpected way (e.g. scoring 'all of the time' to every question in the later stages of the disease) but at this stage in the development process I decided not to remove these people and to focus on item fit instead.

DIF by age and gender was considered in the 'at diagnosis' dataset and by age, gender and duration since diagnosis in the 'now' dataset. In the final four-item scale, no items showed DIF by any of these parameters. In the scale of the remaining items, 'worry about falling' and 'affected your sleep' showed uniform DIF for age.

The person-item threshold distribution was also studied to look at the distribution of items and people. Graphs representing this for the final four-item scale are shown in Figure 8.19, Figure 8.20, Figure 8.21 and Figure 8.22. As for the functional scale, these show that there is sparseness of items at either extreme of the range, with risk of floor and ceiling effects.

Figure 8.19: Person-item threshold distribution for (at diagnosis) emotional and psychological well-being scale for all individuals

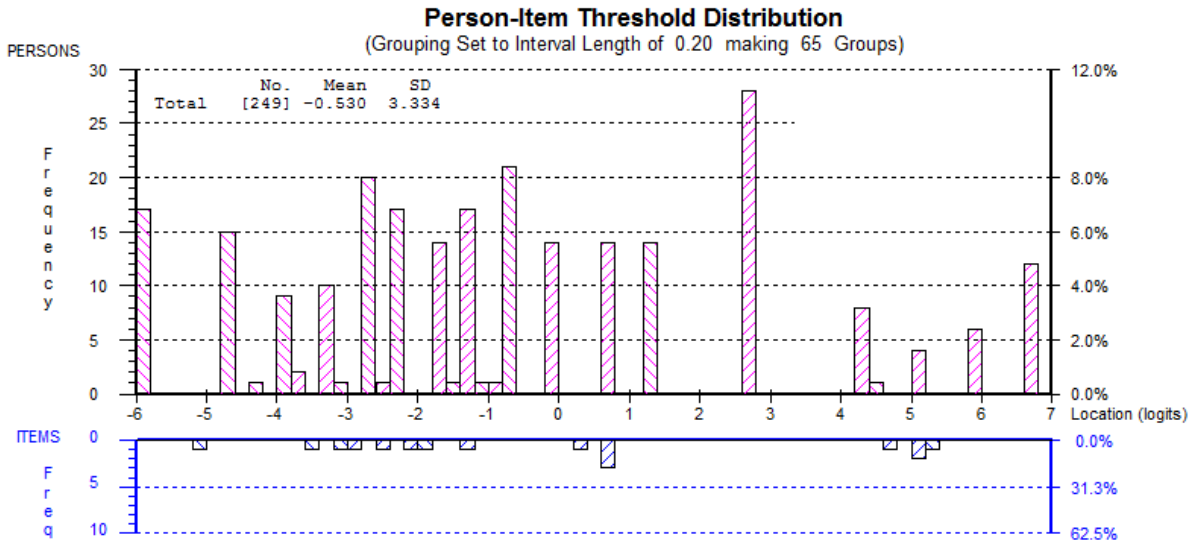


Figure 8.20: Person-item threshold distribution for (at diagnosis) emotional and psychological well-being scale with extreme individuals omitted

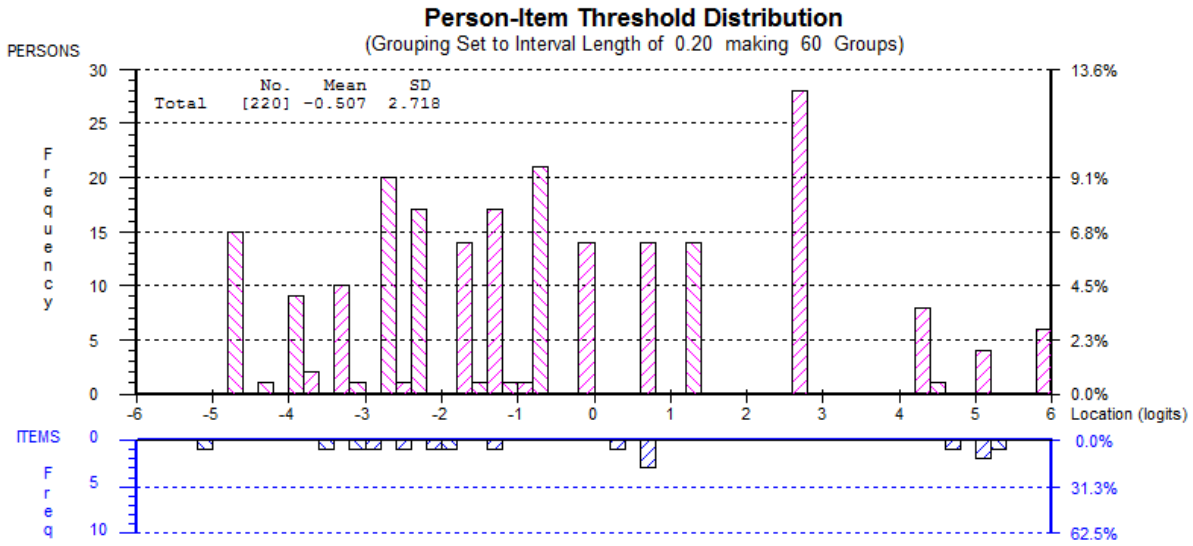


Figure 8.21: Person-item threshold distribution for (now) emotional and psychological well-being scale for all individuals

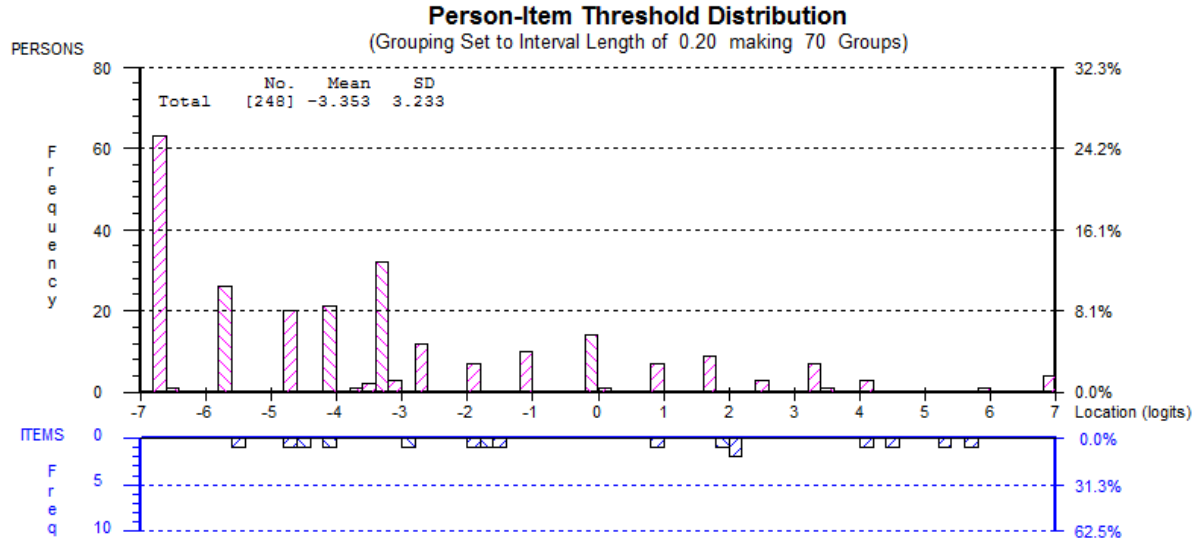
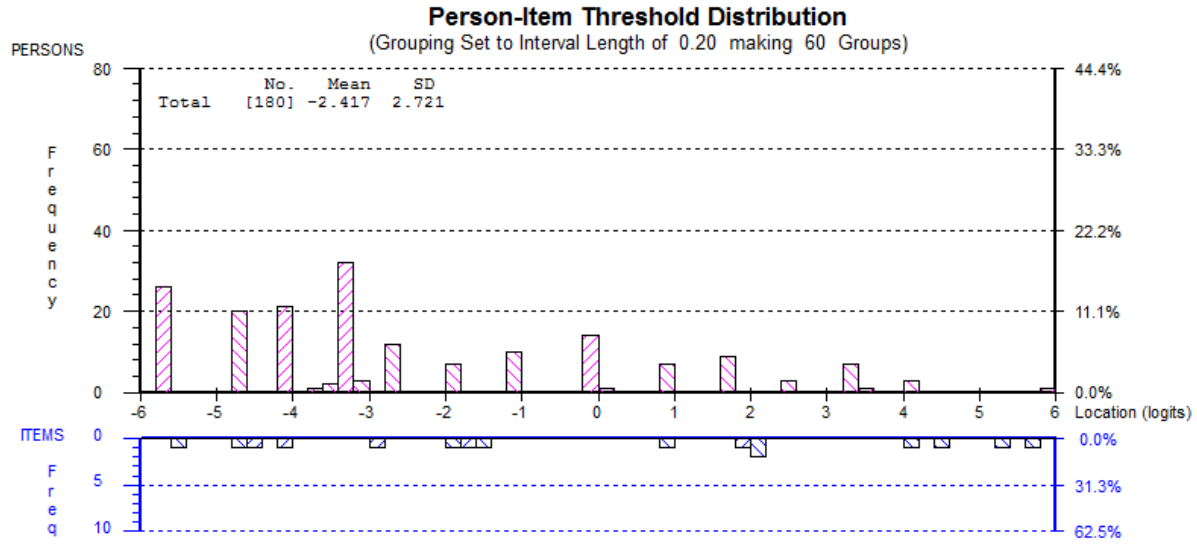


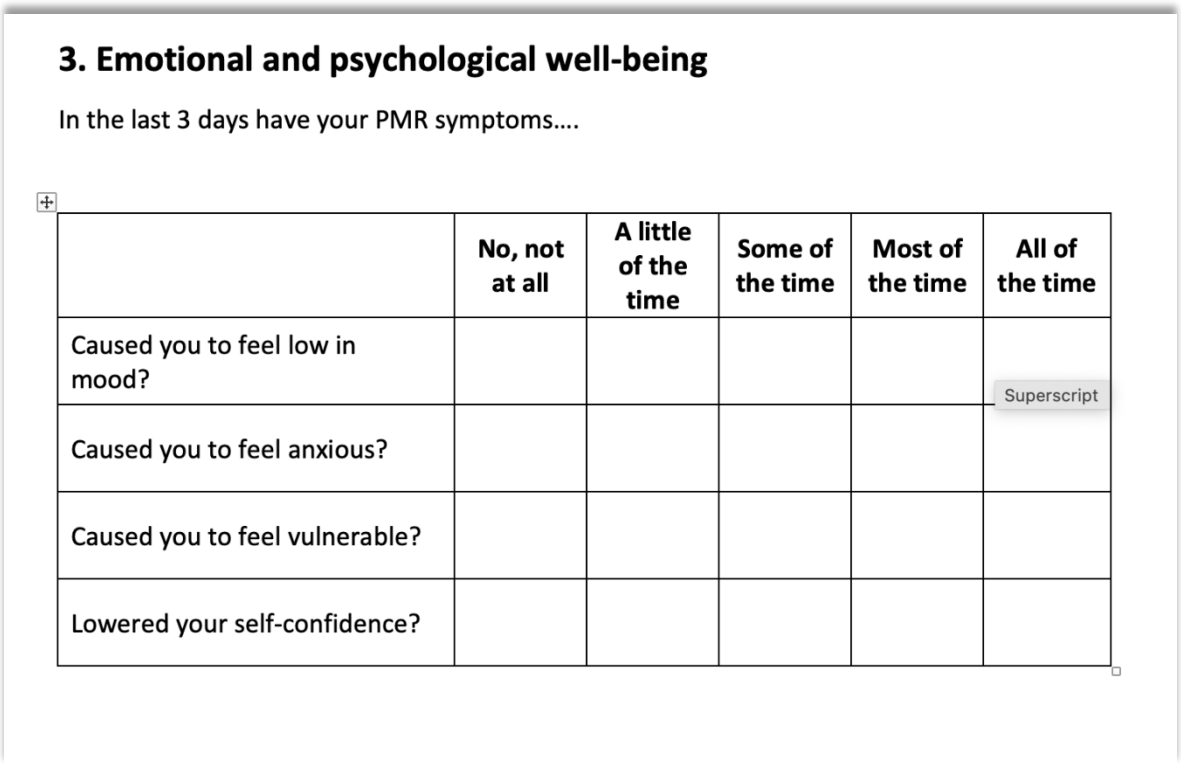
Figure 8.22: Person-item threshold distribution for (now) emotional and psychological well-being scale now with extreme individuals omitted



Final emotional and psychological well-being scale

A four-item scale has been created that is unidimensional and is a reasonable fit to a Rasch model in both datasets (Figure 8.23).

Figure 8.23: Final emotional and psychological well-being scale



3. Emotional and psychological well-being

In the last 3 days have your PMR symptoms....

	No, not at all	A little of the time	Some of the time	Most of the time	All of the time
Caused you to feel low in mood?					
Caused you to feel anxious?					
Caused you to feel vulnerable?					
Lowered your self-confidence?					

A second emotional and psychological well-being scale of five items is unidimensional and fits a Rasch model in the 'at diagnosis' data set. If this is reduced to four items (worry about falling, less interested in hobbies, worry about the future and affected sleep) it is a reasonable fit according to the statistical parameters but the threshold categories are not well spaced.

I therefore decided (after discussion with my supervisors) to retain only the first four-item scale.

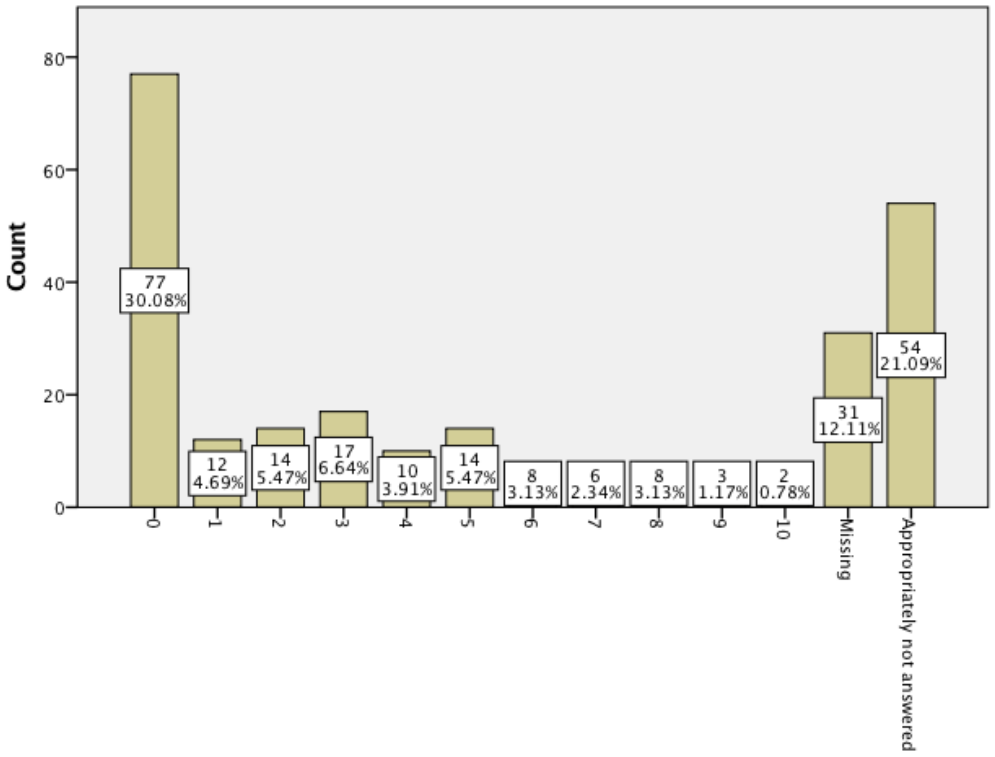
8.5.12 Results of the medication side effects section

74.6% of participants were still taking prednisolone. 21.1% were not and 4.3% did not answer this question.

Of those still taking prednisolone, the mean dose was 6.5mg (SD 5.1).

The third item in this section asked 'How much have you been affected by side effects of prednisolone in the last three days?' with response rated on a 0-10 scale. Results for this are shown in Figure 8.24. Those participants no longer taking prednisolone were asked to skip this question.

Figure 8.24: Impact of prednisolone side effects in the preceding 3 days, where 0= unaffected and 10 = severely affected

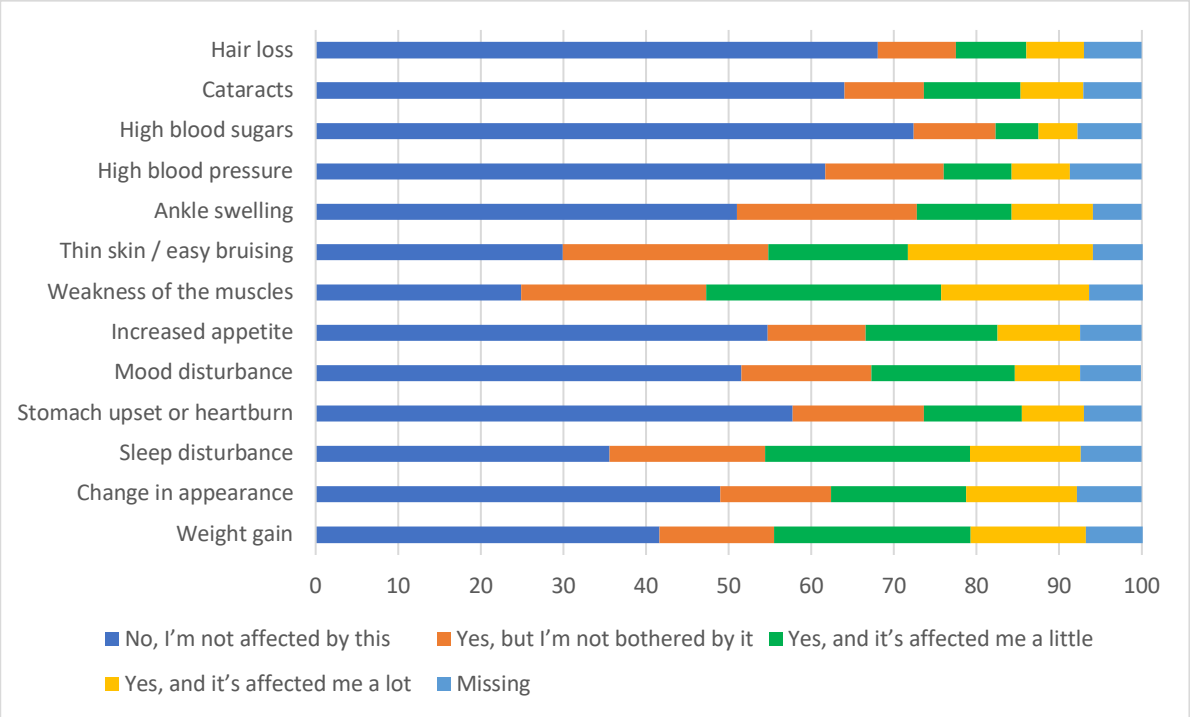


This question has a high rate (>10%) of missing responses and a heavy skew towards the lower end of the scale (though it is important to note that the mean dose of prednisolone that participants were taking was low). On reflection, this 'overall' question adds little to the information gained from asking about the impact of specific side effects (which is more useful to know clinically) and therefore this question will be removed from the PROM.

The distribution of responses to questions about specific side effects are shown in Figure 8.25. The most significant side effects were muscle weakness and thin skin or easy bruising with weight gain, change in appearance and sleep disturbance all also being reported as significantly affecting more than 10% of respondents.

There were fewer than 10% missing responses to all side effect items and all response categories were used.

Figure 8.25: Distribution of responses to side effects items



8.5.13 Results of the final item – the 'back to normal' question

The final item on the questionnaire asked whether respondents felt back to the level of health they were at before they first experienced PMR symptoms.

The distribution of responses is shown in Table 8.20.

Table 8.20: Distribution of response to 'back to normal' question

Response option	Percentage
Yes, completely	18.8
Yes, partially	46.5
No, not at all	31.6
Missing	3.1

8.6 Scoring of the PROM

Having developed the scales for each domain, the scoring system needs to be determined. The PMR-PROM has two domains based on a formative model (symptoms and steroid side effects) and two domains based on a reflective model, which have therefore been developed using measurement theory techniques (functional limitation and emotional and psychological well-being). The impact of the underlying model on the scoring systems for these domains is discussed below.

8.6.1 Symptoms domain

The symptoms section of the PROM contains two questions about each of four symptoms. The severity of each symptom is rated on a 0-10 scale and the duration is scored on a 0-4 scale. Simple sum scoring is therefore appropriate and it seems reasonable clinically to give each symptom equal weight and to give the duration scores equal weight to the severity scores.

Each duration score will therefore be multiplied by 2.5 to create a score out of 10 and the individual scores summed to give a total score. The minimum possible score is therefore 0 and the maximum possible score is 80.

To be consistent with the other domains, this will be converted to a percentage.

8.6.2 Functional and psychological impact domain

The functional and psychological impact scales have been developed using classical and modern test theory. As discussed in Section 8.3.7, Rasch models are probabilistic forms of Guttman scales; that is if a person can 'agree' to an item, there is a high probability that 'easier' items will also be agreed. The logit score equates to the difficulty of the item and a person's location on the same scale represents how much of the variable each respondent possesses. The precision with which an item's scale location has been estimated is represented by the item's standard error of measurement. Likewise, the precision of each individual respondent's estimated location is specified by the standard error of measurement of that person.

The main advantage of having a scale that fits a Rasch model is that it allows interval level measurement and a scoring system can be devised which takes account of the item

difficulties. In a simple sum scoring system if there are items grouped closely together in one part of the scale and someone improves over that range of the scale, they will show a larger improvement in score than if someone improves a similar amount over a range of the scale where there a fewer items. Using logit-based scoring eliminates this problem.

Item locations and their standard errors for the nine-item functional scale and the four-item emotional and psychological well-being scale are presented in Table 8.21 and Table 8.22.

These results show that the items have a different order of difficulty in the two datasets which creates a problem for devising a logit-based scoring system.

Table 8.21: Item locations and their standard errors on the functional scale

Item	At diagnosis			Now		
	Location	SE	Order	Location	SE	Order
Reach above your head	-0.508	0.143	1	-0.672	0.140	3
Shoes / socks on or off	-0.462	0.144	2	-0.534	0.140	5
Carry or lift things	-0.450	0.145	3	-1.433	0.150	1
In / out bed	-0.436	0.145	4	-0.371	0.153	7
Turn over in bed	-0.405	0.143	5	-0.027	0.146	6
Up or down stairs	-0.077	0.144	6	-0.783	0.148	2
In / out car	-0.046	0.140	7	-0.205	0.152	4
Wash fully	1.108	0.128	8	1.583	0.163	8
Get on or off the toilet	1.277	0.128	9	1.701	0.164	9

Table 8.22: Item locations and their standard errors on the functional scale

Item	At diagnosis			Now		
	Location	SE	Order	Location	SE	Order
Low in mood	-0.628	0.121	1	-0.233	0.140	2
Feeling anxious	0.056	0.115	2	-0.283	0.139	3
Low self-confidence	0.141	0.113	3	-0.347	0.136	1
Feeling vulnerable	0.430	0.112	4	0.298	0.139	4

There are two possible reasons for the different ordering. Firstly, it could be due to poor precision of the estimated locations of the items such that the standard errors are large and overlap. To test whether this was the case, a plot was drawn of item locations at each time point with standard errors shown (Figure 8.26 and Figure 8.27). These show that the precision of the estimated locations is reasonably good and unlikely to be the cause of the discrepancy in ordering between the two datasets.

Figure 8.26: Item locations and their standard errors for the functional scale

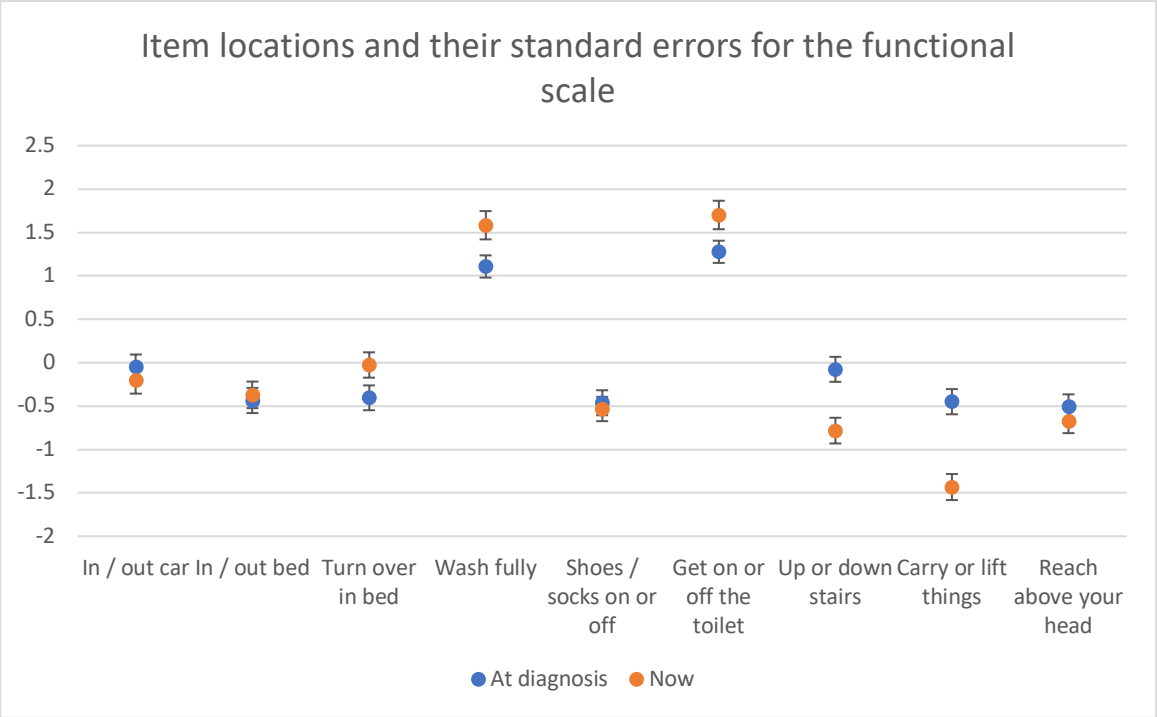
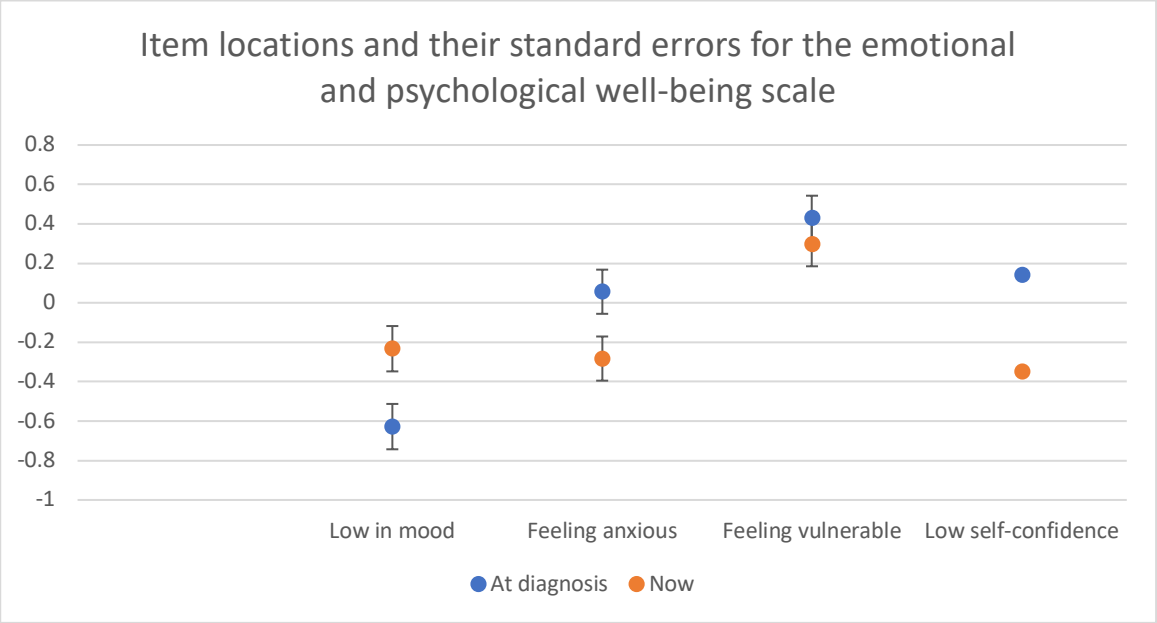


Figure 8.27: Item locations and their standard errors for the emotional and psychological well-being scale



The second possible explanation is that there is differential item functioning by time of some of the items. This seems more likely in this case and indeed, has already been demonstrated to some degree, as when the scales were tested in the combined dataset there was DIF for time for two of the items in the functional scale (in or out of bed and turn over in bed).

To examine the difference in ordering further, plots were drawn of the difference for each item between the two datasets. These are shown in Figure 8.28 and Figure 8.29.

Figure 8.28: Differences in functional item locations between the two datasets

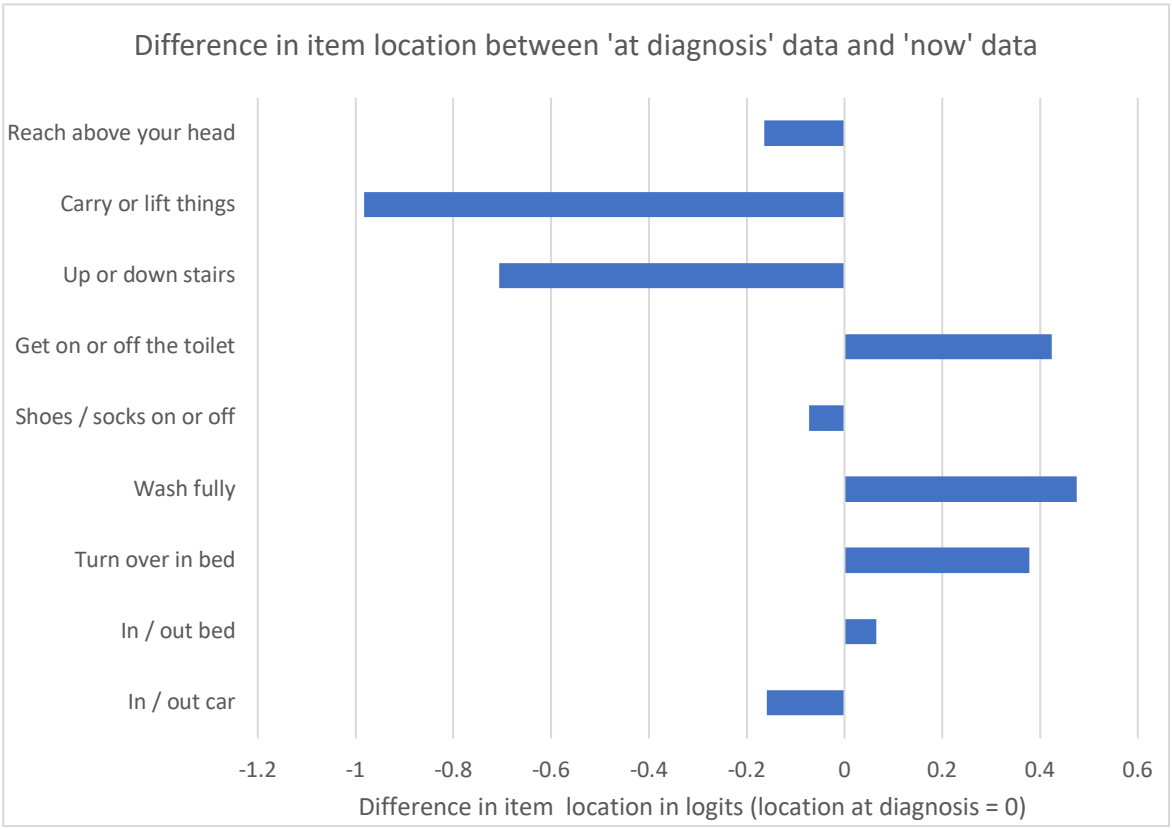
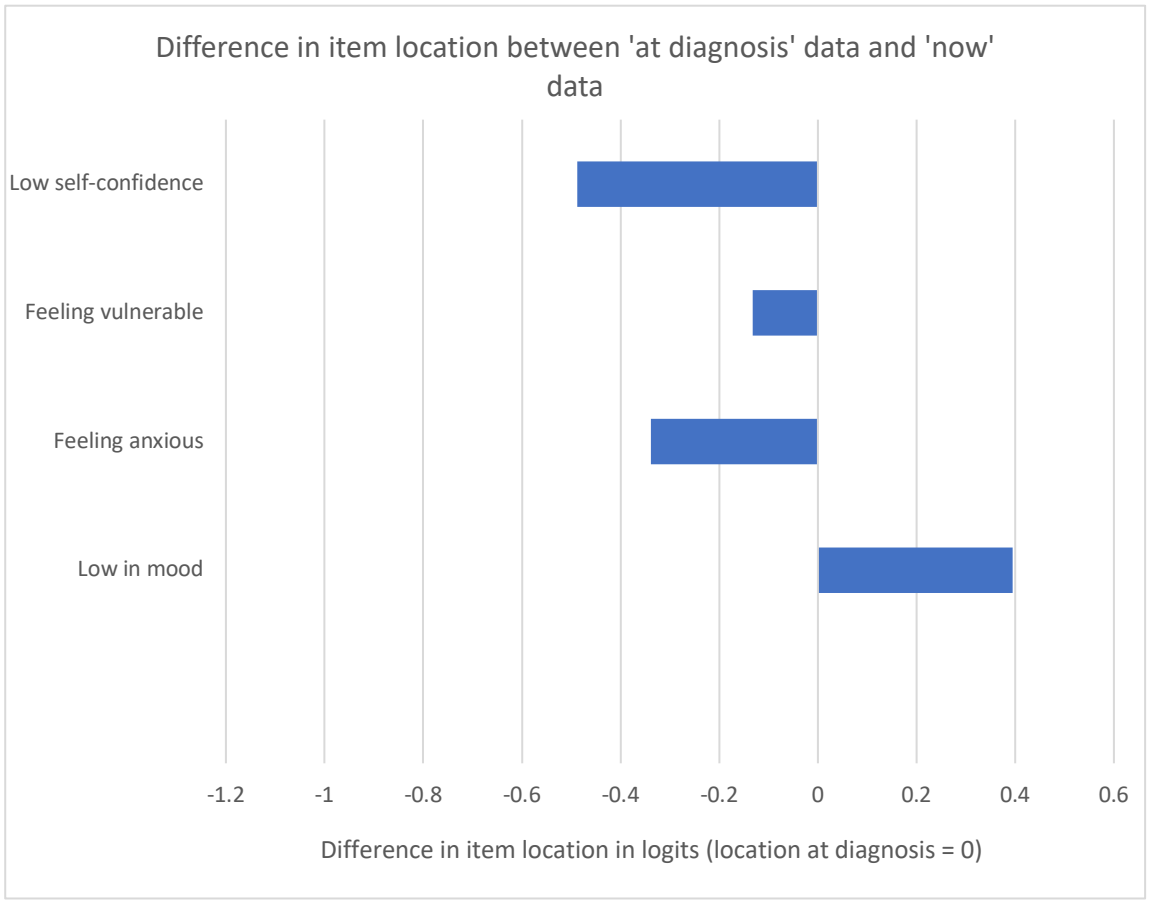


Figure 8.29: Differences in emotional and psychological item locations between the two datasets



These figures show that for the emotional and psychological well-being items, the absolute differences in item location between the two datasets are small (<0.5 logits) and therefore not of significant concern. Of the functional items, 'carry or lift things' and 'walk up or down stairs' both have locations differing by >0.5 logits between the two datasets i.e., these items become easier at a greater rate than the other items at later stages of the condition. This is subtly different to the concept of DIF as DIF is defined as people with the same overall level of functioning, based on their responses to all the items, responding differently to individual items.

In many ways it is perhaps not surprising that ability to complete some activities improves more quickly than in others. However, to resolve this issue to the extent that Rasch modelling can be used for scoring would mean creating two different scoring systems, one for pre-treatment and one for after treatment. This is not the way I intended to use the two datasets so would be a flawed approach. It is also not a good solution to meet the pragmatic aim of creating an instrument that will be user-friendly. It is also important to note that the data from this survey does not simply describe two distinct timepoints – the 'now' data is a from participants ranging from soon after diagnosis up until five years post-diagnosis.

The combined dataset is not unidimensional and does not fit the Rasch model so using the item locations from this dataset is not an option.

One solution is to simply use a sum score of the items. This is the approach taken in the development of scoring for many health measurement instruments, including the HAQ (Fries et al., 1980), the Keele MSK-PROM (Hill et al., 2015) and the SF-36 (Ware & Sherbourne, 1992). A study by McHorney et al. (1997) compared the relative precision of Likert summative scoring and a Rasch scoring model for the functional scale of the SF-36 and found that they were very similar, although there was a small increase in precision in discriminating between clinical groups when the Rasch model was used to compare groups of individuals at the extremes of range of functional ability. Indeed, even PROMs developed to fit a Rasch model such as the EASI-QoL for ankylosing spondylitis (Haywood et al., 2010), and the AAV-PRO for ANCA-associated vasculitis (Robson et al., 2018) tend to use sum scoring for simplicity as calculating a Rasch score necessitates the use of

either computer software or a conversion table, as well as being more complicated conceptually.

Scores for the functional scale will therefore be such that each item is scored 0-2 and the total score is the mean item score as a percentage.

The emotional and psychological well-being scale will be scored such that each item is scored 0-4 and the total score is the mean item score as a percentage.

8.6.3 Steroid side effects domain

The steroid adverse effects section contains 13 items relating to specific side effects, each scored 0-3. The minimum score is therefore 0 (unaffected) and the maximum is 39 (maximally affected) and the result will be given as a percentage.

8.6.4 Back to normal question

The final question, asking about whether a patient feels back to the level of health they were at prior to developing PMR, is only relevant later on in the disease course.

This will not therefore form part of the scoring system but can be used as an add-on to help patients and clinicians make an overall judgement of their health status.

8.6.5 Missing data

The likelihood of having missing data will depend on the context in which the PROM is used. If it is being completed and discussed with a clinician in practice, there is less likely to be missing data. If it is being completed remotely or in a clinical trial situation, scale

scores will be calculated for each domain providing at least half of the items for that domain were answered (5 items for symptoms, 5 items for function, 2 items for psychological and emotional well-being and 6 items for side effects). If fewer than the minimum items were completed, a scale score will not be calculated for this domain.

This approach is consistent with other commonly used scales (including the mHAQ (Fries et al., 1980) and the SF-36 (Ware & Sherbourne, 1992)) and is simple to apply in practice.

8.6.6 Presenting the scores

As discussed in Chapter 6 (Section 6.6.2), overall scores of PROMs can either be presented as their separate components (a profile) or combined to form an index. In the case of the PMR-PROM, it does not make sense to combine the domain scores to present one sum score. Whilst high scores in each scale are a marker of significant impact on disease-related quality of life, each individual scale provides information about different facets of the impact of the condition and high scores in specific sections might require specific responses. For example, a high score in the symptoms section may suggest a need to increase prednisolone dose, a high score in the functional scale might indicate the need to consider physiotherapy or whether extra support at home is needed and a high score in the side effects section may indicate that a reduction in prednisolone dose is indicated. All these scores need to be known and balanced to come to shared decisions about the best management approach.

Scores for the PMR-PROM will therefore be presented as four percentages representing the four domains. Higher scores indicate lower PMR-related quality of life.

8.7 Discussion

In this chapter I have described the process of field testing the PMR-PROM, which has allowed item reduction, formulation and testing of dimension structure and development of the scoring system.

8.7.1 Strengths of the study

The methods of participant recruitment and data collection used in this study meant that I was able to achieve a sample that was of sufficient size for the subsequent statistical processes, and which was demographically representative of the UK population of people with PMR. The support of the NIHR Clinical Research Network facilitated recruitment of practices and participants and the process ran smoothly.

By fully acknowledging the aspects of PMR that make measurement of outcomes difficult at the start of this process, through clinical experience, systematic literature review and participation in the PMR-Special Interest Group of OMERACT, I was able to put in place strategies at the outset to try to mitigate some of these challenges. One major challenge in PMR is the severity of symptoms at the onset (and sometimes during flares) contrasted with a much longer duration of lower-level symptoms for the majority of the disease course. An ideal outcome measure for PMR needs to be able to capture these severe symptoms but also be able to detect fluctuations in lower-level symptoms. In order to ensure that the PMR-PROM was based on data which included people with early stage disease, participants in this state were needed. However, due to the relatively low incidence of the condition, recruiting people as they are diagnosed is difficult. This led to the innovative method of asking participants to complete the PROM retrospectively

based on how they remembered feeling at diagnosis. There is some evidence to support retrospective used of PROMs, as discussed in Section 8.4.3, but to my knowledge, this method has not been used in this context previously.

The willingness to use different psychometric paradigms (Classical Test Theory and Rasch modelling) to achieve a pragmatic outcome rather than being bound to one approach can also be seen as a strength. I initially applied CTT methods but this strategy did not help identify enough items for deletion to create scales of lengths that would be practical to use and also resulted in a functional scale containing more than one factor in one of the datasets (where the two factors were not a meaningful clinical distinction). I therefore went on to apply Rasch modelling as a more rigorous guide to item reduction and assessment of unidimensionality.

The use of Rasch analysis in development of a health instrument is a relatively recent phenomenon (Tennant et al., 2004). Many established measures were developed prior to the recent popularity of Rasch modelling and when researchers have subsequently tested properties of some of these well-established measures using Rasch analysis, problems with dimensionality and assumptions about interval scale scoring have been highlighted. This has been the case for the HAQ (Tennant et al., 1996), the SF-36 (Cordier et al., 2018) the Rowland Morris Disability index (Davidson, 2009; Grotle et al., 2013) and the Neck Disability Index (van der Velde et al., 2009).

Instruments developed more recently, such as the PSORI-QoL for psoriatic arthritis (McKenna et al., 2003), the AS-QoL for ankylosing spondylitis (Doward et al., 2003) and the AAV-PRO for ANCA-associated vasculitis (Robson et al., 2018) have tended to use CTT

and Rasch modelling in their development, as I have used with the PMR-PROM, citing the rigor of this approach.

8.7.2 Limitations of the study

Relying on retrospective report rather than gathering data directly from newly diagnosed participants introduces the possibility of recall bias as already discussed. However, the onset of PMR is often a dramatic event in people's lives with sub-acute onset of significant pain and disability (as described in previous qualitative work) and it is therefore reasonable to presume participants will remember this event well.

Having two, non-independent, datasets increased the complexity of the analysis and created dilemmas about how to manage the differing results. However, it has highlighted some of the complexity of PMR itself - the relative changes in aspects of function / psychological impact over the disease course would not have been apparent without the two datasets and the findings have forced consideration of these issues.

Using two different measurement theory paradigms has also made the process more complex. However, ultimately this has led to a better understanding of the data and ensured a more rigorous process as outlined above.

8.7.3 Strengths and weaknesses of the PMR-PROM

The field-testing process has resulted in amendments to some of the response categories, removal of poorly functioning items to create unidimensional and shorter scales and development of the scoring system.

It has also however, identified some weaknesses of the tool. For the questions on symptom duration and severity there was a ceiling effect in the 'at diagnosis' dataset and a floor effect in the 'now' dataset. This means that the questions cannot discriminate between individuals with very severe symptoms at the onset of the condition or between those with very mild symptoms later on in the disease course. Clinically, the floor effect later on in the disease course is likely to be of more significance as this is the stage where detecting smaller changes in symptom burden could be helpful in decisions about treatment. The likelihood of there being a floor effect was recognised in advance due the nature of PMR itself. The significance of it on the interpretability and responsiveness of the PROM will be evaluated in the testing of the measure that follows.

The person-item threshold distribution for the final functional and emotional and psychological well-being scales show that there are some levels of 'ability' where the items are rather sparse. This indicates that there is likely to be floor and ceiling effects in these domains too and might also cause some problems for the responsiveness of the measure. It could be an area for future work to try to develop items that fill these gaps though this would have to be balanced against additional burden to participants from longer scales.

With regards the scoring system, a pragmatic approach has been taken for all domains. The decision to weight individual items in the symptoms and side effects sections equally was made through discussion with my supervisory team. I did consider going back to a patient group and a wider professional group to gather opinions on this decision but ultimately felt that the small chance of them suggesting significant changes to this approach was outweighed by time and financial constraints.

As already discussed, the scoring system of the functional and emotional and psychological well-being scales was based on sum scoring as a pragmatic response to the finding of having differential item ordering between the two datasets. Whilst it may have been theoretically ideal to have been able to devise a logit-based scoring system, this would have made the PROM more difficult to use in practice. Overall, the process of testing the fit of the data to the Rasch model has still been useful in that it has directed further item reduction by spotlighting poorly functioning items, confirmed category ordering and absence of DIF by age or gender and provided a rigorous test of unidimensionality.

8.8 Further amendments, formatting and naming of the PROM

The version of the PROM produced at the end of the field-testing study was Version 6 (Appendix 8.6: PMR-PROM Version 6).

Further changes were made in an iterative process creating Versions 7-10 (Appendix 8.7: PMR-PROM Version 7 to Appendix 8.10: PMR-PROM Version 10) in response to feedback from the relevant professionals working in the School of Primary, Community and Social Care at Keele University including GPs, rheumatologists, researchers and physiotherapists. The majority of these changes were to the layout and formatting, for example adding tables / boxes, enlarging headings etc. Other changes that were made during this time were:

- Three items were removed from the side effects domain (high blood pressure, high blood sugars and cataracts) as it was felt that these were not easily

identifiable by patients as directly related to prednisolone and may cause difficulties in reporting, e.g., if they were pre-existing conditions.

- The look back period for the questions was changed back to one week rather than three days.
- The instrument was named 'the PMR-impact scale (PMR-IS) and a front page with this title was added.

8.9 Conclusions

The process described in this thesis up until this point has resulted in a PMR-PROM ready for further reliability, responsiveness and validity testing. It has been developed following the principles of COSMIN guidance (Mokkink et al., 2018), which is accepted as the global standard in this field. The original long-list of items was developed from qualitative work and pilot tested with people with the condition. The resultant PROM has now been rigorously field tested in the relevant target population resulting in a shorter measure with scales and scoring system developed. This PROM has been named the PMR-Impact Scale.

Chapter 9: Evaluation of the PMR-Impact Scale

9.1 Introduction

The preceding chapters have described the development of the PMR-IS, a patient-reported outcome measure to evaluate the impact of PMR on people's lives. In order for a PROM to be credibly used in clinical studies, evidence is needed of its reliability, validity and responsiveness. These concepts have been introduced in previous chapters. This chapter discusses the methodological aspects of these measurement properties in more detail and describes the study used to evaluate these properties of the PMR-IS.

9.2 Aims and objectives

To evaluate the psychometric properties of the PMR-IS, assessing its:

- Test-retest reliability
- Construct validity
- Responsiveness
- Smallest detectable change and minimally important change values

9.3 Methodology

9.3.1 Methodology relevant to data collection

This research will collect data using a postal survey as used in the field-work study, but with participants recruited from both primary and secondary care sites. The advantages and disadvantages of postal surveys for data collection have been discussed in Chapter 8

(8.3.1). The decision to expand recruitment for this study to include rheumatology clinics was made to increase the rate of recruitment.

9.3.2 Reliability

As introduced in Chapter 6 (Section 6.7), reliability is defined by COSMIN as

"the extent to which scores for patients who have not changed are the same for repeated measurements under several conditions" (Mokkink et al., 2010a)

Depending on the nature of a measurement instrument, there are several possible components to its overall reliability - internal consistency, test-retest reliability and inter-rater reliability. The internal consistency of the PMR-IS has been demonstrated through the development phase (during which it was referred to as the PMR-PROM) and as it is a self-completed questionnaire, inter-rater reliability is not applicable. The component that needs evaluating in this study is therefore its test-retest reliability.

Parameters of reliability and measurement error

Any observed score (Y), for a given individual, is a function of the true score (η) and the error in measurement (ε).

$$Y = \eta + \varepsilon$$

There is an important distinction between 'agreement' or 'reproducibility', which are terms describing the closeness of the scores on repeated measures, expressed in the unit of the respective measurement scale, and the statistical concept of 'reliability', which

describes how well individuals can be distinguished from each other despite the presence of measurement error (Terwee et al., 2007).

A reliability parameter relates the measurement error to the variability between individuals (de Vet, Terwee, et al., 2011e). Reliability parameters range in value from 0 (unreliable) to 1 (totally reliable). If measurement error is small in comparison with variability between individuals, the reliability parameter will be close to 1. If, however, there is little variation between individuals and the measurement error is large, the test cannot discriminate easily between individuals and its reliability will be closer to 0.

Different parameters for reliability and measurement error are used depending on the type of variable in question, as summarised in Table 9.1. As this study involves continuous data, I will focus on the intra-class correlation coefficient (ICC) and the standard error of the measurement (SEM) / limits of agreement (LoA).

Table 9.1: Parameters of reliability and measurement error according to variable type

(taken from (de Vet, Terwee, et al., 2011e))

	Continuous	Ordinal	Nominal
Parameter of reliability	Intra-class correlation coefficient	Intra-class correlation coefficient or weighted kappa	Unweighted kappa
Parameter of measurement error	Standard error of the measurement or limits of agreement	% agreement	% agreement

Intra-class correlation coefficient

The ICC is the ratio of the inter-individual variance (the variance within a population) to the total variance (the inter-individual variance plus intra-individual variance or measurement error). Which intra-individual (error) variances are included in the calculation of an ICC is determined by the situation of interest. If absolute agreement is important, the error variance due to systematic differences between time points needs to be included but if only ranking (consistency) is important, only random (error) variance needs to be considered (McGraw & Wong, 1996).

$$ICC_{consistency} = \frac{\text{interindividual variance}}{\text{(residual error variance + interindividual variance)}}$$

$$ICC_{agreement} = \frac{\text{interindividual variance}}{\text{(systematic error variance + residual error variance + interindividual variance)}}$$

The ICC approaches 1 when the error variance is negligible compared to the individual variance and approaches 0 when the error variance is large compared to the individual variance. Values >0.7 are generally considered acceptable (Terwee et al., 2007).

Standard error of the measurement

The standard error of the measurement (SEM) is a measure of the distribution of repeatedly measured values on a person on the same instrument around his or her 'true' score (de Vet, Terwee, et al., 2011e) i.e., it is the standard deviation (SD) of a number of measurements made on a single individual.

If there is only one person on whom the measurements are performed, the SEM is the same as the SD of the repeated measurements for that individual (Baker, 2016). If the sample contains multiple individuals, each with repeated measurements performed on the instrument under test, the SEM can be calculated in a number of ways.

One method is to use the root mean square average (Baker, 2016) i.e., the square root of the mean of the squared standard deviations of each individuals' measurements

$$SEM = \sqrt{((SD_1^2 + SD_2^2 + SD_3^2 + \ldots + SD_n^2)/n)}$$

This is similar to one of the methods proposed by de Vet, Terwee et al. (2011e), which uses the SD of the set of pairwise differences ($SD_{difference}$) between two sets of measurements in a sample of stable individuals. As the difference scores are based on two measurements, giving twice the measurement error, the $SD_{difference}$ is divided by $\sqrt{2}$ to give the SEM (the SD around a single measurement).

$$SEM = SD_{difference} / \sqrt{2}$$

With both of these approaches, it is $SEM_{consistency}$ rather than $SEM_{agreement}$ which is calculated as the SD of the differences does not include the systematic error (the error due to systematic differences between the time points).

Alternatively, the SEM can be calculated using the error variances from the ICC calculation (de Vet, Terwee, et al., 2011e), as per the formulae below:

$$SEM_{consistency} = \sqrt{(\text{residual error variance})}$$

$$SEM_{agreement} = \sqrt{(\text{systematic error variance + residual error variance})}$$

The limits of agreement

The 'limits of agreement' are a statistic related to the Bland and Altman plot. On this plot, the mean of the scores from two repeated measurements are plotted against the differences between the scores (Bland & Altman, 1986). Lines are drawn to show the mean systematic difference between the scores (d) and then at $d+/- 1.96xSD_{difference}$. If the difference scores are Normally distributed, 95% will fall between these two lines, which are denoted the limits of agreement. Observed changes that are outside of these limits of agreement are unlikely to be due to measurement error and more likely to represent real change. The limits of agreement can therefore be used to determine the smallest detectable change, which is discussed further in section 9.3.5.

It is important to note that the Bland and Altman method does not determine whether the calculated limits of agreement are acceptable i.e. whether the two tests are sufficiently similar. This can only be judged through knowledge about the construct being measured and the research question or clinical goals (Giavarana, 2015).

Methodological considerations for test-retest reliability studies

1. Population: the reliability of a measurement instrument is dependent on the distribution of the characteristic being studied in the sample population i.e., it is affected by how heterogenous the population is for that characteristic. It is therefore important that a reliability study takes place using participants similar to those in whom the instrument will be used in the future.

2. Time interval: the appropriate time interval for test-retest assessment is a balance between the stability of the characteristic and the independence of repeated tests (which for questionnaires, primarily means any memory effect).
3. Sample size: test-retest reliability studies need a large enough sample to give an acceptable confidence interval around the estimated reliability parameter. It is suggested by experts in the field that 50 is a reasonable number to achieve this. A 95% confidence interval of +/-0.1 can be achieved with 50 participants undergoing two repeat measurements for an ICC of 0.8 (de Vet, Terwee, et al., 2011e).

Test-retest reliability testing of the PMR-IS

Following the principles described above, test-retest reliability of the PMR-IS will be evaluated in the population of interest (patients with PMR), who were stable between the two time points of PROM completion. Data from participants who rate themselves as 'the same' for the domain in question over a time interval of between 2 and 6 weeks will be used. As the scores of the questionnaire are continuous data, the ICC and the SEM / limits of agreement will be calculated to describe the reliability and the measurement error.

9.3.3 Construct validity

As discussed in Chapter 6 (Section 6.8), the validity of an instrument describes whether it measures the things it purports to measure. There are different components of validity (e.g. content validity, criterion validity, structural validity), but the aspect of validity of the PMR-IS that needs assessing in this evaluation phase is construct validity i.e. whether the

instrument produces scores that are expected, based on what is known about the construct it is trying to measure.

Validation of a complex, multi-dimensional measurement instrument is not a 'one-off' test but rather a process of gathering evidence to support hypotheses about relationships between the new measure and scores on instruments measuring related constructs or scores from known sub-groups of patients (e.g. treated vs untreated). The validation process cannot be disentangled from theories about the construct itself and as understanding of the underlying construct develops with time, the validity of ways of measuring it must be reconsidered.

Methodological considerations for studies of construct validity

Construct validity is evaluated through hypothesis testing. Hypotheses need to be specified *a priori*, based on theory or literature findings, and tested using empirical data gathered from the specific population of interest. Details of the comparator instruments need to be clearly described and the hypotheses ideally need to describe the direction and expected magnitude of change (de Vet, Terwee et al., 2011).

When the construct is multidimensional, each scale or part of the instrument that measures a specific dimension needs to be evaluated by forming hypotheses for the different dimensions separately.

As for test-retest reliability studies, there are no set rules on sample size for studies of construct validity but de Vet *et al* recommend a minimum of 50 participants with sample of over 100 preferred (de Vet, Terwee, et al., 2011e).

The strength of relationship between two variables is assessed through calculation of a correlation coefficient. These have values ranging from -1.0 to 1.0 where -1.0 represents perfect negative correlation and 1.0 represents perfect positive correlation. Which correlation coefficient is used is determined by the nature of the data. Pearson's correlation coefficient can be used if the data is continuous and there is a linear relationship between the variables (a change in one produces a proportional change in the other). Spearman's correlation coefficient is based on rank variables rather than raw data and can be used if the data is either continuous or ordinal and there is a monotonic relationship between the variables (an increase / decrease in one variable results in an increase / decrease in the other but the rate of increase or decrease is not necessarily constant) (Ramzai, 2020).

Statistical significance of the correlation is not relevant because the issue is not whether the correlation differs from zero but whether the correlation is of a pre-defined magnitude. In the quality criteria proposed by Terwee et al. (2007), construct validity is rated satisfactory if at least 75% of the results of a study, of at least 50 participants per sub-group, are in agreement with pre-specified hypotheses.

Construct validity testing of the PMR-IS

This will be assessed by testing hypotheses about relationships between scores on the PMR-IS and scores on two other instruments, the SF-36 (Ware & Sherbourne, 1992) and the mHAQ (Pincus et al., 1983). The rationale for choosing these particular instruments is explained in Section 9.4.3.

Scatter plots will be drawn to assess the relationship between the variables for each hypothesis and, providing there is at least a monotonic relationship, the appropriate correlation coefficient (Spearman's or Pearson's) will then be calculated.

9.3.4 Responsiveness

As discussed in Chapter 6 (Section 6.9), responsiveness is defined by COSMIN as:

"the ability of the instrument to detect change over time in the construct to be measured" (Mokkink et al., 2010a).

It is sometimes referred to as longitudinal construct validity (as it can be thought of as the 'validity' of a change score) and aspects its assessment are broadly similar to the methodological approach to validity testing. Again, the process of demonstrating responsiveness is one of continuous evidence gathering rather than a single test.

If there is a gold standard available for comparison, a criterion approach can be used but if not, a construct approach is used.

Methodological considerations for studies of responsiveness

To assess the responsiveness of a measure, a longitudinal study in which at least a proportion of the participants are expected to improve or deteriorate is needed. As for validation studies, sample sizes of at least 50 participants are recommended for studies of responsiveness (de Vet, Terwee, et al., 2011f).

If there is no suitable instrument that can be used as a comparator, 'anchor questions' can be used, in which patients are asked a single question at follow-up to indicate how

much they have changed since baseline. These should be formulated so that they are specific to the construct being measured (e.g. on a 5-point scale ranging from much worse to much better, how has your quality of life changed since your initial assessment?). Hypotheses about expected differences in change score between groups defined by their responses to the anchor questions, can then be formulated and tested. If an instrument has subscales that measure multiple constructs, specific comparator questions can be formulated for each construct measured and multiple hypotheses tested.

There are no set standards on numbers of hypotheses to be tested or of the proportion of hypotheses that need to be accepted to be called a good result (de Vet, Terwee, et al., 2011f).

Statistical tests for assessment of responsiveness

There are a number of statistical approaches and tests used in studies of responsiveness and the choice is, in part, determined by the level of measurement of the instruments (similar to the approach for validity testing set out in Table 6.2).

If the scores on the new measure and a comparator are continuous variables and either a gold standard exists or a construct approach is being followed (in which hypotheses are formulated about relationships between change scores on existing measures and expected change scores on the new measure), correlations between change scores can be used.

If an anchor question is used as a comparator and the new measure is continuous, subjects can be grouped into ordinal categories according to their response to the anchor

question. Hypotheses about the expected direction and magnitude of mean changes in the scores on the new measure for the different categories can then be formulated and tested.

Other measures of responsiveness

There are other statistical tests that are widely reported in responsiveness studies, but which COSMIN recommends are not used for this purpose e.g. the standardised response mean (SRM) and the effect size (ES) statistic (de Vet, Terwee, et al., 2011f).

The SRM is the mean change score of a group of patients divided by the standard deviation of this change score. The ES is usually calculated as the mean change score of a group of patients divided by the standard deviation of the baseline scores of the group. These scale-free statistics facilitate interpretation of the difference between groups following an intervention because they relate the size of the change to the distribution of the sample and thus provide context (Coe, 2002). They are therefore useful for quantifying effects measured on unfamiliar or arbitrary scales. Unlike statistical significance, ES and SRM are not dependent on sample size and they are therefore also useful in quantitatively comparing results from studies done in different settings (McCleod, 2019).

However, an instrument cannot be said to be responsive just because the ES or SRM is high – they are measures of the magnitude of change scores in a given situation, rather than the validity of the change scores. A high magnitude of change does not necessarily mean that the instrument is good at detecting change in the construct being measured. The observed change might actually be smaller than the true change in the construct

being measured (for example if there was a ceiling effect or a lack of relevant items). In addition, because the ES and SRM are dependent on standard deviation of the sample they will be higher if the group is homogenous or the variation in treatment effect is small.

For these reasons, SRM and ES are only appropriate as measures of responsiveness if they are used in conjunction with explicitly stated hypotheses about their expected magnitude (de Vet, Terwee, et al., 2011f).

Responsiveness testing of the PMR-IS

As there is no gold standard measure for PMR, nor another instrument with high quality evidence demonstrating its responsiveness in the relevant constructs in PMR, an anchor-based method will be used.

Hypotheses about the expected direction and degree of change on the PMR-IS between two time points for groups categorized by their response to the anchor question will be formulated and tested.

9.3.5 Smallest detectable change and minimally important change

The principles of the smallest detectable change (SDC) and minimally important change (MIC) values were introduced in Chapter 6 (Section 6.10.4).

The smallest detectable change is closely aligned to the reliability of a measurement tool. As described in Section 9.3.2 on test-retest reliability, data from a sample of people known to have stayed the same on a particular construct between two time points, can

be used to calculate the 'limits of agreement' using the Bland and Altman method (Bland & Altman, 1986). These are the values between which 95% of the differences between the scores at the two time points, are expected to fall. Measurements that fall outside these limits are more likely to represent real change than measurement error.

The limits of agreement (equivalent to mean +/-1.96 x $\sqrt{2}$ x SEM) therefore represent the smallest detectable change at individual level (SDC_{ind}).

At group level, the SDC will be smaller and can be calculated as

$SDC_{group} = SDC_{ind}/\sqrt{n}$

The MIC can be calculated using a variety of different methods and there is no firm consensus on which is best (Spies-Dorgelo et al., 2006; Van Der Roer et al., 2006; Wells et al., 2001). The different approaches can broadly be divided into anchor-based methods or distribution-based methods.

Anchor-based methods

The use of anchor questions in the assessment of responsiveness of a measure was discussed in Section 9.3.4. In addition to allowing analysis of responsiveness, the anchor questions can be used to identify a group of patients who report that their condition has slightly worsened or slightly improved. The mean change score of these groups can be taken as the MIC. This is known as the mean change score method.

Another anchor-based method to determine the MIC is to use receiver-operating characteristic (ROC) curves. This approach is similar to the use of ROC curves in a

diagnostic study - the measurement instrument is considered the 'diagnostic test' and the anchor the 'gold standard'. The instrument's 'sensitivity' is the proportion of importantly improved or deteriorated patients according to the anchor that are correctly identified as such by the instrument i.e., the true positives. The 'specificity' is the proportion of patients with no important change on the anchor that is correctly identified as such by the instrument i.e. the true negatives. A ROC curve plots the sensitivity (true positives) against 1-specificity (false positives). The area under the curve represents the probability of correctly identifying a patient who has improved or deteriorated from randomly selected pairs of improved and stable or deteriorated and stable patients. An AUC of 1.0 indicates that the instrument is able to discriminate perfectly whereas and AUC of 0.5 indicates that the instrument does not discriminate any better than chance. COSMIN standards state an AUC of >0.7 is deemed adequate (Terwee et al., 2007). A cut-off value on the change score of the measurement instrument can be determined at which the balance between sensitivity and specificity is optimal and there is least misclassification. This optimal ROC curve cut-off point is taken as the MIC (Spies-Dorgelo et al., 2006).

Distribution-based methods

These methods estimate the MIC using statistical parameters of the sample. The observed change in the measurement instrument under study is expressed in relation to some form of variation (e.g., the effect size or the SEM) to obtain a standardised metric. However, this means the MIC is related to the heterogeneity of the study population in which it is determined and the same magnitude of change will be considered variably important in populations of different heterogeneity.

Although such methods can demonstrate if a change has occurred and the magnitude of it, they do not provide a good indication of the clinical importance of the observed change and therefore COSMIN do not consider them adequate ways of assessing MIC (de Vet, Terwee, et al., 2011d).

The advantages and disadvantages of using the different types of approach to calculate the MIC are set out in Table 9.2.

Whichever method is used to calculate the MIC, it is important to note that a measurement instrument does not have a fixed MIC value. The MIC will depend on the baseline value of the population in the test (patients who are more severely affected often need a larger change to indicate an important change), the direction of change (improvement versus deterioration) and on the wording used in the anchor question.

Table 9.2: Advantages and disadvantages of anchor-based and distribution-based approaches to calculating the MIC

Approach to identifying the MIC	Advantages	Disadvantages
Anchor-based **(mean change or ROC methods)**	Concept of minimal importance explicitly defined Global rating anchor questions provide the single best measure of the significance of the change from the patient perspective	Do not take the variability of the scores in the sample into account There is typically limited information about the reliability and validity of the anchor question itself.
Distribution-based **(using the ES or SEM)**	Take the variability of the sample into consideration	Do not provide a good indication of the true importance of the change. The same change will be considered to be of differing importance in populations of different heterogeneity

The lack of one clearly superior method has led to the recommendation that an integrated approach in which a combination of anchor-based and distribution-based methods are used to determine clinically meaningful change should be used (Crosby et al., 2003; Dworkin et al., 2008). COSMIN also endorse using multiple methods and advocate being explicit about the diversity of the MIC, but they favour using several different anchor-based approaches over including distribution-based ones (de Vet, Terwee, et al., 2011d).

Interpretability of the PMR-IS

Data collected during this study will be used to provide some indicative information about the interpretability of the PMR-IS, acknowledging that the estimates of these parameters are not fixed measurement properties of the instrument.

The risk of floor and ceiling effects will be assessed by analysing the distribution of responses for each domain.

The SDC at group and individual level will be calculated using the formulae described in Section 9.3.5.

An MIC for each domain will be estimated using two different methods - the mean change method (using the anchor question to identify a group who have experienced a small change on the relevant domain) and the ROC method.

9.4 Methods

9.4.1 Protocol development

The protocol for this study was developed in August 2019. Copies of the practice invitation letter, participant invitation letter and participant information sheet are given in Appendix 9.1: Practice invitation letter, Appendix 9.2: Participant invitation letter and Appendix 9.3: Participant information sheet.

9.4.2 Ethics and governance

A favourable opinion was given by the Proportionate Review Sub-committee of the South Central – Hampshire B Research Ethics Committee on 10th October 2019 (REC reference

19/SC/0525) and NHS Health Research Authority (HRA) approval was given on 17th October 2019 (Appendix 9.5: Confirmation of ethical approval). No amendments were required.

Study recruitment was paused between March 2020 and June 2020 due to the Covid-19 pandemic.

9.4.3 Questionnaires used and process of data collection

Version 10 of the PMR-PROM was used. This is the version created at the end of the field testing and development stage and named the PMR-Impact Scale (PMR-IS) (see Appendix 8.10: PMR-PROM Version 10).

It was collated into a booklet with the other questionnaires needed for the study and distributed by post for participants to complete themselves in their own time.

- Study booklet 1 contained the PMR-IS, the mHAQ and the SF-36.
- Study booklet 2 contained a series of anchor questions, the PMR-IS and the mHAQ.

Each contained a front sheet of demographic questions and booklet 1 also contained a consent page.

Comparator questionnaires

Two comparator questionnaires, the mHAQ (Pincus et al., 1983) and the SF-36 (Ware & Sherbourne, 1992), were chosen for construct validity testing of the PMR-IS. These are included as Appendix 9.6: The mHAQ and Appendix 9.7: The RAND SF-36 Questionnaire.

The mHAQ (Pincus et al., 1983) is described in Chapter 5 (Section 5.1.2). As detailed in that chapter, it was developed for use in rheumatoid arthritis but is widely used as an assessment of function in many different rheumatological conditions (Maska et al., 2011) and some supporting evidence relating to its measurement properties in PMR was found in my systematic review (Chapter 5). This was therefore selected as the best comparator for the function domain. Alternative measures used to assess function in other PMR research studies, such as the HAQ-DI and the assessment of elevation of the upper limbs (EUL), were determined to be less suitable based on the findings from my systematic review.

The mHAQ was included in the second study booklet to provide data for analysis of its test-retest reliability for the ongoing OMERACT work. This analysis is not presented in this chapter as it is not part of the aims and objectives of this PhD.

The SF-36 (Ware & Sherbourne, 1992) is a measure of overall health status and comprises eight scales assessing 1) limitations in physical activities, 2) limitations in social activities because of physical or emotional problems, 3) limitations in role activities because of physical problems, 4) bodily pain, 5) general mental health, 6) limitations in role activities because of emotional problems, 7) vitality (energy and fatigue) and 8) general health perceptions. The version used in this study was produced by RAND Healthcare (https://www.rand.org/health-care/surveys_tools.html) who have made this survey freely available for public use.

The SF-36 is a well-established and highly used generic measure, which assesses similar and related constructs to the domains of the PMR-IS. It is the most commonly used generic QoL measure used in rheumatological clinical trials and observational studies

(Wolfe et al., 2010). A commonly used alternative measure of QoL is the suite of EQ-5D instruments (EuroQol group, 1990). The EQ-5D includes preference weighting for health states and is used in cost effectiveness analyses and where comparisons across disease states is desired. It is, however, less sensitive to change than the SF-36 because it summarises health status in just five questions and it has also been shown to be more sensitive to outliers (Brazier et al., 1993; Wolfe et al., 2010). For the purposes of this validation study therefore, where neither cost effectiveness analysis nor comparison across disease states is needed, I decided to use the SF-36.

Anchor questions

Five anchor questions were written to classify participants into groups who self-identified that a particular aspect of their condition had greatly improved, slightly improved, stayed the same, slightly worsened or greatly worsened (Table 9.3). These groupings were then used in analysis of test-retest reliability and responsiveness.

Table 9.3: Anchor questions used in the second questionnaire booklet

Anchor question	Response options
Compared to when you previously completed this questionnaire, have your symptoms from your PMR…?	1) Improved a lot 2) Improved a little 3) Stayed the same 4) Worsened a little 5) Worsened a lot
Compared to when you previously completed this questionnaire, has the amount your PMR is limiting your activities…..?	1) Improved a lot 2) Improved a little 3) Stayed the same 4) Worsened a little 5) Worsened a lot
Compared to when you previously completed this questionnaire, has the amount PMR is affecting your emotional well-being…..?	1) Improved a lot 2) Improved a little 3) Stayed the same 4) Worsened a little 5) Worsened a lot
Compared to when you previously completed this questionnaire, have the side effects from your prednisolone……?	1) Improved a lot 2) Improved a little 3) Stayed the same 4) Worsened a little 5) Worsened a lot
Compared to when you previously completed this questionnaire, has your quality of life linked to your PMR……?	1) Improved a lot 2) Improved a little 3) Stayed the same 4) Worsened a little 5) Worsened a lot

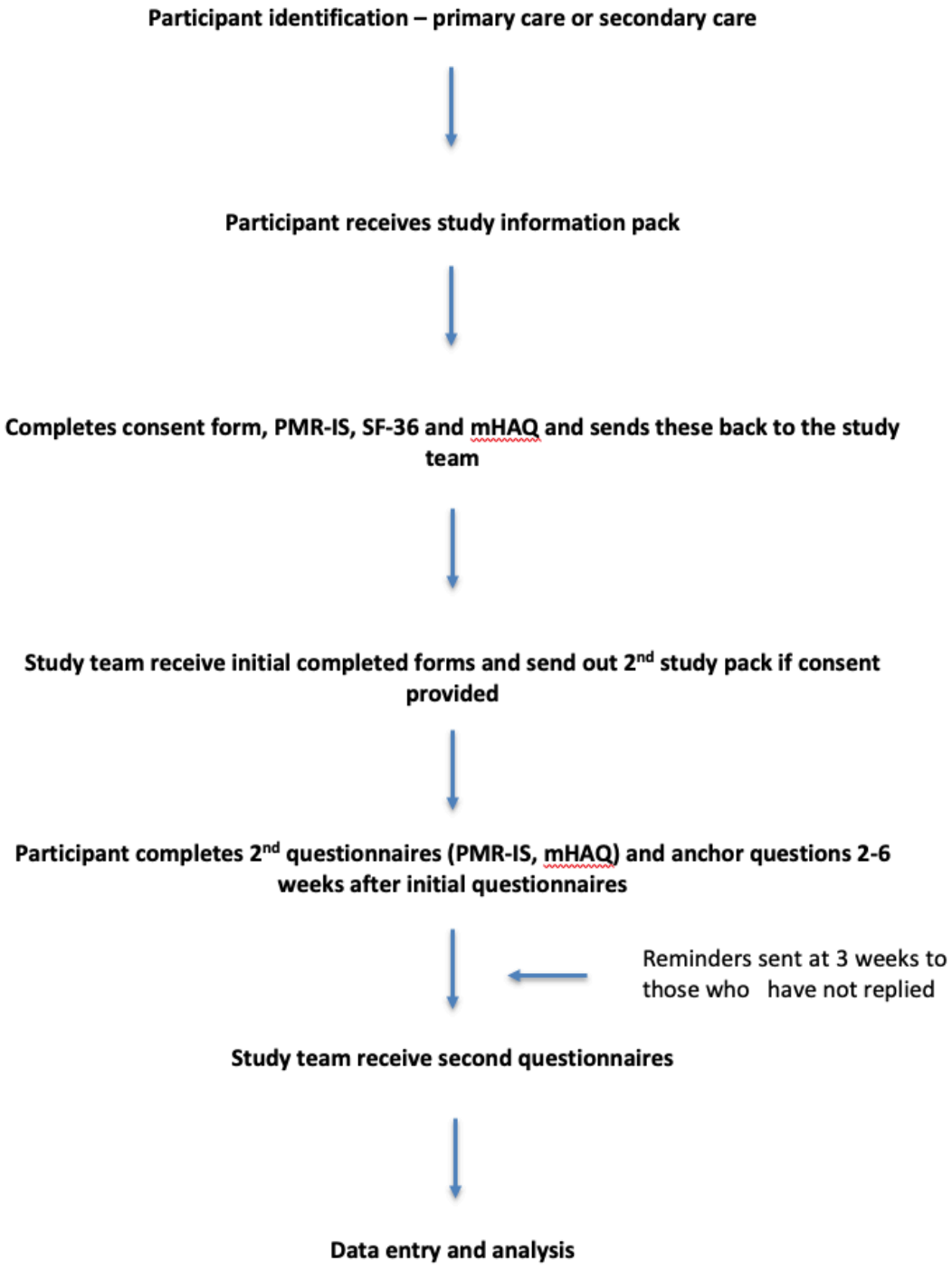
Data collection

Eligible potential participants were sent a study pack containing an invitation letter, a patient information sheet and study booklet 1. They were asked to complete the demographic details section, the consent page (to agree to a second pack to be sent) and the questionnaires and send the booklet back in the enclosed pre-paid envelope, addressed to me at Keele University.

Those that responded and consented were sent the second study pack to be completed between two and six weeks after the first. Reminder letters were sent if this was not received within three weeks of it being sent out.

A flow diagram of the process is given in Figure 9.1.

Figure 9.1: Study flow chart



9.4.4 Sample size

As discussed in the methodology section above, a sample size of 50 is suggested as a minimum for calculation of test-retest reliability (de Vet, Terwee, et al., 2011e). Therefore, this study requires at least 50 participants who report that they 'stayed the same' on the anchor question.

The suggested sample size for construct validity and responsiveness testing is also at least 50 (de Vet, Terwee, et al., 2011; de Vet, Terwee, et al., 2011f). All completed baseline questionnaires will be eligible for construct validity analyses whereas responsiveness calculations will require data from participants who complete the second booklet and have either improved or worsened.

I therefore aimed to recruit 200 people in total to capture sufficient participants who changed (improved or worsened) between the two completion time points as well as those that stayed the same.

9.4.5 Recruitment of research sites

Recruitment took place from both primary and secondary care. The decision to include a rheumatology clinic as a recruitment site was made to maximise participant numbers. Primary care recruitment occurred with the support of the West Midlands NIHR Clinical Research Network (CRN), as for the field-testing stage outlined in Chapter 8. The CRN advertised the study to practices across Staffordshire and if they were willing to participate, they were sent the full study information. Practices were reimbursed for their time following CRN guidance.

Secondary care recruitment took place with the support of the Midlands Partnership NHS Foundation Trust (MPFT) Research and Innovation Department.

9.4.6 Identification of potential participants

Participating general practices were asked to carry out a search of their patient database to identify patients with a new clinical code of "Polymyalgia Rheumatica" entered in the preceding two years (clinical code N20 PMR only, exclude child code N200 GCA with PMR).

A GP from the practice screened the list of identified patients against the inclusion and exclusion criteria (below) to identify suitable participants to invite.

Names and addresses of identified potential participants were uploaded to Docmail, a secure mailing service (https://www.cfhdocmail.com/), and the study pack was sent via this mailing service.

People attending rheumatology clinics at the Haywood Hospital (a specialist secondary care rheumatology centre) who had a diagnosis of PMR made in the preceding two years, were identified by searching the database of clinic attendees and screened using to the same inclusion and exclusion criteria by a rheumatology clinician from the hospital. Eligible participants were sent a study pack via Docmail as above.

Inclusion criteria:

Diagnosis of PMR made within the previous 2 years and not subsequently changed.

The diagnosis should be supported by the following features, which are based on the British Society for Rheumatology / British Society for Health Professionals in Rheumatology guidelines (Dasgupta, Borg, & Hassan, 2010):

- Age > 50 years.
- Bilateral shoulder or pelvic girdle aching or both for at least 2 weeks.
- Morning stiffness.
- Evidence of an acute phase response (raised ESR / CRP).
- Diagnosis made by a rheumatologist despite the presence of atypical features (e.g. normal ESR / CRP).

Exclusion criteria:

- Diagnosis of GCA.
- Inability to read / write English well enough to understand the instructions and complete the questionnaire.
- Comorbidities that made an invitation to participate in the study inappropriate in the view of the screening clinician (dementia, significant anxiety / depression, receiving end of life care etc.).

9.4.7 Data management

Each participant was assigned a unique reference number as study pack one was received. Study IDs, names and addresses were entered into a secure participant database along with the date the booklet was received. The date booklet two was sent, the date of a reminder letter if needed, and the date booklet 2 was received were recorded in the same database to allow mailing to be tracked. The hard copies of pages

from the study booklets containing personally identifiable data were removed and stored separately.

Data from the study booklets were entered into a separate database, in which participants were only identifiable by their study ID number.

9.4.8 Data analysis

Data were analysed using IBM SPSS version 27 (IBM, 2020).

Descriptive statistics were calculated for the sample for age, gender, duration since diagnosis and prednisolone dose.

Mean item scores and percentage scale scores were calculated for each domain of the PMR-IS, providing at least half of the items for each domain were answered (5 items for symptoms, 5 items for function, 2 items for psychological and emotional well-being and 6 items for side effects). If fewer than the minimum items were completed, a scale score was not calculated for this domain.

Scores for the mHAQ were calculated as per the instructions for the tool (Maska et al., 2011) – the mHAQ score is obtained by *"adding all scored items together (at least 6 of the 8 items are required) and dividing by the total number of items answered"*.

Scale scores for the SF-36 were also calculated according to the standard instructions for the tool (RAND, 2000). Numeric values were transformed using the scoring key and then items in the same scale were averaged to create the 8 scale scores. Missing data (items left blank) were not taken into account when calculating the scale scores such that scale scores represent the mean for all the items in the scale that the respondent answered, provided at least half the items were answered.

Test-retest reliability:

Participants who reported that they 'stayed the same' for a particular domain between the two time points on the relevant anchor question were identified. Their scores on the PMR-IS for that domain were then compared. As the scores were continuous data and to ensure that any systematic differences were also captured in the error variance, the parameter of reliability calculated was the $ICC_{agreement}$.

$$ICC_{agreement} = \frac{\text{interindividual variance}}{\text{(systematic error variance + residual error variance + interindividual variance)}}$$

The limits of agreement for each domain were calculated using the Bland and Altman method. The 'difference between the two measurements' and the 'mean of the two measurements' were calculated for each participant who had stayed the same for that domain. To satisfy the requirements of the analysis, the distribution of the mean differences was checked for normality by drawing histograms and visually assessing the fit.

The differences were then plotted against the mean values on a scatter plot. Lines were drawn on the y-axis at the level of the mean difference and at the mean +/-1.96*SD – the limits between which 95% of the differences between the two measurements are contained.

To calculate the $SEM_{agreement}$ for each domain, the individual variance components were estimated using a SPSS VARCOMP analysis (de Vet, Terwee, et al., 2011e) and then used in the following formula:

$SEM_{agreement}$ = √(systematic error variance + residual error variance)

Construct validity

The hypotheses that were tested are set out below. Scatter plots were drawn to test the assumptions of the correlation coefficient and Spearman's r was calculated. Construct validity was deemed satisfactory if at least 75% of the results were in accordance with the hypotheses (Terwee et al., 2007).

1. The symptoms score of the PMR-IS is moderately to strongly negatively correlated (-0.5 to -0.7) with the bodily pain and energy / fatigue scores of the SF-36.
2. The score from the functional domain on the PMR-IS is strongly positively correlated ($r>0.6$) with the score on the mHAQ and strongly negatively correlated with the physical functioning, social functioning and role limitation physical scores of the SF-36 ($r<-0.6$)
3. The score from the emotional and psychological well-being domain of the PMR-IS is strongly negatively correlated ($r<-0.6$) with the emotional well-being, social functioning and role limitation emotional scores of the SF-36.
4. The score from the steroid side effects domain of the PMR-IS score correlates negatively ($r<-0.2$) with the general health scores of the SF-36.
5. The symptoms score of the PMR-IS correlates positively and moderately strongly ($r>0.4$) with the functional domain of the PMR-IS.

Responsiveness

Participants were grouped into those that improved a lot, improved a little, worsened a lot or worsened a little for each domain according to their responses to the anchor questions. If any of the changed groups contained fewer than 35 respondents, the categories were merged so that, for example, the 'slightly improved' and 'significantly improved' groups formed one category of 'improved'. Those reporting that they had stayed the same were not included in responsiveness analyses.

The following hypotheses were tested:

1. There will be a qualitative trend in mean change scores on the PMR-IS symptoms domain from those reporting that their symptoms have improved a lot to those reporting that their symptoms have worsened a lot.
2. There will be a qualitative trend in mean change scores on the PMR-IS function domain from those reporting that their functional ability linked to their PMR has improved a lot to those who report that it has worsened a lot.
3. There will be a qualitative trend in mean change scores on the PMR-IS emotional and psychological well-being domain from those reporting that their psychological well-being linked to their PMR has improved a lot to those who report that it has worsened a lot.
4. There will be a qualitative trend in mean change scores on the PMR-IS steroid side effects domain from those reporting that their steroid side effects have improved a lot to those who report that they have worsened a lot.
5. There will be a qualitative trend in mean change scores on the symptoms, function and psychological and emotional well-being domains between those that report

their overall PMR-related quality of life (PMR-QoL) as improved, the same or worse.

The side effects domain was excluded from the hypothesis 5 because the relationship between the side effects score and overall PMR-related QoL is not as straightforward as the other domains. Improvement in symptoms, functional impairment and the emotional and psychological impact of PMR are expected to correlate with improvement in overall disease related QoL. However, whilst a lower burden of steroid side effects might also correlate with better PMR-related QoL, it is possible that being on a higher dose of steroids causes greater side effects but better disease control and therefore a person rates their overall QoL as better when asked to answer a single question on this.

Interpretability

The mean and standard deviation of the scores for each domain were calculated to allow consideration of the heterogeneity of the sample. The percentage of participants scoring the highest and lowest end of the scale was calculated to assess for risk of floor and ceiling effects.

The smallest detectable change at individual and group level for each domain was calculated from the SEM of the sample of participants who stayed the same on that domain between the two time points. The SEM used in this calculation was calculated from $SD_{difference} / \sqrt{2}$ as per the Bland and Altman method (i.e. it was $SEM_{consistency}$ and assumes the systematic error to be negligible). Appendix 9.10: Calculating the standard error of the measurement, includes data justifying this assumption.

$SDC_{ind} = 1.96 \times \sqrt{2} \times SEM$

$SDC_{group} = SDC_{ind} / \sqrt{n}$

The MIC for each domain was calculated using two different anchor-based approaches – the mean change method and an ROC method. Due to the high level of variability of the mean change scores in the worsened groups, only an MIC for improvement was calculated.

In the mean change method, the group of participants who answered the anchor questions indicating that they had slightly improved were identified for each domain and the mean change score in these groups was calculated.

For the ROC approach, participants were grouped into those who had improved (anchor question response 1 or 2), those who were stable (anchor question response 3). ROC curves were plotted for improved vs stable for each domain with sensitivity (true positives) on the y axis and 1-specificity (false positives) on the x axis. The coordinate points for these curves were examined to identify the cut-off point where sensitivity – specificity was the least. This was taken as the MIC. The AUC was calculated as a measure of the ability of the questionnaire to distinguish between improved and stable patients.

9.5 Results

9.5.1 Response rate

One hospital site and 30 primary care practices carried out searches and screened potential participants. 559 study packs were sent out in total. The details of response rates are given in Table 9.4.

225 responses were received, and 215 second study packs were sent via Docmail (ten respondents did not consent or did not complete address details to receive the second questionnaires). 194 participants completed both sets of questionnaires.

This gives an overall response rate of:

225/559 = **40%** for the first round

194/215 = **90%** for the second round

Table 9.4: Response rates to first and second study booklets

	Sent	**Responses**	**Response rate**
Booklet 1 – primary care (30 practices)	**425**	**161**	**37.9%**
Booklet 1 – MPFT	**134**	**64**	**47.8%**
Booklet 2*	**215**	**194**	**90.2%**

*41 reminders were sent to those that did not respond to the second booklets within 3 weeks and 26 subsequently replied

One study inclusion criterion was that participants had been diagnosed with PMR in the preceding two years. However, some of the participants indicated in their responses that

they had been diagnosed with PMR more than 2 years ago. This could have occurred for a number of reasons – error in the screening process, mismatch between 'coded' date of PMR diagnosis in the records and patient's recollection of events, coding of PMR as a new diagnosis when it was an episode of relapse or simply a lag time between identification of the participant and completion of the questionnaires. After discussion with my supervisors, I decided to include participants diagnosed up to three years ago. This strikes a balance between optimising participant numbers whilst not including participants who reported they had been diagnosed such a long time ago that they are likely to be off treatment and asymptomatic or very atypical in their disease course.

9.5.2 Demographics of the sample

After removing the 15 respondents who were diagnosed more than 36 months ago or who returned questionnaires that were blank or uninterpretable, there were **210** first booklets for analysis and **179** paired questionnaires for analysis.

Table 9.5 shows the demographics of the study participants.

The demographic distribution of participants in this study is similar to that of the largest UK cohort study of PMR (Muller et al., 2016) which reported a mean age of 72.4 years with 62.2% being female.

Table 9.5: Summary of study participant characteristics

	Mean (SD)	Range	Missing
Age	72.2 years (8.14)	52-90	6
Duration since diagnosis	16.1 months (8.93)	1-36	10
Dose of prednisolone	5.76g (4.3)	0-20	14
Gender	57.1% Female		0

9.5.3 Test-retest reliability

Responses to the anchor questions

For analysis of test-retest reliability, only participants who rated themselves as being unchanged on the domain being evaluated were included.

The frequencies of responses to each anchor question are shown in Table 9.6. Bar charts visually depicting the frequencies can be found in Appendix 9.8: Bar charts of frequencies of responses to the anchor questions.

Table 9.6: Frequencies of responses to each anchor question

Domain	Frequency of anchor response				
	Improved a lot (1)	Improved a little (2)	Stayed the same (3)	Worsened a little (4)	Worsened a lot (5)
Symptoms	33	32	60	34	16
Function	26	28	80	29	14
Emotional well- being	27	19	95	27	8
Steroid side effects	19	16	107	19	5
QoL linked to PMR	25	26	75	40	11

Time interval between questionnaires

The mean time interval between completion of the questionnaires was **18 days** (range 7-45).

I initially specified that the time interval between questionnaires should be 2-6 weeks. However, some participants did not fill in the date when they completed the questionnaire. Early on in recruitment, the questionnaires were logged as soon as they were received by the study team and it was therefore possible to use the date that they were received as an approximation. During the later stages of recruitment, due to the Coronavirus pandemic, there was a variable delay between the questionnaires arriving in the department and them being logged in the study database. In this case, if a participant had not completed the date then the length of the interval between questionnaires was

unknown (although all participants included the test-retest analysis were those that had rated themselves as unchanged on the domain in question between the two time points of questionnaire completion). For this reason, the mean time interval presented here is based on the 102 participants who completed the study before the pandemic restrictions came into force, or who completed the date on both questionnaires.

Intraclass correlation coefficients

The $ICC_{agreement}$ for each scale is presented in Table 9.7. In each case, the result is >0.8 suggesting good reliability.

Standard Error of Measurement

The $SEM_{agreement}$ for each domain is also given in Table 9.7. The calculation of this is explained in more detail in Appendix 9.10: Calculating the standard error of the measurement.

The $SEM_{agreement}$ values for the function, emotional and psychological well-being and steroid side effects scales are fairly consistent though slightly higher for the symptoms scale.

Limits of agreement

Histograms drawn to check for normality of the data are included in Appendix 9.9: Testing for normality of the differences between the measurements for each scale.

Bland and Altman plots for each domain variable are presented in Figure 9.2 and the limits of agreement are given in Table 9.7.

The distribution of points on the Bland and Altman plots is slightly 'funnelled' in each case i.e., the differences between the scores are smaller when the mean of the two measurements is lower. This is likely to be because there is less opportunity for variability of responses when scores are lower (i.e., the response option to choose is clearer to the participant when the condition is better controlled) even in a group who have all judged their overall state to be the same as the previous time they completed the questionnaire. It also highlights that many participants scored at the lower end of the scales, suggesting there may be a floor effect.

Table 9.7: Intraclass correlation, standard error of measurement and limits of agreement for each scale

Scale	**n**	**$ICC_{agreement}$ (95% CI)**	**$SEM_{agreement}$**	**Mean and LoA**
Symptoms	59	0.83 (0.73, 0.90)	11.85	4.22 (-27.88, 36.32)
Function	80	0.85 (0.77, 0.90)	8.44	0.67 (-22.83, 24.16)
Emotional and psychological well-being	95	0.81 (0.73, 0.87)	9.72	1.05 (-25.97, 28.07)
Steroid side effects	100	0.83 (0.76, 0.88)	9.31	-1.94 (-27.59, 23.72)

n = number of participants reporting they had 'stayed the same' on this scale between completing the two questionnaires

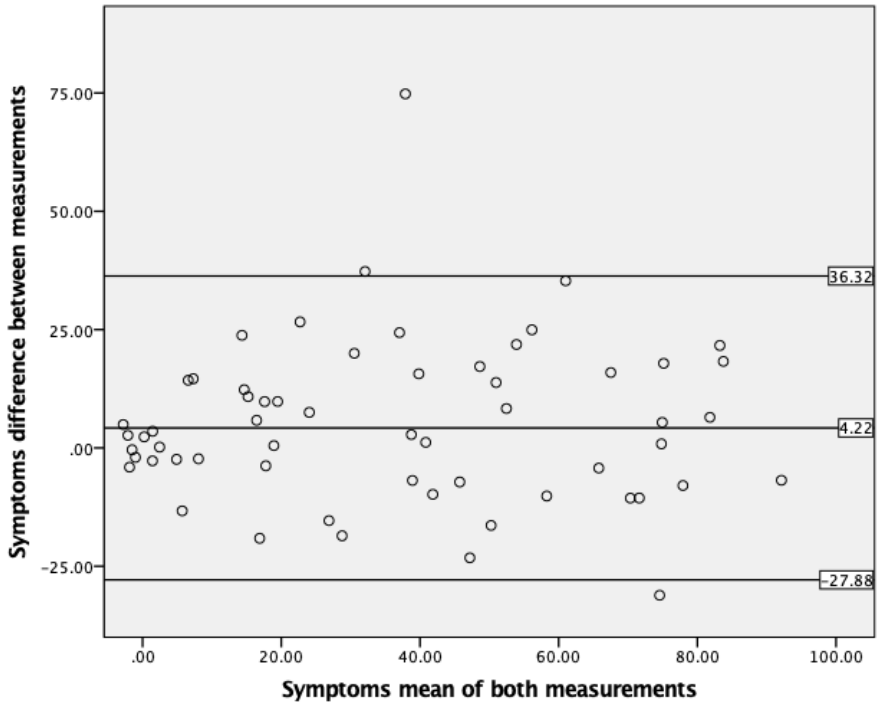
$SEM_{agreement}$ = standard error of the measurement, calculated for agreement

$ICC_{agreement}$ = intraclass correlation coefficient, two-way, calculated for agreement

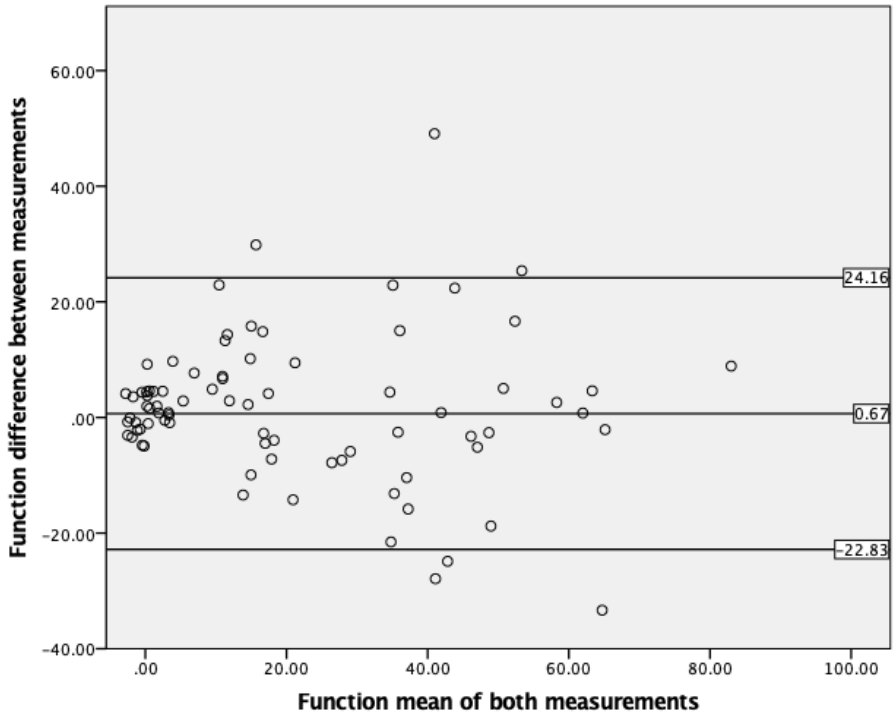
LoA = limits of agreement between which 95% of second values are expected to fall calculated using the Bland and Altman method

Figure 9.2: Bland and Altman plots for each domain

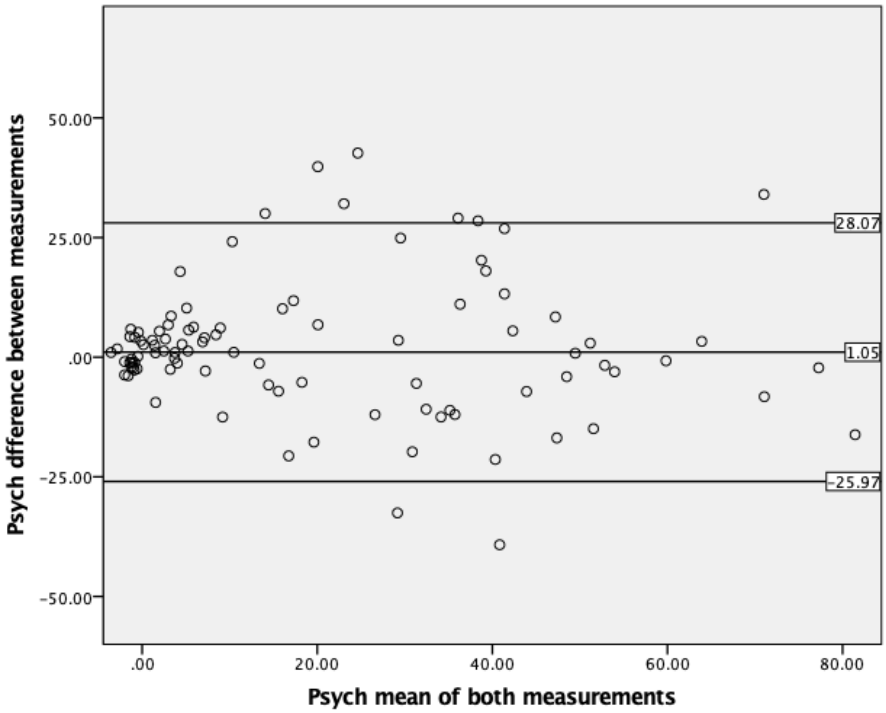
Symptoms



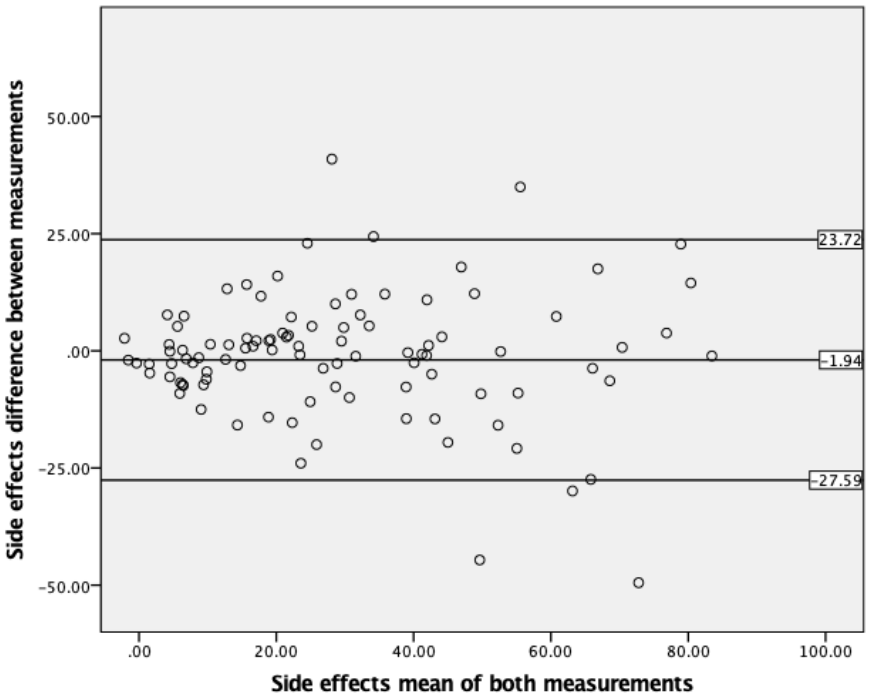
Function



Emotional and psychological well-being



Steroid side effects



9.5.4 Construct validity

Scatter plots for the paired variables tested for each hypothesis are included in Appendix 9.11: Scatter plots for correlation. They confirm a monotonic relationship in each case, satisfying the assumptions of Spearman's correlation coefficient.

Table 9.8 shows the Spearman's r for each pair tested and summarises the outcomes of the hypotheses tested. Overall, only one out of 11 hypotheses was rejected. The PMR-IS therefore meets the pre-specified criteria for good construct validity.

Table 9.8: Results of hypothesis testing for construct validity

Comparator construct	Hypotheses	Results (Spearman correlation)	Interpretation
Symptom severity	The symptoms score of the PMR-IS is moderately to highly negatively correlated (-0.5 to -0.7) with the bodily pain and energy / fatigue scores of the SF-36.	-0.81 with SF-36 bodily pain (n=206) -0.66 with SF-36 energy / fatigue (n=204)	2 of 2 hypotheses met
Physical function	The score from the functional domain on the PMR-IS is strongly positively correlated (r>0.6) with the score on the mHAQ and strongly negatively correlated with the physical functioning, social functioning and role limitation physical scores of the SF-36 (r<-0.6)	0.897 with the mHAQ (n=206) -0.774 with SF-36 physical functioning (n=206) -0.603 with SF-36 social functioning (n=206) -0.549 with SF-36 role limitation physical (n=198)	3 of 4 hypotheses met.

Mental and emotional state	The score from the emotional and psychological well-being domain of the PMR-IS is strongly negatively correlated (r<-0.6) with the emotional well-being, social functioning and role limitation emotional scores of the SF-36.	-0.784 with emotional well-being (n=199) -0.732 with SF-36 social functioning (n=204) -0.610 with SF-36 role limitation emotional (n=193)	3 of 3 hypotheses met
Steroid side effects	The score from the steroid side effects domain of the PMR-IS score correlates negatively (r<-0.2) with the general health scores of the SF-36.	-0.593 with general health (n=185)	Hypothesis met
Symptoms and function – internal relationship	The symptoms score of the PMR-IS correlates positively and moderately strongly (r>0.4) with the functional domain of the PMR-IS.	0.834 with PMR-IS function (n=206)	Hypothesis met

9.5.5 Responsiveness

For each domain, as there were fewer than 35 participants in each of the worsened a little / a lot and improved a little / a lot groups, the groups were merged into a combined 'worsened' or 'improved' category.

The mean change scores for each of the groups defined by the response to the domain specific anchor questions are shown in the Table 9.9 and Figure 9.3. The mean change scores for groups defined by their answer to the overall PMR-QoL anchor question are show in Table 9.10 and Figure 9.4.

For each of the domains, whether the domain specific anchor or the QoL anchor is considered, the distinction is greater between the 'improved' and 'stayed the same' groups than between the 'stayed the same' and 'worsened' groups. The variability (shown by the standard error bars on the charts) is much greater in the 'worsened' groups.

A summary of the results of the hypothesis testing is given in Table 9.11. Overall, four out of the five hypotheses were satisfied. However, for the steroid side effects domain, even though the hypothesis concerning the trend in expected changes was satisfied, the scores improved even for the 'worsened' group. The small magnitude of change with the high level of variability in the worsened group suggest that this finding should be interpreted with caution.

Table 9.9: Mean change scores for each domain for groups defined by participants' response to the domain-specific anchor question

Domain-specific anchor response	Symptoms		Function		Psychological and emotional		Steroid side effects	
	n	Mean (SD) change score	n	Mean (SD) change score	n	Mean (SD) change score	n	Mean (SD) change score
Improved	65	-7.57 (20.57)	52	-4.60 (19.11)	44	-7.67 (15.18)	29	-3.89 (10.66)
Stayed the same	59	4.22 (16.38)	80	0.67 (11.99)	95	1.05 (13.79)	100	-1.93 (13.09)
Worsened	50	7.06 (11.78)	43	0.88 (15.28)	35	-0.89 (19.18)	23	-1.16 (15.18)

Table 9.10: Mean change scores for each domain for groups defined by participants' response to the anchor question on overall PMR-QoL

PMR-QoL anchor response	Symptoms		Function		Psychological and emotional well-being	
	n	Mean (SD) change score	n	Mean (SD) change score	n	Mean (SD) change score
Improved	51	-6.77 (19.80)	49	-5.20 (17.91)	49	-5.61 (15.64)
Stayed the same	74	2.36 (16.48)	75	0.81 (11.81)	75	-0.08 (11.74)
Worsened	51	4.69 (17.30)	51	0.93 (16.62)	50	0.25 (19.96)

Figure 9.3: Bar chart showing mean change scores for each domain for groups defined by participants' response to the domain-specific anchor question

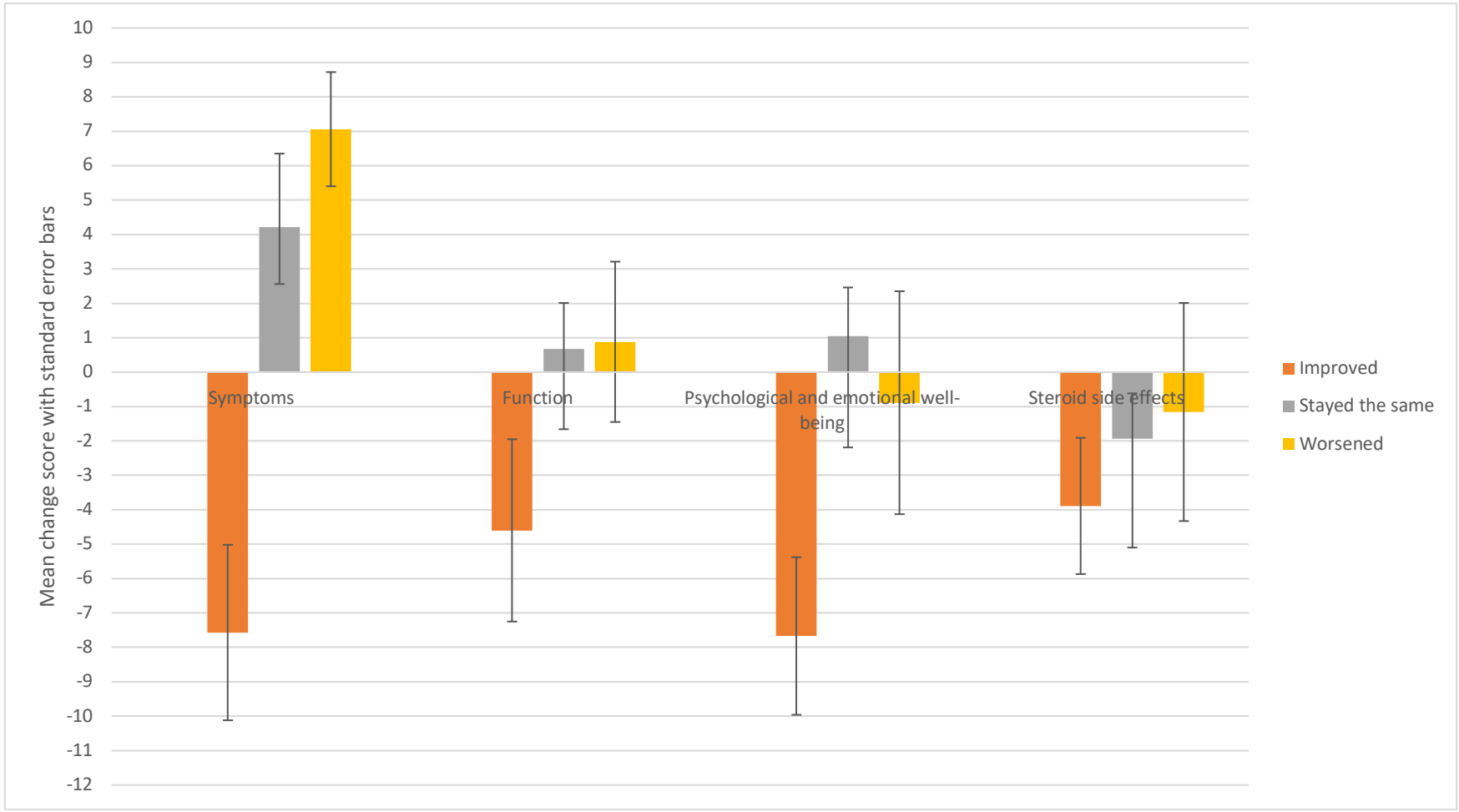


Figure 9.4: Bar chart showing mean change scores per domain for groups defined by participants' response to the PMR-QoL anchor question

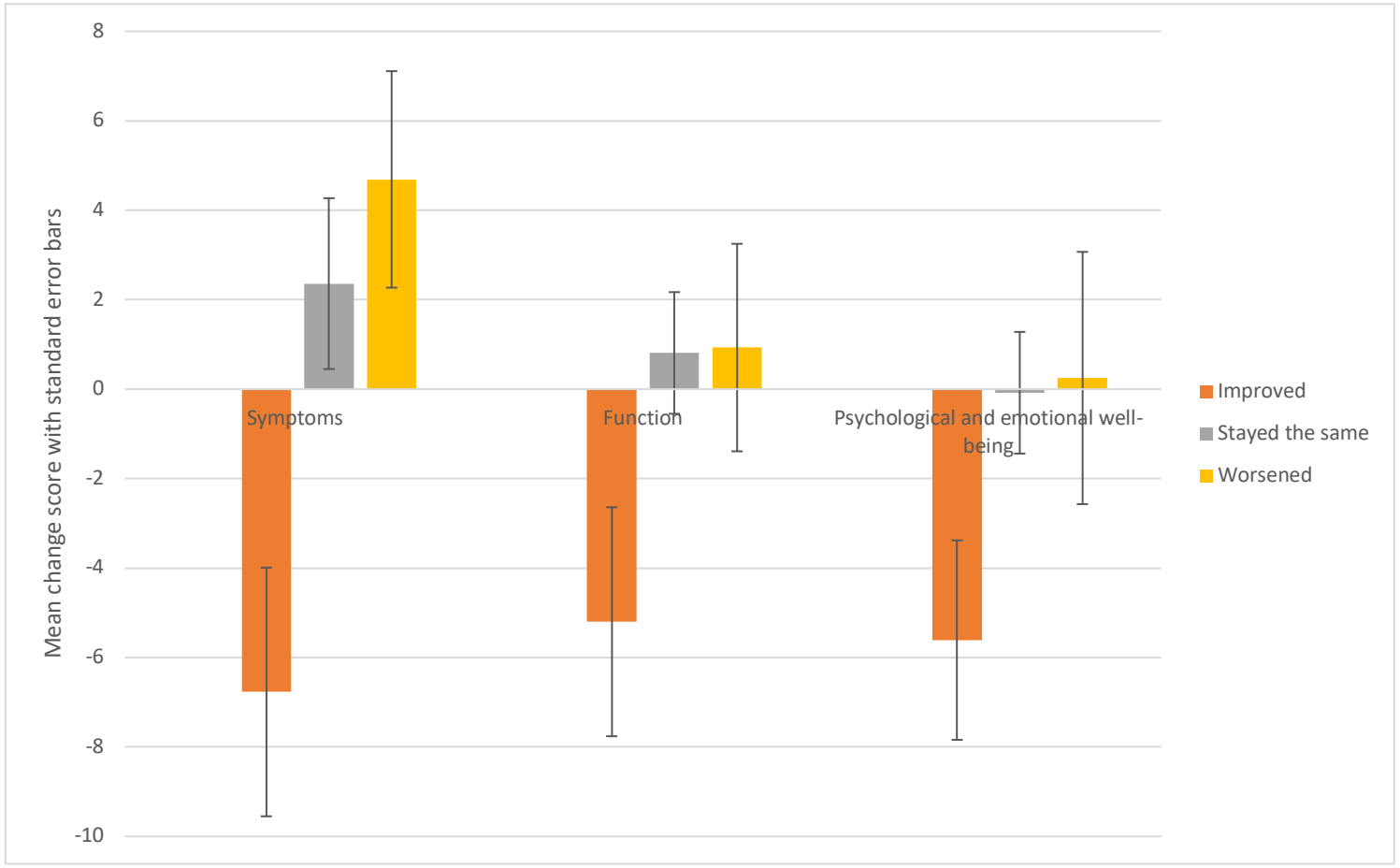


Table 9.11: Results of hypothesis testing for responsiveness

Hypothesis	Satisfied?	Comments
There will be a qualitative trend in mean change scores on the PMR-IS symptoms domain from those reporting that their symptoms have improved a lot to those reporting that their symptoms have worsened a lot.	**Yes**	Relatively high mean change score for the group that rated themselves as staying the same. Proportionately smaller change between the 'worsened' and 'stayed the same' groups than between the 'improved' and 'stayed the same' groups.
There will be a trend in mean change scores on the PMR-IS function domain from those reporting that their functional ability linked to their PMR has improved a lot to those who report that it has worsened a lot.	**Yes**	Very small difference between the 'worsened' and 'stayed the same' groups.
There will be a trend in mean change scores on the PMR-IS emotional and psychological well-being domain from those reporting that their psychological well-being linked to their PMR has improved a lot to those who report that it has worsened a lot.	**No**	The 'worsened' group showed a small improvement in their mean PMR-IS score.
There will be a trend in mean change scores on the PMR-IS steroid side effects domain from those reporting that their steroid side effects have improved a lot to those who report that they have worsened a lot.	**Yes**	The trend between the groups is in the expected direction but all groups showed improvement in their PMR-IS score.

There will be a trend in mean change scores on the symptoms, function and emotional and psychological well-being domains between those that report their overall PMR-related quality of life (PMR-QoL) as improved, the same or worsened.	**Yes**	The difference in mean change score for those in the 'stayed the same' as compared to the 'worsened' group is small, particularly for the function and emotional and psychological well-being domains.

9.5.6 Interpretability

Risk of floor and ceiling effects was assessed by considering the sample distribution and the percentage of participants scoring the highest and lowest possible scores at baseline (Table 9.12). The function and emotional and psychological well-being domains both have risk of floor effects as more than 15% of participants scored the lowest possible score at baseline, thus making it impossible to detect any improvement in these individuals.

Table 9.12: Sample distribution and assessment for risk of floor and ceiling effects

Domain	n	Mean score (SD)	Range	% scoring 0	% scoring 100
Symptoms	209	46.1 (29.3)	0-100	9.5	1.0
Function	207	30.3 (27.2)	0-100	21.0	1.0
Emotional and psychological well-being	205	28.6 (27.2)	0-100	27.1	1.0
Steroid side effects	190	33.2 (24.6)	0-100	6.2	0.5

The smallest detectable change (SDC) at group and individual level and the minimally important change for improvement (MIC), calculated by two different methods, are shown in Table 9.13.

The description and plots are included in Appendix 9.12: Anchor based ROC method to calculate the MIC improvement for each domain.

Table 9.13: Smallest detectable change and minimally important change for each domain

Domain	Smallest detectable change			MIC improvement (mean change method)		MIC improvement (anchor-based ROC method)		
	n	Individual level	Group level	n	Mean change score (SD)	n	MIC	AUC
Symptoms	59	32.10	4.18	32	-8.91 (19.4)	124	-0.63	0.66
Function	80	23.50	2.63	27	-2.73 (16.83)	132	-0.69	0.58
Psychological and emotional well-being	95	27.02	2.77	19	-10.20 (12.88)	139	-3.13	0.68
Steroid side effects	100	25.66	2.57	15	-2.96 (8.45)	129	-1.30	0.53

*the group used for this calculation were those that rated themselves as 'improved a little' on the anchor question

MIC = minimally important change

AUC = area under the curve

At group level, for each domain, the MIC for improvement (mean change method) is greater than the SDC indicating that the instrument is able to capture meaningful change in this direction. However, the numbers of participants in the groups used in these calculations were low.

The MICs for improvement calculated using the ROC method are smaller than those calculated using the mean change method, and only the one for the psychological and emotional well-being domain exceeds the SDC. The AUC result is less that 0.7 in each case suggesting that the tool is not able to clearly discriminate between those that have improved and those that have remained stable.

At individual level, the SDC in each domain is much higher (as expected) and exceeds the MIC for improvement and deterioration in all domains.

9.6 Discussion

This study is the first evaluation of the measurement properties of the newly developed PMR-IS. The results demonstrate satisfactory achievements of COSMIN standards for some measurement properties but not for all.

In line with recommendations for studies of this type (Mokkink et al., 2019), it was carried out in the population of interest (people with PMR) and the sample achieved was representative of this population. However, the heterogeneity of the study population, in terms of time since diagnosis, disease course and dose of steroids, adds complexity to the interpretation of the data, particularly when considering responsiveness and interpretability, as discussed further below.

Test-retest reliability

The methods used in evaluation of test-retest reliability meet all of the criteria specified in the COSMIN checklist for designing studies of measurement properties (Mokkink et al., 2019). There were difficulties in determining the time interval between completion of the two sets of questionnaires for some participants, linked to the COVID-19 pandemic as discussed in Section 9.5.3, but the mean time interval calculated on the basis of the available data was satisfactory at 18 days (de Vet, Terwee, et al., 2011e). Given that the criteria for the assumption of stability was a self-rating anchor question rather than purely relying on the time interval, the uncertainty regarding the time interval for some individuals should not affect the results. The group size was adequate for each domain (n = 60-107) and the statistical methods used were those recommended by the COSMIN group. Participants were only included in the analysis if they had a score for the domain in question at both time points so missing data was not a concern.

Each of the domains of the PMR-IS demonstrated good reliability in this study population ($ICC_{agreement}$>0.8). However, the measurement error indicated by the $SEM_{agreement}$ and LoA was fairly large in each case – the SEM being about 10% of the range of the scale and the LoA around +/-25%. There are no set standards for these parameters, with judgement of whether they are satisfactory being guided by clinical interpretation. The context for understanding the implications of the size of the measurement error comes from considering the expected change in score in response to change in disease activity e.g. during a flare or in response to treatment, and this is discussed further below. Due to the limited evidence on psychometric properties of other measures in PMR, there is little against which to compare the findings of this study. However, Matteson et al. (2012) calculated the limits of agreement using the Bland and Altman method for the mHAQ and

for the SF-36 physical component scale and the SF-36 mental component scale in PMR finding them to be 25%, 9.2% and 16.7% of the scales respectively. The limitations of this study have been outlined in Chapter 5 (Section 5.4.2).

Construct validity

Construct validity of the PMR-IS was evaluated according to the COSMIN standards (Mokkink et al., 2019), using specific hypotheses about expected relationships between it and other outcome measurement instruments. The population used was a good representation of that in which the tool will be used and the sample size was adequate. In an ideal situation, a new measurement instrument will be tested against an existing 'gold standard'. If a gold standard is not available, as is the case with PMR, a new instrument has to be tested against selected comparator instruments. These comparator instruments should ideally measure the same construct as the new instrument and should have undergone adequate validation themselves. However, in PMR, as I have demonstrated in my systematic review, there are no adequately validated instruments for any of the domains measured by the PMR-IS. I therefore used a generic health measurement instrument, the SF-36 (Ware & Sherbourne, 1992), which has been shown to be a good measure of perceived health in the general population (Brazier et al., 1992) and subsequently used and validated in many different conditions (Ware, 2000). Alongside this, I used the mHAQ (Pincus et al., 1983), which has good evidence of its psychometric properties in other inflammatory musculoskeletal conditions (Maska et al., 2011) and some evidence supporting its use as a measure of function in PMR (Matteson et al., 2012; McCarthy et al., 2014).

Eleven hypotheses about expected relationships between specific domains were set and ten of these were satisfied.

The PMR-IS symptoms domain covers pain, stiffness, weakness and fatigue with questions about severity and duration of each. This correlated strongly, as predicted, with the pain and energy / fatigue scores of the SF-36 and also with the function score on the PMR-IS.

The function score on the PMR-IS correlated very strongly with the mHAQ. There is some overlap of questions between the two measures (both ask about getting in and out of bed, getting in and out of a car and being able to fully wash yourself) but the PMR-IS asks distinct questions specifically linked to functional impairment due to the PMR disease process (asking about turning over in bed and reaching above your head). It was anticipated therefore that the two measures would correlate strongly but not perfectly, and this was found to be the case.

The PMR-IS function score also correlated strongly with the physical functioning score of the SF-36. The wording of the question on the physical function stem question on the SF-36 asks about whether the persons 'health' limits the listed activities and this may have caused some difference in responses to questions that focused on specific limitations due to PMR. The questions also have more of a focus on exercise tolerance than on discrete physical activities. For both of these reasons, the correlation would not be expected to be as strong as with the mHAQ and this was indeed the case.

I had anticipated that the PMR-IS function score would also correlate strongly with the SF-36 social functioning score and the SF-36 role limitation physical score. These questions assess the extent to which a person's health is interfering with their work, regular daily activities and social activities. The correlation here was found to be less strong than with the mHAQ, which on reflection could be due to the demographic of the people affected

by PMR. It may also have been influenced by the COVID-19 pandemic, which restricted everyone's daily activities. There was still a reasonably high correlation in the expected direction with both of these scales, but the magnitude of the correlation did not meet the hypothesis in the case of the role limitation physical scale.

The emotional and psychological well-being score of the PMR-IS correlated strongly with the relevant three scales of the SF-36, meeting the specified hypotheses. The correlation was strongest with the SF-36 emotional well-being scale and least strong with the SF-36 role limitation emotional scale. The explanation for this is likely to be the same as discussed for the corresponding physical scales.

When it came to assessing the construct validity of the steroid side effects domain, there were few options for comparators. A hypothesis was set about the relationship with the general health scale of the SF-36 as it seemed reasonable that the multitude of possible adverse effects from steroids would impact in a cumulative way on self-rated general health. The results showed a moderate correlation between the PMR-IS steroid side effects score and the SF-36 general health score and this hypothesis was met.

Overall, the satisfaction of ten out of eleven hypotheses (91%), in a study meeting the methodological criteria, demonstrates that the PMR-IS has good construct validity in this population.

Responsiveness

The responsiveness of the PMR-IS was assessed by testing hypotheses about expected mean change scores in groups defined by their response to an anchor question.

The anchor question acts as a proxy for a gold standard measure of the construct in question as in PMR, no such gold standard exists. However, it is important to

acknowledge that the anchor question itself is not validated and different wording of the stem and response options could elicit different results.

The time interval between the two data collection points in this study was short, primarily because the same study was being used to collect data for test-retest reliability (where a short time interval is needed). Using the anchor method to define sub-groups ensured that the participants included in the responsiveness analysis were ones in whom a change had occurred, and the short time interval is therefore not problematic in itself. However, the numbers of participants in these 'changed' groups were small and this is a significant limitation. If a longer time interval were used, more participants would be expected to experience a change in their condition and the mean change scores would be more reliable.

Hypotheses about expected changes in mean scores in groups defined by a domain specific anchor as well those relating to an anchor question about overall PMR-QoL were tested.

Whilst four out of five hypotheses were met, this encouraging result belies underlying complexity in the data.

For each domain, the ability to detect improvement is reasonable - the mean score on each domain decreased in those that rated themselves as having improved, either on that domain or in terms of their overall PMR-related QoL, and this change was greater than in those that rated themselves as having stayed the same. The ability to detect deterioration however, was less good. For the symptoms domain, the mean change score in those that rated their symptoms as 'worse' was similar in magnitude to the change in score for those that had improved, but in the opposite direction i.e. it responded as expected. For the other three domains, the mean change score was very small and in the

case of the emotional and psychological well-being and steroid side effects domains, the mean score actually reduced even in those who reported that they had worsened. Similarly, in those that rated their overall PMR-related quality of life as worse, the PMR-IS symptoms domain score changed as expected but there was minimal change on the function or emotional and psychological well-being domains. In addition to the magnitude of change being small for the worsened groups in these domains, the variability is high and the range of standard error crosses zero.

Whilst these results may in part be influenced by the low number of participants who reported a change in their condition, they might also be reflecting something about PMR itself.

In general, once someone is on treatment for PMR, the expected direction of change is that the condition improves with time. A person may experience relapses, but these will not be well captured in a cross-sectional study of a heterogenous sample of people at all different stages of the disease course. Participants rating themselves as 'worse' in this study may have been relapsing but could have been at the start or end of this process, and thus not very different from baseline, or at could have been at their very worst. The proportion of participants who reported they had worsened or improved on the anchor question for any of the domains are actually fairly even but the degree of this change is different (Table 9.6). In the improved group approximately as many have improved a lot as improved a little whereas in the worsened group, approximately three times as many had only worsened a little as worsened a lot. The numbers in the 'worsened a lot' group were therefore particularly small compared to the other groups. In order to assess the ability of the PMR-IS to detect relapse, a study would need to be designed to specifically capture adequate numbers of participants in this state.

The other issue that could affect the results of the responsiveness analysis is the validity of the measure compared to the anchor question. If the participant responds to an anchor question about function and then answers a series of question about function that do not accurately capture their experience of this construct, the change in these two scores would not have a predictable relationship. However, if this were the case, the relationship would be affected for both improvement and worsening and this is not the situation here. It is interesting however, that the mean change scores matched most closely with what was expected when an anchor question about overall PMR related QoL was used. It may be that participants found it easier to respond to this question than questions about individual symptoms or effects of their PMR.

In summary therefore, this study provides some evidence that the PMR-IS is a responsive measure in detecting improvement in PMR but further studies are needed to determine whether it is able to detect relapse.

Interpretability

Whilst interpretability is not a measurement property of an instrument, this study was used to calculate some parameters of the PMR-IS that facilitate interpretation of the meaning of scores on the instrument.

The distribution of the scores shows that there is a risk of floor effects for the function and emotional and psychological well-being domains. This means that if people with low scores for these domains improve further, this will not be captured by the instrument. This has implications for the mean change scores calculated in this study in that it may have made the mean change score for improvement smaller than it might have been if the floor effect was not present. However, in the context of using this tool to evaluate

the impact of PMR on a person's life (whether in a research study or in clinical practice), this floor effect might not be a problem. Once the impact on function or emotional and psychological well-being reaches a certain low level, it may be reasonable to deem the condition 'controlled' and improvement beyond this may not necessitate any change in management. Whether this is true for the level of floor effect of these domains needs further study.

The SDC and MIC are not fixed properties of a measurement instrument and will vary in different populations and with different methodological approaches. For this reason, the results for these parameters from this study are only a first estimate and need corroboration in larger studies, with participants at different stages of their disease course and with a variety of approaches.

I had planned to calculate an MIC for improvement and for deterioration but the degree of variability in the change scores for the 'worsened' groups precluded this. Studies reporting responsiveness and MIC of other musculoskeletal outcome measures (Haywood et al., 2010; Jordan et al., 2006; Spies-Dorgelo et al., 2006) solely report MIC for improvement, presumably because in many trials, this is what is necessary to detect. In PMR the OMERACT Special Interest Group are working to understand how best to define and measure relapse and future studies to look at the utility of the PMR-IS for this need consideration.

The two different approaches used to estimate MIC in this data yielded quite different sets of results. The groups of participants used in the two MIC calculations in were different – the mean change method was based on data from participants who rated themselves as slightly improved or slightly worsened on the anchor question whereas the ROC method used data from all participants rating themselves as improved or

deteriorated. This meant that the numbers of participants included in the mean change method calculation were small, reducing confidence in the results. Another limiting factor for the mean change method in this study was that the anchor question asked about 'slight change' rather than 'important' change. To truly determine the minimal 'clinically important' change, a question specifically targeted to this should be used.

The MIC calculated using the anchor-based ROC method, was smaller for each domain than that calculated using the mean change method. The values for the symptoms and function domains in particular are very low suggesting small changes in score could be significant. However, for such small changes to be detectable in a study, very large numbers of participants would be needed to bring the SDC down sufficiently. In their study of two different measurements of hand function, Spies-Dorgelo et al., (2006) note that the MIC calculated using the anchor-based ROC method tends to be lower than that found using other approaches and comment that the rationale for choosing the cut-off point will be determined by the context in which the questionnaire is being used.

The results here suggest that overall, at group level, the PMR-IS may have the ability to detect minimally important improvements in the assessed domains in people with PMR but further studies are needed before a definitive conclusion can be drawn.

At individual level, instruments need to be able to detect change more sensitively and the SDC for each domain is much higher. Further studies are therefore required to determine the usefulness of the PMR-IS in guiding treatment decisions for an individual in clinical practice.

9.7 Conclusions

This chapter describes the first evaluation of the psychometric properties of the PMR-IS. It shows this new, PMR-specific PROM to be a reliable and valid measure of the impact of

the condition, in a population representative of the overall UK PMR population. Scores on the PROM respond as expected to improvement in the condition and there is some evidence to suggest it is able to distinguish between improved and stable states at group level. However, further studies are needed to determine its ability to detect relapse and its clinical utility. Studies of people at different stages of their PMR disease course, rather than the heterogenous sample used in this study, will be important to evaluate responsiveness and interpretability more precisely.

Chapter 10: Discussion

10.1 Introduction

I began this thesis with an overview of polymyalgia rheumatica, highlighting the limited evidence base for this condition and the many unanswered questions that remain about its aetiology, pathogenesis, clinical course and management. I then discussed patient reported outcome measures and their increasing recognition as key tools in facilitating patient-centred research and clinical practice.

My systematic review confirmed the paucity of validated outcome measures for the condition and the primary research studies that followed enabled the development and evaluation of a new PROM specifically for PMR, the PMR-IS.

I have discussed the strengths, weaknesses and implications of each stage of the research within the relevant chapters. In this chapter, I will therefore take a holistic view of the strengths and weaknesses of the overall process and the instrument that has been developed and conclude by considering the further steps that are needed to ensure this work results in tangible improvements to care for patients with PMR.

10.2 Novelty and importance of the work in this thesis

This research arose because of a perceived gap in our ability, as clinicians and researchers, to reliably measure the impact of PMR on patients' lives. My systematic review was the first to formally confirm this, identifying instruments currently used to measure outcomes in research studies of PMR and examining the existing evidence for the measurement properties of these instruments. The finding that none of the

commonly used instruments have good evidence to support their use in PMR was important in highlighting the work that needed to be done to improve the assessment of the condition and therefore improve the quality of evidence that ultimately informs care. The research described in Chapters 8 and 9 is original research designed and executed specifically to develop and evaluate the first disease-specific patient reported outcome measure for PMR.

During the time that I have been carrying out the work for this PhD, the PMR-SIG of OMERACT have continued to work towards agreeing a core outcome measurement set for research studies into PMR. At each stage of this process, the lack of evidence for current instruments and in particular, the lack of a patient-reported outcome measure, has been emphasised (Owen, Yates, et al., 2019; Yates et al., 2020). Studies by the group are ongoing to try to gather supporting evidence for current instruments including pain and stiffness VAS / NRS and the mHAQ (and some of my findings will contribute to this evidence gathering process) but the absence of a comprehensive patient reported outcome measure for PMR persists. The work reported in this thesis therefore remains current and important.

10.3 Reflection on the need for a PMR specific outcome measure

The impetus to develop an instrument to measure the impact of PMR on a person's life initially arose from clinical experience and was subsequently affirmed and shaped by my qualitative work on patient experiences of PMR. Mine and others' qualitative work, plus ongoing work with patient partners in the OMERACT PMR-SIG, have consistently emphasised that PMR has much broader effects than simply 'pain and stiffness', that inflammatory markers are not reliable objective measures of disease activity and that

treatment decisions need to take into account the full range of symptoms / effects and balance these with the adverse effects of steroid treatment. We do therefore need a better way than we currently have of assessing the impact of PMR in clinical practice and to facilitate further research to improve management of the condition.

Clinical research studies require precisely defined outcomes and instruments used to measure these need to be able to capture disease-specific changes. There is a tension however, between the time and resource needed to develop disease-specific instruments, which capture more nuanced outcomes, and the cheaper and easier option of using established measures from other conditions, which may sacrifice precision. Multiple PROMs for musculoskeletal conditions already exist and a challenge could therefore be made that developing a new measure rather than adapting an existing one was not necessary. However, PMR is an unusual condition which, along with GCA, sits slightly separately to other rheumatological conditions such as the inflammatory arthritides, autoimmune rheumatic diseases and other vasculitides. The distribution of the pathology of PMR, predominantly affecting proximal muscles and joints, means that the functional effects are different to those of other arthritides, which commonly affect peripheral joints. The pattern of functional limitation and the age group it affects, also means that the psychological effects of PMR differ from other rheumatological conditions – an older adult experiencing difficulty turning over in bed, getting in and out of a bath or getting up off the floor is likely to experience to fear of falls and a sense of vulnerability that a younger person with difficulty with hand function from peripheral arthritis would not.

The disease course of PMR is also unusual. The subacute onset of usually quite severe symptoms, followed by a prolonged duration of much milder symptoms, gradually

improving to a point of complete resolution over a number of years, but with an unpredictable pattern of flares / relapses, is different to many other rheumatological conditions. As there is not the associated risk of joint damage that there is with other inflammatory arthritides, nor the risk of sight loss or other end organ damage that there can be with GCA or other vasculitidies, there is not the same drive to treat hard and early and decisions about treatment can be more holistic.

The PROM that I have developed is a composite measure using some standard assessment tools, such as numeric rating scales for symptom severity and categories for symptom duration, but also containing unique scales for function, emotional and psychological well-being and steroid side effects. I believe that the unique features of PMR mean that developing this disease-specific PROM was warranted and that it has the potential to allow more comprehensive and subtle assessment of the impact of the condition than existing measures.

10.4 Strengths and weakness of the development process

The strengths and weaknesses of the methodological approach taken to each stage of development of the PMR-IS have been discussed within each chapter.

The detailed documentation of the overall process from start to finish, with each version of the PROM included for transparency, allows for scrutiny of the process and enables future users to fully understand the benefits and limitations of the instrument.

As outlined in Chapter 2, there are several organisations committed to improving the use of PROMs in research (COSMIN, CPROR, ISOQOL) and several different guidelines on evaluating PROMs with a view to selecting high quality PROMs for research studies (Mokkink et al., 2019; OMERACT, 2019; Patrick et al., 2007). There is not however, a

definitive guideline on how to develop a new PROM and there are a wide variety of methods reported in the literature. I used several sources of evaluative guidance to design my overall approach and then focussed on COSMIN terminology and methodology for the detail of each stage.

10.4.1 Early development work and pilot testing

The initial development work for the PMR-IS was carried out prior to starting my PhD. The qualitative study of patient experiences of PMR stands alone as an important piece of work documenting the patient voice in this condition as well as informing the development of the conceptual framework and long-list of items for the future PROM development work. It was published in a high impact-factor journal demonstrating the rigour of the methods and analysis. The subsequent pilot testing study was carried out in a separate group of patients with PMR using a validated tool for assessing the face validity, feasibility and utility of healthcare questionnaires (the QQ-10 (Moores et al., 2012)) but it was a relatively small sample and was not as rigorous a method as the alternative approach to pilot testing, cognitive interviewing.

Whilst the pilot testing stage comes after the initial development of the items in both development models depicted in Figures 6.1 and 6.2, this process arguably needs to be repeated at least once during the development of a PROM. At this early stage, phrasing of individual items can be refined and an initial assessment of the relevance and comprehensiveness of the items can be made. However, given that significant changes are likely to be made through the subsequent field-testing process as items are reduced and scales formed, it seems appropriate that a further assessment of comprehensibility, relevance and comprehensiveness is made once the final instrument has been developed.

I sought multi-professional opinions on the PROM after the field-testing stage and several changes were made, but I did not repeat a formal face validity, comprehensiveness and comprehensibility assessment with patients. Since I carried out my qualitative study, pilot testing and field-testing work, the COSMIN Risk of Bias checklist (developed for assessing the methodological quality of single studies included in the systematic reviews of PROMs) has been published (Mokkink et al., 2018). This includes standards for the pilot testing stage that the PMR-IS would not easily satisfy (including testing all the items in their final form in a group of at least 30 participants in a survey study or using qualitative interview methods) and this is therefore something I need to address in future work.

10.4.2 Field testing

The broad methodological approach taken to field testing was based on guidance from the COSMIN group and included application of defined statistical processes (factor analysis and Rasch analysis). The rationale for decisions made at each step and the implications of the findings have been discussed in detail within Chapter 8.

The strong theoretical understanding of the construct and its likely component domains arising from the qualitative work meant that it was clear that the overall PROM would contain both formative and reflective elements. The detailed study of the symptoms and steroid side effects domains, analysing the distribution of item responses and missing data to improve item selection and response categories, in addition to the use of standardised statistical processes to develop the reflective scales, strengthens the overall PROM.

With hindsight, it might have been reasonable to solely apply Rasch analysis rather than beginning with Classical Test Theory methods. However, had factor analysis resulted in clearly defined unidimensional scales with good internal consistency, it would have been possible to move on to psychometric evaluation at that point and this would have been a quicker and simpler approach. It was only due to the uncertainty that remained after factor analysis that it was apparent further analysis was needed and Rasch analysis was therefore employed. As discussed extensively in Chapter 8, the benefits of Rasch analysis in this study were in providing a more powerful study of item functioning, testing for differential item functioning and allowing assessment of unidimensionality.

Whilst the COSMIN risk of bias checklist was not available at the time I was planning my field testing study, if the criteria were applied to my methods now, the category of 'adequate' or 'very good' would be reached for each criterion for the assessment of structural validity and internal consistency.

10.4.3 Evaluation of the measurement properties

Once an instrument has been developed, it needs to be evaluated in the population in which it will be used. This is not a one-off assessment; it is a process of gathering evidence to support or refute the reliability, validity and responsiveness of the instrument in defined circumstances. In addition to these psychometric properties, consideration needs to be given to the interpretability of the scores in the population of interest.

The study that I have carried out to evaluate the measurement properties of the PMR-IS is an initial step in this process of generating evidence to support its use. As previously discussed, there are many different methodological approaches described in the literature for each measurement property and the choices made at each stage have been

justified in Chapter 9. The study was designed with the COSMIN study design checklist (Mokkink et al., 2019) in mind, to give the final tool the best chance of meeting the criteria against which it is likely to be assessed.

One of the challenges faced was in trying to evaluate test-retest reliability, construct validity and responsiveness within one study. Both test-retest reliability and responsiveness assessment rely on data from completion of a PROM at two time points but the requirements of each vary. Test-retest reliability needs to be assessed in a group of patients who are stable in the construct being measured over the time interval between the two measurements and this typically necessitates a short time interval, although long enough to reduce any memory effect. For responsiveness testing, the participants need to have either improved or deteriorated between the two measurements and often, therefore the two measurements are months apart. I ratified the different requirements of these two assessments by using anchor questions to identify people who had stayed stable and those who had changed, within a maximum of 6 weeks between the two data collections. This was a pragmatic approach in terms of time and resources but meant that the numbers of participants for the responsiveness analysis ended up being low, which was a significant limitation.

The attempt to use one study to evaluate multiple measurement properties was also a constraint when considering SDC and MIC. The small size of the 'changed' groups had an impact here too, as did the fact that the anchor question was not written to ask specifically about the degree of change that was important to the participant. Future studies to address responsiveness and interpretability in defined populations are needed.

10.5 Strengths and weakness of the final PROM

The PMR-IS has been developed with input from patients with the condition at every stage – people with PMR were involved in planning and designing the research as well as being research participants, ensuring that the data used to develop and refine the tool was truly relevant to PMR. The inclusion of weakness and fatigue in the symptoms domain and the inclusion of a psychological well-being domain are testament to this, as these were included as a direct result of qualitative and survey studies and discussion with patient partners through the OMERACT process.

The rigorous development process enabled the creation of short, unidimensional scales for function and emotional and psychological well-being that are specific to the effects of PMR. The inclusion of a steroid side effects domain allows the balance of treatment benefits versus harms to be considered within one assessment and the scoring system is straightforward and will be easy to use in research or clinical practice.

The initial evaluation study shows that the PMR-IS has good construct validity and test-retest reliability, with parameters meeting the COSMIN standards for these measurement properties (Terwee et al., 2007). However, the measurement error from the reliability analysis is quite high at approximately 10% of the scale for each domain. This means that the SDC at individual level is high although at group level, it is reasonable at between 2-4% for each domain. The ability of the PROM to detect improvement at group level looks reasonable but it is not possible to be sure that it can detect worsening in the condition / relapse.

One of the main limitations of the PMR-IS is its floor effects. More than 10% of participants in the evaluation study scored at the lowest level in the function and psychological domains (i.e., were functioning above the threshold at which the PROM

could further distinguish) and there was sparseness of items at the lower end of the scale evident in the person-item threshold distribution for these same domains in the field testing study. However, this same limitation has been found for pain and stiffness VAS / NRS and for the HAQ and mHAQ in PMR (Owen, Yates, et al., 2019) and is to be expected given the nature of the clinical course of the condition. It may be that in a trial of a new treatment / novel treatment regimen, this does not cause significant difficulty as once the participant is scoring within the 'floor effect' margins, the condition might reasonably be considered to be 'under control' and further differentiation may not be needed.

One way to combat the current floor effect would be to develop further items spanning the lower end of the 'ability' scale. However, this would increase burden on those completing the PROM and may put people off as for some, this could mean answering many questions that are not relevant to them. In the future, the use of item banks and computer adaptive testing may allow targeted questions but the technology to do this is not currently widely available.

10.6 The PMR-IS as an outcome measure for use in clinical practice

Whilst the aims of this thesis centred on developing an outcome measure for PMR for use in clinical research, ultimately this tool could also be useful in clinical practice.

PROMs have the potential to improve patient-centredness of consultations through enabling patients to report symptoms that they may not otherwise and empowering patients to be more involved in decisions about their management and engage with supported self-management. PROMs can also facilitate remote monitoring of long-term conditions and could be a really useful tool in ensuring that clinical appointments are scheduled at times when a review is needed, rather than being offered at a set time

interval. They could also help determine whether a telephone or face to face appointment is needed. This is particularly relevant currently, in the time of the COVID pandemic, and in the context of the increasingly urgent importance of improving the sustainability of healthcare and reducing its environmental impact.

However, PROMs that are used to guide treatment decisions at an individual level in clinical practice need to have greater sensitivity to change with smaller measurement error. In the case of the PMR-IS, this requires further development work and subsequent evaluation for this specific context.

10.7 Conclusions

This thesis had three overall aims – 1) to establish how PMR is currently assessed in clinical research and summarise the evidence that supports the use of existing outcome measures, 2) to develop a patient-reported outcome measure that assesses the impact of PMR on a person's life, and 3) to evaluate this new outcome measure to establish its suitability for use in clinical research. I feel that these aims have been met through the work presented in the preceding chapters and that the tool that has been developed, the PMR-IS, has the potential to bridge the gap between patient and clinician perspectives on the condition and ensure that future PMR research measures what matters to patients.

There are aspects of the PMR-IS which require further work, particularly with regards its responsiveness to change and ability to detect flares, and further evidence needs to be gathered to support its use in clinical trials. As this work progresses, this new instrument will hopefully meet the requirements to be included in the OMERACT core instrument set ensuring its widespread uptake. With further work, the PMR-IS could also become routinely used in clinical practice, as an aid to improving person-centred care for PMR.

References

Anderson, J., Sayles, H., Curtis, J. R., Wolfe, F., & Michaud, K. (2010). Converting modified Health Assessment Questionnaire (HAQ), multidimensional HAQ, and HAQII scores into original HAQ scores using models developed with a large cohort of rheumatoid arthritis patients. *Arthritis Care and Research*, *62*(10), 1481–1488.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573.

Andrich, D. (2003). Controversy and the Rasch Model. *Medical Care*, *42* (Supplement), I–7.

Andrich, D., Lyne, A., Sheridan, B., & Luo, G. (2003). *RUMM2020*. RUMM Laboratory. Rasch Unidimensional Measurement Model. [Computer software].

Au, H.J., Ringash, J., Brundage, M., Palmer, M., Richardson, H., & Meyer, R. M. (2010). Added value of health-related quality of life measurement in cancer clinical trials: the experience of the NCIC CTG. *Expert Review of Pharmacoeconomics & Outcomes Research*, *10*(2), 119–128.

Baker, R. (2016). *Calculating the SEM*. Walking with Richard. https://wwrichard.net/2016/06/01/calculating-the-sem/

Barber H.S. (1957). Myalgic syndrome with constitutional effects. Polymyalgia rheumatica. *Annals of the Rheumatic Diseases*, *16*, 230–237.

Barraclough, K., Liddell, W. G., du Toit, J., Foy, C., Dasgupta, B., Thomas, M., & Hamilton, W. (2008). Polymyalgia rheumatica in primary care: a cohort study of the diagnostic criteria and outcome. *Family Practice*, *25*(5), 328–333.

Basch, E., Deal, A. M., Dueck, A. C., Scher, H. I., Kris, M. G., Hudis, C., & Schrag, D. (2017). Overall Survival Results of a Trial Assessing Patient-Reported Outcomes for Symptom Monitoring During Routine Cancer Treatment. *JAMA*, *318*(2), 197.

Basch, E., Deal, A. M., Kris, M. G., Scher, H. I., Hudis, C. A., Sabbatini, P., Rogak, L., Bennett, A. v, Dueck, A. C., Atkinson, T. M., Chou, J. F., Dulko, D., Sit, L., Barz, A., Novotny, P., Fruscione, M., Sloan, J. A., & Schrag, D. (2016). Symptom Monitoring With Patient-Reported Outcomes During Routine Cancer Treatment: A Randomized Controlled Trial. *Journal of Clinical Oncology*, *34*(6), 557–565.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In *Lord FM, Novick MR. Statistical Theories of Mental Test Scores* (pp. 397–545). Addison-Wesley.

Black, R. J., Goodman, S. M., Ruediger, C., Lester, S., Mackie, S. L., & Hill, C. L. (2017). A survey of glucocorticoid adverse effects and benefits in rheumatic diseases: The patient perspective. *Journal of Clinical Rheumatology*, *23*(8), 416–420.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet (London, England)*, *1*(8476), 307–310.

Blockmans, D., de Ceuninck, L., Vanderschueren, S., Knockaert, D., Mortelmans, L., & Bobbaers, H. (2007). Repetitive 18-fluorodeoxyglucose positron emission tomography in isolated polymyalgia rheumatica: a prospective study in 35 patients. *Rheumatology (Oxford, England)*, *46*(4), 672–677.

Boers, M., Beaton, D. E., Shea, B. J., Maxwell, L. J., Bartlett, S. J., Bingham, C. O., Conaghan, P. G., D'Agostino, M. A., De Wit, M. P., Gossec, L., March, L., Simon, L. S., Singh, J. A., Strand, V., Wells, G. A., & Tugwell, P. (2019). OMERACT filter 2.1:

Elaboration of the conceptual framework for outcome measurement in health intervention studies. *Journal of Rheumatology*, *46*(8), 1021–1027.
Boers, M., Brooks, P., Simon, L. S., Strand, V., Tugwell, P., Boers, M., Brooks, P., Simon, L. S., Strand, V., & Idzerda, L. (2007). OMERACT: An international initiative to improve outcome measurement in rheumatology. *Trials*, *8*(38), 1–6.
Boers, M., Brooks, P., Strand, C. v., & Tugwell, P. (1998). The OMERACT filter for Outcome Measures in Rheumatology. *The Journal of Rheumatology*, *25*(2), 198–199.
Boers, M., Kirwan, J. R., Wells, G., Beaton, D., Gossec, L., D'Agostino, M. A., Conaghan, P. G., Bingham, C. O., Brooks, P., Landewé, R., March, L., Simon, L. S., Singh, J. A., Strand, V., & Tugwell, P. (2014). Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. *Journal of Clinical Epidemiology*, *67*(7), 745–753.
Bowling, A. (2001). *Measuring disease: A review of disease specific quality of life measurement scales* (2nd ed.). Open University Press.
Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health*, *27*(3), 281–291.
Boyce, M. B., Browne, J. P., & Greenhalgh, J. (2014). The experiences of professionals with using information from patient-reported outcome measures to improve the quality of healthcare: a systematic review of qualitative research. *BMJ Quality & Safety*, *23*(6), 508–518.
Brazier, J. E., Harper, R., Jones, N. M., O'Cathain, A, Thomas, K. J., Usherwood, T., & Westlake, L. (1992). Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *BMJ (Clinical Research Ed.)*, *305*(6846), 160–164.
Brazier, J., Jones, N., & Kind, P. (1993). Testing the validity of the Euroqol and comparing it with the SF-36 health survey questionnaire. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *2*(3), 169–180.
Brook, R. H. (2010). The End of the Quality Improvement Movement. *JAMA*, *304*(16), 1831.
Brooks, R. (1996). EuroQol: the current state of play. *Health Policy (Amsterdam, Netherlands)*, *37*(1), 53–72.
Bruce, B., & Fries, J. F. (2003). The Stanford Health Assessment Questionnaire: dimensions and practical applications. *Health and Quality of Life Outcomes*, *1*, 20.
Bruce, B., & Fries, J. F. (2005). The Health Assessment Questionnaire (HAQ). *Clinical & Experimental Rheumatology*, *23*(5 Suppl 39), S14-8.
Bruce, W. (1888). Senile rheumatic gout. *British Medical Journal*, *2*(1450), 811–813.
Busse, J. W., Bartlett, S. J., Dougados, M., Johnston, B. C., Guyatt, G. H., Kirwan, J. R., Kwoh, K., Maxwell, L. J., Moore, A., Singh, J. A., Stevens, R., Strand, V., Suarez-Almazor, M. E., Tugwell, P., & Wells, G. A. (2015). Optimal strategies for reporting pain in clinical trials and systematic reviews: Recommendations from an OMERACT 12 workshop. *Journal of Rheumatology*, *42*(10), 1962–1970.
Bylicki, O., Gan, H. K., Joly, F., Maillet, D., You, B., & Péron, J. (2015). Poor patient-reported outcomes reporting according to CONSORT guidelines in randomized clinical trials evaluating systemic cancer therapy. *Annals of Oncology*, *26*(1), 231–237.
Calvert, M., Blazeby, J., Altman, D. G., Revicki, D. A., Moher, D., Brundage, M. D., & Consort Pro Group, F. T. (2013). Reporting of Patient-Reported Outcomes in Randomized Trials. *JAMA*, *309*(8), 814.

Calvert, M., Kyte, D., Mercieca-Bebber, R., Slade, A., Chan, A.-W., King, M. T., Hunn, A., Bottomley, A., Regnault, A., Chan, A.-W., Ells, C., O'Connor, D., Revicki, D., Patrick, D., Altman, D., Basch, E., Velikova, G., Price, G., Draper, H., … Groves, T. (2018). Guidelines for Inclusion of Patient-Reported Outcomes in Clinical Trial Protocols. *JAMA*, *319*(5), 483.

Calvert, M., Kyte, D., Price, G., Valderas, J. M., & Hjollund, N. H. (2019). Maximising the impact of patient reported outcome assessment for patients and society. *BMJ*, k5267.

Calvert, M., Thwaites, R., Kyte, D., & Devlin, N. (2015). Putting patient-reported outcomes on the 'Big Data Road Map.' *Journal of the Royal Society of Medicine*, *108*(8), 299–303.

Cantini, F., Niccoli, L., Nannini, C., Padula, A., Olivieri, I., Boiardi, L., & Salvarani, C. (2005). Inflammatory changes of hip synovial structures in polymyalgia rheumatica. *Clinical and Experimental Rheumatology*, *23*(4), 462–468.

Cantini, F., Salvarani, C., Olivieri, I., Macchioni, L., Ranzi, A., Niccoli, L., Padula, A., & Boiardi, L. (2000). Erythrocyte sedimentation rate and C-reactive protein in the evaluation of disease activity and severity in polymyalgia rheumatica: a prospective follow-up study. *Seminars in Arthritis & Rheumatism*, *30*(1), 17–24.

Cantini, F., Salvarani, C., Olivieri, I., Niccoli, L., Padula, A., Macchioni, L., Boiardi, L., Ciancio, G., Mastrorosato, M., Rubini, F., Bozza, A., & Zanfranceschi, G. (2001). Shoulder ultrasonography in the diagnosis of polymyalgia rheumatica: a case-control study. *The Journal of Rheumatology*, *28*(5), 1049–1055.

Caporali, R., Cimmino, M. A., Ferraccioli, G., Gerli, R., Klersy, C., & Salvarani, C. (2004). Prednisone plus Methotrexate for Polymyalgia Rheumatica. A Randomized, double blind, placebo controlled trial. *Annals of Internal Medicine*, *141*(7), 493–500.

Caporali, R., Montecucco, C., Epis, O., Bobbio-Pallavicini, F., Maio, T., & Cimmino, M. A. (2001). Presenting features of polymyalgia rheumatica (PMR) and rheumatoid arthritis with PMR-like onset: a prospective study. *Annals of the Rheumatic Diseases*, *60*(11), 1021–1024.

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, *1*(2), 245–276.

Cawley, A., Prior, J. A., Muller, S., Helliwell, T., Hider, S. L., Dasgupta, B., Barraclough, K., & Mallen, C. D. (2017). Association between characteristics of pain and stiffness and the functional status of patients with incident polymyalgia rheumatica from primary care. *Clinical Rheumatology*, *37*(6), 1639–1644.

Centres for Medicare and Medicaid Services. (2017). *MEDCAC Meeting 3/22/2017 - Health Outcomes in Heart Failure Treatment Technology Studies*.

Chuang, T. Y., Hunder, G. G., Ilstrup, D. M., & Kurland, L. T. (1982). Polymyalgia rheumatica: a 10-year epidemiologic and clinical study. *Annals of Internal Medicine*, *97*(5), 672–680.

Cimmino, M. A., Salvarani, C., Macchioni, P., Gerli, R., Bartoloni Bocci, E., Montecucco, C., Caporali, R., & Systemic Vasculitis Study Group of the Italian Society for, R. (2008). Long-term follow-up of polymyalgia rheumatica patients treated with methotrexate and steroids. *Clinical & Experimental Rheumatology*, *26*(3), 395–400.

Coe, R. (2002). It's the effect size, stupid - What effect size is and why it is important. *Annual Conference of the British Educational Research Association*, 1–12.

Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research*, *12*(3), 229–238.
Cordier, R., Brown, T., Clemson, L., & Byles, J. (2018). Evaluating the Longitudinal Item and Category Stability of the SF-36 Full and Summary Scales Using Rasch Analysis. *BioMed Research International*, *2018*, 1013453.
COSMIN. (2019a). *About the initiative.* Cosmin. https://www.cosmin.nl/about/?fbclid=IwAR0DHu6cInjIF-EqL0K-urlhZJPP-uVgtj9uYcPnc2yCsoOs1YHq5cUzWRs
COSMIN. (2019b). *COSMIN Taxonomy of Measurement Properties*. Cosmin. https://www.cosmin.nl/tools/cosmin-taxonomy-measurement-properties/
Crohnbach, I. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.
Crosby, R. D., Kolotkin, R. L., & Williams, G. R. (2003). Defining clinically meaningful change in health-related quality of life. *Journal of Clinical Epidemiology*, *56*(5), 395–407.
Crowson, C. S., & Matteson, E. L. (2017). Contemporary prevalence estimates for giant cell arteritis and polymyalgia rheumatica, 2015. *Seminars in Arthritis & Rheumatism*, *47*(2), 253–256.
Crowson, C. S., Matteson, E. L., Myasoedova, E., Michet, C. J., Ernste, F. C., Warrington, K. J., Davis, J. M., Hunder, G. G., Therneau, T. M., & Gabriel, S. E. (2011). The lifetime risk of adult-onset rheumatoid arthritis and other inflammatory autoimmune rheumatic diseases. *Arthritis & Rheumatism*, *63*(3), 633–639.
Cutolo, M., Cimmino, M. A., & Sulli, A. (2009). Polymyalgia rheumatica vs late-onset rheumatoid arthritis. *Rheumatology*, *48*(2), 93–95.
Damman, O. C., Verbiest, M. E. A., Vonk, S. I., Berendse, H. W., Bloem, B. R., de Bruijne, M. C., & Faber, M. J. (2019). Using PROMs during routine medical consultations: The perspectives of people with Parkinson's disease and their health professionals. *Health Expectations*, *22*(5), 939–951.
Dasgupta, B., Borg, F. A., Hassan, N., Barraclough, K., Bourke, B., Fulcher, J., Hollywood, J., Hutchings, A., Kyle, V., Nott, J., Power, M., Samanta, A., Bsr, Bhpr Standards, G., & Audit Working, G. (2010). BSR and BHPR guidelines for the management of polymyalgia rheumatica. *Rheumatology*, *49*(1), 186–190.
Dasgupta, B., Borg, F., & Hassan, N. (2010). BSR and BHPR guidelines for the management of giant cell arteritis. *Rheumatology*, *49*(8), 1594–1597.
Dasgupta, B., Cimmino, M. A., Maradit-Kremers, H., Schmidt, W. A., Schirmer, M., Salvarani, C., Bachta, A., Dejaco, C., Duftner, C., Jensen, H. S., Duhaut, P., Poór, G., Kaposi, N. P. N. P., Mandl, P., Balint, P. V, Schmidt, Z., Iagnocco, A., Nannini, C., Cantini, F., … Matteson, E. L. (2012). 2012 provisional classification criteria for polymyalgia rheumatica: a European League Against Rheumatism/American College of Rheumatology collaborative initiative. *Annals of the Rheumatic Diseases*, *71*(4), 484–492.
Davidson, M. (2009). Rasch Analysis of 24-, 18- and 11-item Versions of the Roland-Morris Disability Questionnaire. *Quality of Life Research*, *18*(4), 473–481.
Dawson, J., Fitzpatrick, R., Murray, D., & Carr, A. (1996). Comparison of measures to assess outcomes in total hip replacement surgery. *Quality in Health Care : QHC*, *5*(2), 81–88.

Dawson, J., Fitzpatrick, R., Murray, D., & Carr, A. (1998). Questionnaire on the perceptions of patients about total knee replacement. *The Journal of Bone and Joint Surgery. British Volume*, *80*(1), 63–69.

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011a). Concepts, theories and models. In *Measurement in medicine* (8th ed., pp. 7–29). Cambridge University Press.

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011b). Development of a measurement instrument. In *Measurement in medicine* (1st ed., pp. 30–64). Cambridge University Press.

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011c). Field-testing: item reduction and data structure. In *Measurement in medicine* (1st ed., pp. 65–92). Cambridge University Press.

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011d). Interpretability. In *Measurement in medicine* (1st ed., pp. 227–274). Cambridge University Press.

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011e). Reliability. In *Measurement in medicine* (1st ed., pp. 96–149). Cambridge University Press.

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011f). Responsiveness. In *Measurement in medicine* (1st ed., pp. 202–226). Cambridge University Press.

de Vet, H. C. W., Terwee, Caroline B, Mokkink, L. B., & Knol, D. L. (2011). Validity. In *Measurement in medicine* (1st ed., pp. 150–201). Cambridge University Press.

Dejaco, C., Brouwer, E., Mason, J. C., Buttgereit, F., Matteson, E. L., & Dasgupta, B. (2017). Giant cell arteritis and polymyalgia rheumatica: Current challenges and opportunities. *Nature Reviews Rheumatology*, *13*(10), 578–592.

Dejaco, C., Duftner, C., Buttgereit, F., Matteson, E. L., & Dasgupta, B. (2017). The spectrum of giant cell arteritis and polymyalgia rheumatica: Revisiting the concept of the disease. *Rheumatology (United Kingdom)*, *56*(4), 506–515.

Dejaco, C., Duftner, C., Dasgupta, B., Matteson, E. L., & Schirmer, M. (2011). Polymyalgia rheumatica and giant cell arteritis: management of two diseases of the elderly. *Aging Health*, *7*(4), 633–645.

Dejaco, C., Singh, Y. P., Perel, P., Hutchings, A., Camellino, D., Mackie, S., Abril, A., Bachta, A., Balint, P., Barraclough, K., Bianconi, L., Buttgereit, F., Carsons, S., Ching, D., Cid, M., Cimmino, M., Diamantopoulos, A., Docken, W., Duftner, C., … Dasgupta, B. (2015). 2015 recommendations for the management of polymyalgia rheumatica: A European League Against Rheumatism/American College of Rheumatology collaborative initiative. *Annals of the Rheumatic Diseases*, *74*(10), 1799–1807.

Dejaco, C., Singh, Y. P., Perel, P., Hutchings, A., Camellino, D., Mackie, S., Matteson, E. L., & Dasgupta, B. (2015). 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 po. *Annals of the Rheumatic Diseases*, *74*(10), 1808–1817.

Department of Health. (2009). *Guidance on the routine collection of Patient Reported Outcome Measures (PROMs)*.

Devauchelle-Pensec, V., Berthelot, J. M., Cornec, D., Renaudineau, Y., Marhadour, T., Jousse-Joulin, S., Querellou, S., Garrigues, F., De Bandt, M., Gouillou, M., & Saraux, A. (2016). Efficacy of first-line tocilizumab therapy in early polymyalgia rheumatica: A prospective longitudinal study. *Annals of the Rheumatic Diseases*, *75*(8), 1506–1510.

Devlin, N. J., Appleby, J., Buxton, M., & Vallance-Owen, A. (2010). Getting the most out of PROMS. Putting health outcomes at the heart of NHS decision making. In *Health Economics*. www.kingsfund.org.uk/publications

Di Munno, O., Imbimbo, B., Mazzantini, M., Milani, S., Occhipinti, G., & Pasero, G. (1995). Deflazacort versus methylprednisolone in polymyalgia rheumatica: clinical equivalence and relative antiinflammatory potency of different treatment regimens. *Journal of Rheumatology*, *22*(8), 1492–1498.

Doward, L. C., Spoorenberg, A., Cook, S. A., Whalley, D., Helliwell, P. S., Kay, L. J., McKenna, S. P., Tennant, A., van der Heijde, D., & Chamberlain, M. A. (2003). Development of the ASQoL: a quality of life instrument specific to ankylosing spondylitis. *Annals of the Rheumatic Diseases*, *62*(1), 20–26.

Duarte, C., Ferreira, R. J. d. O., Mackie, S. L., Kirwan, J. R., & Pereira da Silva, J. A. (2015). Outcome Measures in Polymyalgia Rheumatica. A Systematic Review. *The Journal of Rheumatology*, *42*(12), 2503–2511.

Duncan, P. W., Bode, R. K., Lai, S. M., & Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: The stroke impact scale. *Archives of Physical Medicine and Rehabilitation*, *84*(7), 950–963.

Dworkin, R. H., Turk, D. C., Wyrwich, K. W., Beaton, D., Cleeland, C. S., Farrar, J. T., Haythornthwaite, J. A., Jensen, M. P., Kerns, R. D., Ader, D. N., Brandenburg, N., Burke, L. B., Cella, D., Chandler, J., Cowan, P., Dimitrova, R., Dionne, R., Hertz, S., Jadad, A. R., … Zavisic, S. (2008). Interpreting the Clinical Importance of Treatment Outcomes in Chronic Pain Clinical Trials: IMMPACT Recommendations. *Journal of Pain*, *9*(2), 105–121.

Edlund, M., & Tancredi, L. R. (1985). Quality of life: an ideological critique. *Perspectives in Biology and Medicine*, *28*(4), 591–607.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174.

Eghtedari, A. A., Esselinckx, W., & Bacon, P. A. (1976). Circulating immunoblasts in polymyalgia rheumatica. *Annals of the Rheumatic Diseases*, *35*(2), 158–162.

EuroQol group. (1990). EuroQol - a new facility for the measurement of health-related quality of life. *Health Policy*, *16*(3), 199–208.

Feinberg, H. L., Schrepferman, C. G., Sherman, J. D., Dietzen, C. J., & Feinberg, G. D. (1995). Pharmacotherapy. Steroid treatment of polymyalgia rheumatica. *American Journal of Pain Management*, *5*(2), 52–54.

Ferraccioli, G., Salaffi, F., Salvatore, D. V., Casatta, L., & Bartoli, E. (1996). Methotrexate in PMR: Preliminary Results of an Open Randomized Study. *The Journal of Rheumatology*, *23*, 624–628.

Field, A. (2009). Exploratory factor analysis. In *Discovering statistics using SPSS* (3rd ed., pp. 627–685). SAGE Publications.

Fitzpatrick, R., Davey, C., Buxton, M. J., & Jones, D. R. (1998). Evaluating patient-based outcome measures for use in clinical trials. In *Health technology assessment (Winchester, England)* (Vol. 2, Issue 14).

Friedlander, M., Mercieca-Bebber, R. L., & King, M. T. (2016). *Patient-reported outcomes (PRO) in ovarian cancer clinical trials—lost opportunities and lessons learned.* *27*(suppl 1), i66–i71.

Fries, J. F., Spitz, P., Kraines, R. G., & Holman, H. R. (1980). Measurement of patient outcome in arthritis. *Arthritis & Rheumatism*, *23*(2), 137–145.

Fu, H. (2007). Clinical observation on effect of modified Yanghe Decoction combined with glycocorticoid for treatment of polymyalgia rheumatica. *Chinese Journal of Integrated Traditional and Western Medicine*, *27*(10), 894–897.

Gargon, E., Gorst, S. L., & Williamson, P. R. (2019). Choosing important health outcomes for comparative effectiveness research: 5th annual update to a systematic review of core outcome sets for research. *PLOS ONE*, *14*(12).

Gargon, E., Gurung, B., Medley, N., Altman, D. G., Blazeby, J. M., Clarke, M., & Williamson, P. R. (2014). Choosing Important Health Outcomes for Comparative Effectiveness Research: A Systematic Review. *PLoS ONE*, *9*(6).

General Practice Notebook. (2021). *Polymyalgia Rheumatica - General Practice Notebook*. https://gpnotebook.com/en-gb/simplepage.cfm?ID=-1254817780

Giavarana, D. (2015). Understanding Bland Altman Analysis. *Biochemia Medica*, *25*(2), 141–151.

Gonzalez-Gay, M. A. (2004). Giant cell arteritis and polymyalgia rheumatica: two different but often overlapping conditions. *Seminars in Arthritis & Rheumatism*, *33*(5), 289–293.

González-Gay, M. A., Amoli, M. M., Garcia-Porrua, C., & Ollier, W. E. R. (2003). Genetic markers of disease susceptibility and severity in giant cell arteritis and polymyalgia rheumatica. *Seminars in Arthritis and Rheumatism*, *33*(1), 38–48.

Gonzalez-Gay, M. A., Garcia-Porrua, C., & Vazquez-Caruncho, M. (1998). Polymyalgia rheumatica in biopsy proven giant cell arteritis does not constitute a different subset but differs from isolated polymyalgia rheumatica. *The Journal of Rheumatology*, *25*(9), 1750–1755.

Gonzalez-Gay, M. A., Garcia-Porrua, C., Vazquez-Caruncho, M., Dababneh, A., Hajeer, A., & Ollier, W. E. (1999). The spectrum of polymyalgia rheumatica in northwestern Spain: incidence and analysis of variables associated with relapse in a 10 year study. *Journal of Rheumatology*, *26*(6), 1326–1332.

González-Gay, M. A., Matteson, E. L., & Castañeda, S. (2017). Polymyalgia rheumatica. *The Lancet*, *390*(10103), 1700–1712.

Gonzalez-Gay, M. A., Rodriguez-Valverde, V., Blanco, R., Fernandez-Sueiro, J. L., Armona, J., Figueroa, M., & Martinez-Taboada, V. M. (1997). Polymyalgia rheumatica without significantly increased erythrocyte sedimentation rate. A more benign syndrome. *Archives of Internal Medicine*, *157*(3), 317–320.

Gonzalez-Gay, M. A., Vazquez-Rodriguez, T. R., Lopez-Diaz, M. J., Miranda-Filloy, J. A., Gonzalez-Juanatey, C., Martin, J., & Llorca, J. (2009). Epidemiology of giant cell arteritis and polymyalgia rheumatica. *Arthritis & Rheumatism*, *61*(10), 1454–1461.

Gran, J. T., & Myklebust, G. (2000). The incidence and clinical characteristics of peripheral arthritis in polymyalgia rheumatica and temporal arteritis: a prospective study of 231 cases. *Rheumatology*, *39*(3), 283–287.

Gran, J. T., Myklebust, G., Wilsgaard, T., & Jacobsen, B. K. (2001). Survival in polymyalgia rheumatica and temporal arteritis: a study of 398 cases and matched population controls. *Rheumatology*, *40*(11), 1238–1242.

Gray, M. (2017). Value based healthcare. *BMJ (Online)*, *356*, 1–2. https://doi.org/10.1136/bmj.j437

Gray, M., & el Turabi, A. (2012). Optimising the value of interventions for populations. *BMJ (Online)*, *345*(7877), 1–2. https://doi.org/10.1136/bmj.e6192

Greenhalgh, J., Dalkin, S., Gooding, K., Gibbons, E., Wright, J., Meads, D., Black, N., Valderas, J. M., & Pawson, R. (2017). Functionality and feedback: a realist synthesis of the collation, interpretation and utilisation of patient-reported outcome measures data to improve patient care. *Health Services and Delivery Research*, *5*(2), 1–280.

Grotle, M., Wilkens, P., Garratt, A. M., Scheel, I., & Storheim, K. (2013). Which Roland-Morris Disability Questionnaire? Rasch analysis of four different versions tested in a Norwegian population. *Journal of Rehabilitation Medicine*, *45*(7), 670–677.

Guttman, L. (1950). The basis for scalogram analysis. In S. Stouffer, L. Guttman, F. Suchman, P. Lazarsfeld, S. Star, & J. Clausen (Eds.), *Studies in Social Psychology in World War II. Measurement and Prediction* (pp. 60–90). Princeton University.

Halls, S., Sinnathurai, P., Hewlett, S., Mackie, S. L., March, L., Bartlett, S. J., Bingham, C. O., Alten, R., Campbell, I., Hill, C. L., Holt, R. J., Hughes, R., Kirwan, J. R., Leong, A. L., Leung, Y. Y., Lyddiatt, A., Neill, L., & Orbai, A. M. (2017). Stiffness is the cardinal symptom of inflammatory musculoskeletal diseases, yet still variably measured: Report from the OMERACT 2016 Stiffness Special Interest Group. *Journal of Rheumatology*, *44*(12), 1904–1910.

Hayden, J. A., van der Windt, D. A., Cartwright, J. L., Côté, P., & Bombardier, C. (2013). Assessing bias in studies of prognostic factors. *Annals of Internal Medicine*, *158*(4), 280–286.

Haywood, K. L., Garratt, A. M., Jordan, K. P., Healey, E. L., & Packham, J. C. (2010). Evaluation of ankylosing spondylitis quality of life (EASi-QoL): Reliability and validity of a new patient-reported outcome measure. *Journal of Rheumatology*, *37*(10), 2100–2109.

Helfgott, S. M., & Kieval, R. I. (1996). Polymyalgia rheumatica in patients with a normal erythrocyte sedimentation rate. *Arthritis & Rheumatism*, *39*(2), 304–307.

Helliwell, T., Brouwer, E., Pease, C. T., Hughes, R., Hill, C. L., Neill, L. M., Halls, S., Simon, L. S., Mallen, C. D., Boers, M., Kirwan, J. R., & Mackie, S. L. (2016). Development of a provisional core domain set for polymyalgia rheumatica: Report from the OMERACT 12 Polymyalgia Rheumatica Working Group. *Journal of Rheumatology*, *43*(1), 182–186.

Helliwell, T., Hider, S. L., & Mallen, C. D. (2013). Polymyalgia rheumatica: diagnosis, prescribing, and monitoring in general practice. *British Journal of General Practice*, *63*(610), e361–e366.

Hendriks, J., Fyfe, S., Styles, I., Skinner, S. R., Merriman, G., & Hendriks, J. (2012). Scale construction utilising the Rasch unidimensional measurement model. *Australasian Medical Journal*, *5*(5), 251–261.

Higgins, P., Savovic, H., Page, M., & Sterne, J. (2016). A revised tool for assessing risk of bias in randomised trials. *Cochrane Database of Systematic Reviews*, *10*.

Hill, J. C., Thomas, E., Hill, S., Foster, N. E., & van der Windt, D. A. (2015). Development and validation of the keele musculoskeletal patient reported outcome measure (MSK-PROM). *PLoS ONE*, *10*(4), 1–14.

Hoes, J. N., Jacobs, J. W. G., Verstappen, S. M. M., Bijlsma, J. W. J., & van der Heijden, G. J. M. G. (2009). Adverse events of low- to medium-dose oral glucocorticoids in inflammatory diseases: a meta-analysis. *Annals of the Rheumatic Diseases*, *68*(12), 1833–1838.

Huang, A., & Castrejon, I. (2016). Patient-reported outcomes in trials of patients with polymyalgia rheumatica: a systematic literature review. *Rheumatology International*, *36*(7), 897–904.

Hunder, G. G. (2006). The early history of giant cell arteritis and polymyalgia rheumatica: First descriptions to 1970. *Mayo Clinic Proceedings*, *81*(8), 1071–1083.

Hutchings, A., Hollywood, J., Lamping, D. L., Pease, C. T., Chakravarty, K., Silverman, B., Choy, E. H. S., Scott, D. G. I., Hazleman, B. L., Bourke, B., Gendi, N., & Dasgupta, B. (2007). Clinical outcomes, quality of life, and diagnostic uncertainty in the first year of polymyalgia rheumatica. *Arthritis & Rheumatism*, *57*(5), 803–809.

IBM. (2014). *SPSS Statistics Desktop* (No. 24). [Computer software]. http://www14.software.ibm.com/download/data/web/en_US/trialprograms/W110742E06714B29.html

IBM. (2020). *SPSS Statistics Desktop* (No. 27). [Computer software]. http://www14.software.ibm.com/download/data/web/en_US/trialprograms/W110742E06714B29.html

International Society for Quality of Life Research. (2018). *What Is QOL? | ISOQOL*. https://www.isoqol.org/what-is-qol/

Izumi, K., Kuda, H., Ushikubo, M., Kuwana, M., Takeuchi, T., & Oshima, H. (2015). Tocilizumab is effective against polymyalgia rheumatica: Experience in 13 intractable cases. *RMD Open*, *1*(1). https://doi.org/10.1136/rmdopen-2015-000162

Jordan, K., Dunn, K. M., Lewis, M., & Croft, P. (2006). A minimal clinically important difference was derived for the Roland-Morris Disability Questionnaire for low back pain. *Journal of Clinical Epidemiology*, *59*(1), 45–52.

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika*, *35*(4), 401–415.

Kalke, S., Mukerjee, D., & Dasgupta, B. (2000). A study of the health assessment questionnaire to evaluate functional status in polymyalgia rheumatica. *Rheumatology*, *39*(8), 883–885.

Kelfve, S., Kivi, M., Johansson, B., & Lindwall, M. (2020). Going web or staying paper? The use of web-surveys among older people. *BMC Medical Research Methodology*, *20*(252), 1–12.

Kirshner, B., & Guyatt, G. (1985). A methodological framework for assessing health indices. *Journal of Chronic Diseases*, *38*(1), 27–36.

Kirwan, J. R. (2013). OMERACT - Where it came from and what it is trying to do. *Indian Journal of Rheumatology*, *8*(SUPPL.1), 8–11.

Kreiner, F., & Galbo, H. (2010). Effect of etanercept in polymyalgia rheumatica: a randomized controlled trial. *Arthritis Research & Therapy*, *12*(5), R176.

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B. W., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, *114*(1–3), 163–173.

Kwong, E., & Black, N. (2017). Retrospectively patient-reported pre-event health status showed strong association and agreement with contemporaneous reports. *Journal of Clinical Epidemiology*, *81*, 22–32.

Kyle, V., & Hazleman, B. L. (1993). The clinical and laboratory course of polymyalgia rheumatica/giant cell arteritis after the first two months of treatment. *Annals of the Rheumatic Diseases*, *52*(12), 847–850.

Kyte, D., Cockwell, P., Lencioni, M., Skrybant, M., Hildebrand, M. von, Price, G., Squire, K., Webb, S., Brookes, O., Fanning, H., Jones, T., & Calvert, M. (2016). Reflections on the

national patient-reported outcome measures (PROMs) programme: Where do we go from here? *Journal of the Royal Society of Medicine*, *109*(12), 441–445.

Lally, L., Forbess, L., Hatzis, C., & Spiera, R. (2016). Brief Report: A Prospective Open-Label Phase IIa Trial of Tocilizumab in the Treatment of Polymyalgia Rheumatica. *Arthritis & Rheumatology*, *68*(10), 2550–2554.

Laporte, J.-P., Garrigues, F., Huwart, A., Jousse-Joulin, S., Marhadour, T., Guellec, D., Cornec, D., Devauchelle-Pensec, V., & Saraux, A. (2019). Localized Myofascial Inflammation Revealed by Magnetic Resonance Imaging in Recent-onset Polymyalgia Rheumatica and Effect of Tocilizumab Therapy. *The Journal of Rheumatology*, *46*(12), 1619 LP – 1626.

Leeb, B. F., & Bird, H. A. (2004). A disease activity score for polymyalgia rheumatica. *Annals of the Rheumatic Diseases*, *63*(10), 1279–1283.

Leeb, B. F., Bird, H. A., Nesher, G., Andel, I., Hueber, W., Logar, D., Montecucco, C. M., Rovensky, J., Sautner, J., & Sonnenblick, M. (2003). EULAR response criteria for polymyalgia rheumatica: results of an initiative of the European Collaborating Polymyalgia Rheumatica Group (subcommittee of ESCISIT). *Annals of the Rheumatic Diseases*, *62*(12), 1189–1194.

Lewis, S. (2019). *Patient reported outcome measures enhance communication with patients*. BMJ. https://blogs.bmj.com/bmj/2019/05/28/sally-lewis-patient-reported-outcome-measures-enhance-communication-with-patients/

Lim, C. R., Harris, K., Dawson, J., Beard, D. J., Fitzpatrick, R., & Price, A. J. (2015). Floor and ceiling effects in the OHS: An analysis of the NHS PROMs data set. *BMJ Open*, *5*(7). https://doi.org/10.1136/bmjopen-2015-007765

Linacre, J. M. (1994). *Sample Size and Item Calibration or Person Measure Stability*. Rasch Measurement Transactions. http://www.rasch.org/rmt/rmt74m.htm

Lohr, K. N., & Zebrack, B. J. (2009). Using patient-reported outcomes in clinical practice: Challenges and opportunities. *Quality of Life Research*, *18*(1), 99–107. https://doi.org/10.1007/s11136-008-9413-7

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). Statistical theories of mental test scores. In *Statistical theories of mental test scores.* Addison-Wesley.

Lund, B., Egsmose, C., Jørgensen, S., & Krogsgaard, M. (1987). Establishment of the relative antiinflammatory potency of deflazacort and prednisone in polymyalgia rheumatica. *Calcified Tissue International*, *41*(6), 316–320.

Mackie, S. L., Arat, S., da Silva, J., Duarte, C., Halliday, S., Hughes, R., Morris, M., Pease, C. T., Sherman, J. W., Simon, L. S., Walsh, M., Westhovens, R., Zakout, S., & Kirwan, J. R. (2014). Polymyalgia Rheumatica (PMR) Special Interest Group at OMERACT 11: outcomes of importance for patients with PMR. *J Rheumatol*, *41*(4), 819–823.

Mackie, S. L., Hensor, E. M. A., Haugeberg, G., Bhakta, B., & Pease, C. T. (2010). Can the prognosis of polymyalgia rheumatica be predicted at disease onset? Results from a 5-year prospective study. *Rheumatology*, *49*(4), 716–722.

Mackie, S. L., Hughes, R., Walsh, M., Day, J., Newton, M., Pease, C., Kirwan, J., & Morris, M. (2015). An impediment to living Life": Why and how should we measure stiffness in polymyalgia rheumatica? *PLoS ONE*, *10*(5), 1–13.

Mackie, S. L., Koduri, G., Hill, C. L., Wakefield, R. J., Hutchings, A., Loy, C., Dasgupta, B., & Wyatt, J. C. (2015). Accuracy of musculoskeletal imaging for the diagnosis of polymyalgia rheumatica: systematic review. *RMD Open*, *1*(1), e000100. https://doi.org/10.1136/rmdopen-2015-000100

Mackie, S. L., Pease, C. T., Fukuba, E., Harris, E., Emery, P., Hodgson, R., Freeston, J., & McGonagle, D. (2015). Whole-body MRI of patients with polymyalgia rheumatica identifies a distinct subset with complete patient-reported response to glucocorticoids. *Annals of the Rheumatic Diseases*, *74*(12), 2188–2192.

Mackie, S. L., Twohig, H., Neill, L. M., Harrison, E., Shea, B., Black, R. J., Kermani, T. A., Merkel, P. A., Mallen, C. D., Buttgereit, F., Mukhtyar, C., Simon, L. S., & Hill, C. L. (2017). The OMERACT core domain set for outcome measures for clinical trials in polymyalgia rheumatica. *Journal of Rheumatology*, *44*(10), 1515–1521.

Manzo, C., & Emamifar, A. (2019). Polymyalgia Rheumatica and Seronegative Elderly-Onset Rheumatoid Arthritis : Two Different Diseases with Many Similarities. *European Medical Journal*, *4*(September), 111–119.

Manzo, C., & Milchert, M. (2018). Polymyalgia rheumatica with normal values of both erythrocyte sedimentation rate and C-reactive protein concentration at the time of diagnosis: a four-point guidance. *Reumatologia*, *56*(1), 1–2.

Maska, L., Anderson, J., & Michaud, K. (2011). Measures of functional status and quality of life in rheumatoid arthritis: Health Assessment Questionnaire Disability Index (HAQ), Modified Health Assessment Questionnaire (MHAQ), Multidimensional Health Assessment Questionnaire (MDHAQ), Health Assessment. *Arthritis Care and Research*, *63*(SUPPL. 11), 4–13.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

Matteson, E. L., Maradit-Kremers, H., Cimmino, M. A., Schmidt, W. A., Schirmer, M., Salvarani, C., Bachta, A., Dejaco, C., Duftner, C., Slott Jensen, H., Poor, G., Kaposi, N. P., Mandl, P., Balint, P. v, Schmidt, Z., Iagnocco, A., Cantini, F., Nannini, C., Macchioni, P., … Dasgupta, B. (2012). Patient-reported outcomes in polymyalgia rheumatica. *Journal of Rheumatology*, *39*(4), 795–803.

Mazzantini, M., Torre, C., Miccoli, M., Baggiani, A., Talarico, R., Bombardieri, S., & di Munno, O. (2012). Adverse Events During Longterm Low-dose Glucocorticoid Treatment of Polymyalgia Rheumatica: A Retrospective Study. *The Journal of Rheumatology*, *39*(3), 552–557.

McCarthy, E. M., MacMullan, P. A., Al-Mudhaffer, S., Madigan, A., Donnelly, S., McCarthy, C. J., Molloy, E. S., Kenny, D., & McCarthy, G. M. (2013). Plasma fibrinogen is an accurate marker of disease activity in patients with polymyalgia rheumatica. *Rheumatology (United Kingdom)*, *52*(3), 465–471.

McCarthy, E. M., MacMullan, P. A., Al-Mudhaffer, S., Madigan, A., Donnelly, S., McCarthy, C. J., Molloy, E. S., Kenny, D., & McCarthy, G. M. (2014). Plasma fibrinogen along with patient-reported outcome measures enhances management of polymyalgia rheumatica: A prospective study. *Journal of Rheumatology*, *41*(5), 931–937.

McCleod, S. (2019). *What does effect size tell you?* Simply Psychology. https://www.simplypsychology.org/effect-size.html

McColl, E., Jacoby, A., Thomas, L., Soutter, J., Bamford, C., Steen, N., Thomas, R., Harvey, E., Garratt, A., & Bond, J. (2001). Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients. *Health Technology Assessment*, *5*(31).

McGonagle, D., Pease, C., Marzo-Ortega, H., O'Connor, P., Gibbon, W., & Emery, P. (2001). Comparison of extracapsular changes by magnetic resonance imaging in

patients with rheumatoid arthritis and polymyalgia rheumatica. *The Journal of Rheumatology*, *28*(8), 1837–1841.
McGraw, K. O., & Wong, S. P. (1996). Forming Inferences about Some Intraclass Correlation Coefficients. *Psychological Methods*, *1*(1), 30–46.
McHorney, C. A., Haley, S. M., & Ware, J. E. (1997). Evaluation of the MOS SF-36 physical functioning scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. *Journal of Clinical Epidemiology*, *50*(4), 451–461.
McHorney, C. A., & Tarlov, A. R. (1995). Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *4*(4), 293–307.
McKenna, S. P., Cook, S. A., Whalley, D., Doward, L. C., Richards, H. L., Griffiths, C. E. M., & van Assche, D. (2003). Development of the PSORIQoL, a psoriasis-specific measure of quality of life designed for use in clinical practice and trials. *The British Journal of Dermatology*, *149*(2), 323–331.
Meliconi, R., Pulsatelli, L., Uguccioni, M., Salvarani, C., Macchioni, P., Melchiorri, C., Focherini, M. C., Frizziero, L., & Facchini, A. (1996). Leukocyte infiltration in synovial tissue from the shoulder of patients with polymyalgia rheumatica. Quantitative analysis and influence of corticosteroid treatment. *Arthritis and Rheumatism*, *39*(7), 1199–1207.
Mercieca-Bebber, R., King, M. T., Calvert, M. J., Stockler, M. R., & Friedlander, M. (2018). The importance of patient-reported outcomes in clinical trials and strategies for future optimization. *Patient Related Outcome Measures*, *Volume 9*, 353–367.
Michet, C. J., & Matteson, E. L. (2017). Polymyalgia rheumatica. *BMJ*, *336*(7647), 765–769.
Mitchell, C., Dwyer, R., Hagan, T., & Mathers, N. (2011). Impact of the QOF and the NICE guideline in the diagnosis andmanagement of depression: a qualitative study. *British Journal of General Practice*, *61*(586), e279–e289.
Miyake, K., & Katsuyama, T. (2014). Analysis of the relationship between polymyalgia rheumatica and matrix metalloproteinase-3 levels during the first medical examination and during treatment. *Japanese Journal of Clinical Immunology*, *37*(1), 48–54.
Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*(1), 1.
Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Quality of Life Research*, *27*(5), 1171–1179.
Mokkink, L. B., Prinsen, C. A., Patrick, D., Alonso, J., Bouter, L. M., de Vet, H. C., & Terwee, C. B. (2019). COSMIN Study Design checklist for Patient-reported outcome measurement instruments. In *Department of Epidemiology and Biostatistics Amsterdam Public Health research institute Amsterdam University Medical Centers, location VUmc* (Issue July). www.cosmin.nl
Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010a). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research*, *19*(4), 539–549.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010b). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, *63*(7), 737–745.

Moores, K. L., Jones, G. L., & Radley, S. C. (2012). Development of an instrument to measure face validity, feasibility and utility of patient questionnaire use during health care: The QQ-10. *International Journal for Quality in Health Care*, *24*(5), 517–524.

Mori, S., Koga, Y., & Ito, K. (2007). Clinical characteristics of polymyalgia rheumatica in Japanese patients: evidence of synovitis and extracapsular inflammatory changes by fat suppression magnetic resonance imaging. *Modern Rheumatology*, *17*(5), 369–375.

Muller, S., Hider, S. L., Helliwell, T., Lawton, S., Barraclough, K., Dasgupta, B., Zwierska, I., & Mallen, C. D. (2016). Characterising those with incident polymyalgia rheumatica in primary care: Results from the PMR Cohort Study. *Arthritis Research and Therapy*, *18*(1), 1–9.

Muller, S., Whittle, R., Hider, S. L., Belcher, J., Helliwell, T., Morton, C., Hughes, E., Lawton, S. A., & Mallen, C. D. (2019). Longitudinal clusters of pain and stiffness in polymyalgia rheumatica: 2-year results from the PMR Cohort Study. *Rheumatology*.

Narvaez, J. A., Nolla-Sole, J. M., Narvaez, J. A., Clavaguera, M. T., Valverde-Garcia, J., Roig-Escofet, D., Narváez, J. A., Nolla-Solé, J. M., Narváez, J. A., Clavaguera, M. T., Valverde-García, J., & Roig-Escofet, D. (2001). Musculoskeletal manifestations in polymyalgia rheumatica and temporal arteritis. *Annals of the Rheumatic Diseases*, *60*(11), 1060 LP – 1063.

Narvaez, J., Clavaguera, M. T., Nolla-Sole, J. M., Valverde-Garcia, J., & Roig-Escofet, D. (2000). Lack of association between infection and onset of polymyalgia rheumatica. *Journal of Rheumatology*, *27*(4), 953–957.

Narvaez, J., Nolla-Sole, J. M., Clavaguera, M. T., Valverde-Garcia, J., & Roig-Escofet, D. (1999). Longterm therapy in polymyalgia rheumatica: effect of coexistent temporal arteritis. *Journal of Rheumatology*, *26*(9), 1945–1952.

Nazarinia, AM., Moghimi, J., & Toussi, J. (2012). Efficacy of methotrexate in patients with polymyalgia rheumatica. *Koomesh*, *14*(3), 265–270.

Nelson, E. C., Eftimovska, E., & Lind, C. (2015). *Patient reported outcome measures in practice*. *7818*(February), 1–3.

NICE. (2021). *Polymyalgia rheumatica | Health topics A to Z | CKS | NICE*. https://cks.nice.org.uk/topics/polymyalgia-rheumatica/

Norman, G., & Streiner, D. (2008). *Biostatistics: the bare essentials* (3rd ed.). B.C.Decker Inc.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Office for National Statistics. (2018). *Internet users, UK: 2018*. https://www.ons.gov.uk/businessindustryandtrade/itandinternetindustry/bulletins/internetusers/2018

OMERACT. (2017). Chapter 4: Instrument selection for Core Outcome Measurement Sets Instrument selection: Three pillars, four questions, one answer. In *The OMERACT Handbook*.

OMERACT. (2019). OMERACT Handbook Instrument Selection Chapter 5 Mar 2019. In *OMERACT handbook*. https://omeracthandbook.org/handbook

Orr, C. K., Najm, A., Young, F., McGarry, T., Biniecka, M., Fearon, U., & Veale, D. J. (2018). The utility and limitations of CRP, ESR and DAS28-CRP in appraising disease activity in rheumatoid arthritis. *Frontiers in Medicine*, *5*(185), 1–8.

Otteva, E. N., & Kocherova, T. I. (2008). Activity index in rheumatic polymyalgias. *Klinicheskaia Meditsina*, *86*(1), 41–44.

Owen, C. E., Liew, D. F. L., & Buchanan, R. R. C. (2019). Musculotendinous Inflammation: The Defining Pathology of Polymyalgia Rheumatica? *The Journal of Rheumatology*, *46*(12), 1552 LP – 1555.

Owen, C. E., Yates, M., Twohig, H., Muller, S., Neill, L. M., Harrison, E., Shea, B., Simon, L. S., Hill, C. L., & Mackie, S. L. (2019). Toward a Core Outcome Measurement Set for Polymyalgia Rheumatica: Report from the OMERACT 2018 Special Interest Group. *The Journal of Rheumatology*, jrheum.181050.

Ozen, G., Inanc, N., Unal, A. U., Bas, S., Kimyon, G., Kisacik, B., Onat, A. M., Murat, S., Keskin, H., Can, M., Mengi, A., Cakir, N., Balkarli, A., Cobankara, V., Yilmaz, N., Yazici, A., Dogru, A., Sahin, M., Sahin, A., … Direskeneli, H. (2016). Assessment of the New 2012 EULAR/ACR Clinical Classification Criteria for Polymyalgia Rheumatica: A Prospective Multicenter Study. *The Journal of Rheumatology*, *43*(5), 893–900.

Palard-Novello, X., Querellou, S., Gouillou, M., Saraux, A., Marhadour, T., Garrigues, F., Abgral, R., Salaün, P. Y., & Devauchelle-Pensec, V. (2016). Value of 18F-FDG PET/CT for therapeutic assessment of patients with polymyalgia rheumatica receiving tocilizumab as first-line treatment. *European Journal of Nuclear Medicine and Molecular Imaging*, *43*(4), 773–779.

Pamuk, O. N., Donmez, S., Karahan, B., Pamuk, G. E., & Cakir, N. (2009). Giant cell arteritis and polymyalgia rheumatica in northwestern Turkey: Clinical features and epidemiological data. *Clinical & Experimental Rheumatology*, *27*(5), 830–833.

Partington, R. J., Muller, S., Helliwell, T., Mallen, C. D., & Abdul Sultan, A. (2018). Incidence, prevalence and treatment burden of polymyalgia rheumatica in the UK over two decades: A population-based study. *Annals of the Rheumatic Diseases*, *77*(12), 1750–1756.

Partington, R., Muller, S., Mallen, C. D., Abdul Sultan, A., & Helliwell, T. (2020). Mortality among patients with polymyalgia rheumatica: A retrospective cohort study. *Arthritis Care & Research*.

Patrick, D. L., Burke, L. B., Powers, J. H., Scott, J. A., Rock, E. P., Dawisha, S., O'Neill, R., & Kennedy, D. L. (2007). Patient-reported outcomes to support medical product labeling claims: FDA perspective. *Value in Health*, *10*(SUPPL. 2).

Pease, C. T., Haugeberg, G., Montague, B., Hensor, E. M. A., Bhakta, B. B., Thomson, W., Ollier, W. E. R., & Morgan, A. W. (2009). Polymyalgia rheumatica can be distinguished from late onset rheumatoid arthritis at baseline: results of a 5-yr prospective study. *Rheumatology*, *48*(2), 123–127.

Perfetto, F., Moggi-Pignone, A., Becucci, A., Cantini, F., di Natale, M., Livi, R., Tempestini, A., & Matucci-Cerinic, M. (2005). Seasonal pattern in the onset of polymyalgia rheumatica. *Annals of the Rheumatic Diseases*, *64*(11), 1662–1663.

Pett, M., Lackey, N., & Sullivan, J. (2011a). An Overview of Factor Analysis. In *Making Sense of Factor Analysis* (pp. 2–12).

Pett, M., Lackey, N., & Sullivan, J. (2011b). Evaluating and refining the factors. In *Making Sense of Factor Analysis* (pp. 1–33).

Pett, M., Lackey, N., & Sullivan, J. (2011c). Extracting the initial factors. In *Making Sense of Factor Analysis* (pp. 1–38).

Pincus, T., Summey, J. A., Soraci, S. A., Wallston, K. A., & Hummon, N. P. (1983). Assessment of patient satisfaction in activities of daily living using a modified stanford health assessment questionnaire. *Arthritis & Rheumatism*, *26*(11), 1346–1353.

Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research*, *27*(5), 1147–1157.

Raheel, S., Shbeeb, I., Crowson, C. S., & Matteson, E. L. (2017). Epidemiology of Polymyalgia Rheumatica 2000–2014 and Examination of Incidence and Survival Trends Over 45 Years: A Population-Based Study. *Arthritis Care and Research*, *69*(8), 1282–1285.

Ramzai, J. (2020). *Clearly explained: Pearson V/S Spearman Correlation Coefficient*. Towards Data Science. https://towardsdatascience.com/clearly-explained-pearson-v-s-spearman-correlation-coefficient-ada2f473b8

RAND. (2000). *36-Item Short Form Survey (SF-36) Scoring Instructions*. RAND Corporation. https://www.rand.org/health-care/surveys_tools/mos/36-item-short-form/scoring.html

Rasch, G. (1960). *Probabalistic models for some intelligence and achievement tests*. MESA Press.

Rea, L., & Parker, R. (1997). *Designing and conducting survey research: a comprehensive guide* (2nd ed.). Jossey-Bass.

Reeve, B. B., Wyrwich, K. W., Wu, A. W., Velikova, G., Terwee, C. B., Snyder, C. F., Schwartz, C., Revicki, D. A., Moinpour, C. M., Mcleod, L. D., Lyons, J. C., Lenderking, W. R., Hinds, P. S., Hays, R. D., Greenhalgh, J., Gershon, R., Feeny, D., Fayers, P. M., Cella, D., … Butt, Z. (2013). ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Quality of Life Research*, *22*(8), 1889–1905.

Robson, J. C., Dawson, J., Doll, H., Cronholm, P. F., Milman, N., Kellom, K., Ashdown, S., Easley, E., Gebhart, D., Lanier, G., Mills, J., Peck, J., Luqmani, R. A., Shea, J., Tomasson, G., & Merkel, P. A. (2018). Validation of the ANCA-associated vasculitis patient-reported outcomes (AAV-PRO) questionnaire. *Annals of the Rheumatic Diseases*, *77*(8), 1157–1164.

Ruof, J., & Stucki, G. (1999). Validity aspects of erythrocyte sedimentation rate and C-reactive protein in ankylosing spondylitis: a literature review. *The Journal of Rheumatology*, *26*(4), 966–970.

Rutherford, C., Costa, D., Mercieca-Bebber, R., Rice, H., Gabb, L., & King, M. (2016). Mode of administration does not cause bias in patient-reported outcome results: a meta-analysis. *Quality of Life Research*, *25*(3), 559–574.

Ryan, K., Gannon-Slater, N., & Culbertson, M. J. (2012). Improving Survey Methods With Cognitive Interviews in Small- and Medium-Scale Evaluations. *American Journal of Evaluation*, *33*(3), 414–430.

Salvarani, C., Cantini, F., & Hunder, G. G. (2008). Polymyalgia rheumatica. *Lancet*, *372*(9070), 234–245.

Salvarani, C., Cantini, F., Macchioni, P., Olivieri, I., Niccoli, L., Padula, A., & Boiardi, L. (1998). Distal musculoskeletal manifestations in polymyalgia rheumatica: a prospective followup study. *Arthritis & Rheumatism*, *41*(7), 1221–1226.

Salvarani, C., Cantini, F., Niccoli, L., Catanoso, M. G., I, P. M., Pulsatelli, L. I. A., Padula, A., Olivieri, I., Boiard, L., Rezunatologica, U., Medicina, D., Emilia, R., Macchioni, P., Pulsatelli, L. I. A., Padula, A., Olivieri, I., & Boiardi, L. (2003). Treatment of refractory polymyalgia rheumatica with infliximab: a pilot study. *Journal of Rheumatology*, *30*(4), 760–763.

Salvarani, C., Cantini, F., Niccoli, L., Macchioni, P., Consonni, D., Bajocchi, G., Vinceti, M., Catanoso, M. G., Pulsatelli, L., Meliconi, R., & Boiardi, L. (2005). Acute-phase reactants and the risk of relapse/recurrence in polymyalgia rheumatica: a prospective followup study. *Arthritis & Rheumatism*, *53*(1), 33–38.

Salvarani, C., Gabriel, S. E., O'Fallon, W. M., & Hunder, G. G. (1995). Epidemiology of polymyalgia rheumatica in Olmsted County, Minnesota, 1970-1991. *Arthritis & Rheumatism*, *38*(3), 369–373.

Salvarani, C., Macchioni, P., Manzini, C., Paolazzi, G., Trotta, A., Manganelli, P., Cimmino, M., Gerli, R., Catanoso, M. G., Boiardi, L., Cantini, F., Klersy, C., & Hunder, G. G. (2007). Infliximab plus Prednisone or Placebo plus Prednisone for the Initial Treatment of Polymyalgia Rheumatica. *Annals of Internal Medicine*, *146*, 631–640.

Schmidt, W. A., Seifert, A., Gromnica-Ihle, E., Krause, A., & Natusch, A. (2008). Ultrasound of proximal upper extremity arteries to increase the diagnostic yield in large-vessel giant cell arteritis. *Rheumatology (Oxford, England)*, *47*(1), 96–101.

Schmier, J. K., & Halpern, M. T. (2004). Patient recall and recall bias of health state and health status. *Expert Review of Pharmacoeconomics & Outcomes Research*, *4*(2), 159–163.

Seyfarth, B., Harten, P., & Loffler, H. (1996). Thrombocytosis in PMR. *Deutsche Medizinische Wochenschrift*, *121*(41), 1255–1260.

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., Altman, D. G., Booth, A., Chan, A. W., Chang, S., Clifford, T., Dickersin, K., Egger, M., Gøtzsche, P. C., Grimshaw, J. M., Groves, T., Helfand, M., … Whitlock, E. (2015). Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015: Elaboration and explanation. *BMJ (Online)*, *349*(December 2014), 1–25.

Shbeeb, I., Challah, D., Raheel, S., Crowson, C. S., & Matteson, E. L. (2018). Comparable Rates of Glucocorticoid-Associated Adverse Events in Patients With Polymyalgia Rheumatica and Comorbidities in the General Population. *Arthritis Care & Research*, *70*(4), 643–647.

Smeeth, L., Cook, C., & Hall, A. J. (2006). Incidence of diagnosed polymyalgia rheumatica and temporal arteritis in the United Kingdom, 1990-2001. *Annals of the Rheumatic Diseases*, *65*(8), 1093–1098.

Snyder, C. F., & Aaronson, N. K. (2009). Use of patient-reported outcomes in clinical practice. *The Lancet*, *374*(9687), 369–370.

Spies-Dorgelo, M. N., Terwee, C. B., Stalman, W. A., & Van Der Windt, D. A. (2006). Reproducibility and responsiveness of the Symptom Severity Scale and the hand and finger function subscale of the Dutch arthritis impact measurement scales (Dutch-AIMS2-HFF) in primary care patients with wrist or hand problems. *Health and Quality of Life Outcomes*, *4*(1).

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Lowe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097.

St Sauver, J. L., Grossardt, B. R., Yawn, B. P., Joseph Melton, L., Pankratz, J. J., Brue, S. M., & Rocca, W. A. (2012). Data resource profile: The rochester epidemiology project (REP) medical records-linkage system. *International Journal of Epidemiology*, *41*(6), 1614–1624.

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care and Research*, *57*(8), 1358–1362.

Tennant, A., Hillman, M., Fear, J., Pickering, A., & Chamberlain, M. A. (1996). Are we making the most of the Stanford Health Assessment Questionnaire? *British Journal of Rheumatology*, *35*(6), 574–578.

Tennant, A., McKenna, S. P., & Hagell, P. (2004). Application of Rasch analysis in the development and application of quality of life instruments. *Value in Health*, *7*(SUPPL. 1), S22–S26.

Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. *Statistics in Medicine*, *19*(11–12), 1651–1683.

Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., Bouter, L. M., & de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*(1), 34–42.

Tidy, C., & Knott, L. (2021). *Polymyalgia Rheumatica.* https://patient.info/doctor/polymyalgia-rheumatica-pro

Twohig, H., Jones, G., Mackie, S., Mallen, C., & Mitchell, C. (2018). Assessment of the face validity, feasibility and utility of a patient-completed questionnaire for polymyalgia rheumatica: A postal survey using the QQ-10 questionnaire. *Pilot and Feasibility Studies*, *4*(1).

Twohig, H., Mitchell, C., Mallen, C., Adebajo, A., & Mathers, N. (2015). "I suddenly felt I'd aged": A qualitative study of patient experiences of polymyalgia rheumatica (PMR). *Patient Education and Counseling*, *98*(5), 645–650. https://doi.org/10.1016/j.pec.2014.12.013

U.S Department of Health and Human Services. (2009). Guidance for Industry Use in Medical Product Development to Support Labeling Claims Guidance for Industry. In *Clinical/Medical Federal Register* (Issue December).

Van Der Roer, N., Ostelo, R. W. J. G., Bekkering, G. E., Van Tulder, M. W., & De Vet, H. C. W. (2006). Minimal Clinically Important Change for Pain Intensity, Functional Status, and General Health Status in Patients With Nonspecific Low Back Pain. *Spine*, *31*(5), 578–582.

van der Velde, G., Beaton, D., Hogg-Johnston, S., Hurwitz, E., & Tennant, A. (2009). Rasch analysis provides new insights into the measurement properties of the neck disability index. *Arthritis Care and Research*, *61*(4), 544–551.

Vodicka, E., Kim, K., Devine, E. B., Gnanasakthy, A., Scoggins, J. F., & Patrick, D. L. (2015). Inclusion of patient-reported outcome measures in registered clinical trials: Evidence from ClinicalTrials.gov (2007–2013). *Contemporary Clinical Trials*, *43*, 1–9.

Wagener, P. (1995). Treatment of rheumatic polymyalgia with methotrexate. *Zeitschrift Fur Rheumatologie*, *54*(6), 413–416.

Ware, J. E. (2000). SF-36 Health Survey update. *Spine*, *25*(24), 3130–3139.

Ware, J. E. J., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care*, *30*(6), 473–483.

Wells, G., Beaton, D., Shea, B., Boers, M., Simon, L., Strand, V., Brooks, P., & Tugwell, P. (2001). Minimal clinically important differences: review of methods. *The Journal of Rheumatology*, *28*(2), 406 LP – 412.

Willis, G. B., Royston, P., & Bercini, D. (1991). The use of verbal report methods in the development and testing of survey questionnaires. *Applied Cognitive Psychology*, *5*(3), 251–267.

Wolfe, F., Michaud, K., Li, T., & Katz, R. S. (2010). EQ-5D and SF-36 quality of life measures in systemic lupus erythematosus: Comparisons with rheumatoid arthritis, noninflammatory rheumatic disorders, and fibromyalgia. *Journal of Rheumatology*, *37*(2), 296–304.

Yamane, T., Yamauchi, H., Abe, N., Torio, N., Shimada, R., Senba, T., Imaizumi, Y., & Nomura, T. (2003). Serum amyloid A as a useful index of disease activity in polymyalgia rheumatica. *Ryumachi*, *43*(3), 544–548.

Yates, M., Graham, K., Watts, R. A., & MacGregor, A. J. (2016). The prevalence of giant cell arteritis and polymyalgia rheumatica in a UK primary care population. *BMC Musculoskeletal Disorders*, *17*(1), 1–9.

Yates, M., Owen, C. E., Muller, S., Graham, K., Neill, L., Twohig, H., Boers, M., Rodriguez, M. P., Goodman, S. M., Cheah, J., Dejaco, C., Mukhtyar, C., Nielsen, B. D., Robson, J., Simon, L. S., Shea, B., Mackie, S. L., & Hill, C. L. (2020). Feasibility and Face Validity of Outcome Measures for Use in Future Studies of Polymyalgia Rheumatica: An OMERACT Study. *Journal of Rheumatology*, *47*(9), 1379–1384.

Appendix 4.1: Protocol for systematic review of outcome measures in PMR

Title of the review	A systematic review of outcome measures used in research studies of polymyalgia rheumatica (PMR).
First reviewer	Helen Twohig
Other reviewers (with role/contribution in the review)	Claire Owen Sarah L Mackie Catherine Hill Elisabeth Brouwer Samantha Hider Sara Muller, Christian Mallen, and Caroline Mitchell (PhD supervisors)
Funding source	HT is funded by a Welcome Trust Primary Care Doctoral Fellowship SLM is funded by an NIHR Clinician Scientist award CDM is funded by the National Institute for Health Research (NIHR) Collaborations for Leadership in Applied Health Research and Care West Midlands, the NIHR School for Primary Care Research and a NIHR Research Professorship in General Practice (NIHRRP-2014-04-026). This article/paper/report presents independent research funded by the National Institute for Health Research (NIHR). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.
PROSPERO registration number	CRD42017080058
Amendments to the protocol	In the event of protocol amendments the date of each amendment will be accompanied by a description of the change and the rationale and a new version number applied.

1. Background to review

Polymyalgia rheumatica (PMR) is a chronic inflammatory condition characterised by the subacute-onset of shoulder and pelvic girdle pain, and early morning stiffness in men and women over the age of 50 years.[1] In the absence of a gold standard test, diagnosis is based upon a clinical construct supported by laboratory evidence of systemic inflammation.

Despite certain advances, including the development of standardised classification criteria and increasing use of imaging modalities to characterise its precise pathology, PMR remains poorly understood. Glucocorticoids represent the mainstay of treatment, however the disease course of patients with PMR typically exhibits heterogeneity. Some individuals respond rapidly and require therapy for less than one year, whereas others experience initial treatment response failure or subsequent disease relapse.[2] The use of disease modifying anti-rheumatic drugs (DMARDs) has been examined, however their role is not altogether clear owing to partly contradictory results from randomised trials.[3] Further complicating matters, the morbidity associated with low-moderate doses of corticosteroids is both greater and occurs more frequently in PMR than other rheumatic conditions.[4]

If strong evidence-based recommendations are to be made concerning management, clinical trials must utilise valid and reliable outcome measures. In 2015 a systematic review of 35 studies conducted by the OMERACT PMR Working Group found significant variability in the assessment of PMR in research settings.[5] Additionally, most of the instruments identified were deemed to have been insufficiently validated according to the OMERACT Filter 2.0. This systematic review however was limited in its scope and did not make any assessment of the quality of the relevant studies. At the OMERACT 2016 meeting, a core domain set for future research studies of PMR was agreed on. We intend to carry out a systematic review of outcome measures and instruments used in research studies of PMR, which is more rigorous and broader in scope than that done previously, and map the identified instruments to the core domains. This will help to determine which instruments to take forwards to full

evaluation under the OMERACT filter 2.1 requirements of truth, discrimination and feasibility.

2. Specific objectives/questions the review will address

By systematically reviewing the relevant literature we will:

5. Identify all of the outcomes which have been measured in studies of PMR and the instruments used to assess them.
6. Categorise these into the domains defined in the core domain set agreed by the OMERACT PMR Working Group in 2016 (laboratory markers of systemic inflammation, pain, stiffness, physical function).
7. Assess the number and quality of studies using each outcome / instrument and whether the instruments have been validated for use in PMR.
8. Identify research gaps with respect to outcomes and instruments used in studies of PMR.

This review will inform which instruments to take forwards and consider under the OMERACT Filter 2.1 requirements of truth, discrimination and feasibility and ultimately determine their suitability for inclusion in a core outcome set for future PMR research studies.

3. a) Eligibility Criteria for including studies in the review
If the PICOS format does not fit the research question of interest, please split up the question into separate concepts and put one under each heading

i. Population, or participants and conditions of interest	**Patients with polymyalgia rheumatica**
ii. Interventions/Exposure/item of interest	**Research studies reporting original quantitative data**
iii. Comparisons or control groups, if any	**N/A**

iv. Outcomes of interest	**Outcomes measured and instruments used**
v. Setting	**Primary or secondary care**
vi. Study designs	**Randomised controlled trials** **Other interventional trials** **Longitudinal observational studies** **Cohort studies** **Cross sectional studies** **Case control studies**

3. b) Criteria for excluding studies not covered in inclusion criteria Any specific populations excluded, date range, language, whether abstracts or full text available, etc
Studies will be excluded if they do not study patients with PMR or if they consider patients with PMR and GCA as a single group and don't present disease specific data. Diagnostic studies will be excluded. Editorials, commentaries, review articles, case reports and letters without original quantitative patient data will be excluded. A full text version of the article must be available. Non-English publications will be permitted on the proviso translation can be achieved with the assistance of either Google Translate or the involvement of an investigator from the OMERACT PMR Working Group or affiliated academic institutions who is a competent speaker of the language concerned.

4. Search methods

Electronic databases & websites Please list all databases that are to be searched and include the interface (eg NHS HDAS, EBSCO, OVID etc) and date ranges searched for each.	The following databases will be searched from database inception to Sep 30th 2017: • MEDLINE via OVID • Embase via HDAS • Cumulative Index of Nursing and Allied Health (CINAHL) via EBSCO • Web of Science • Cochrane library (Cochrane central register of controlled trials and Cochrane database of systematic reviews) A search strategy will be designed for each database using the thesauri, text words, truncated text words and abbreviations of key words.
Other methods used for identifying relevant research ie contacting experts and reference checking, citation tracking	Trial registries such as ClinicalTrials.gov, ISRCTN and the EU Clinical Trials Register will be searched to identify ongoing and completed trials and investigators will be contacted to request any available information on outcomes being measured. Additional papers will be retrieved by searching the reference lists of key papers and review articles and by identifying papers that cite these key papers using the facility on Web of Science. Experts in the field will be contacted to identify any unpublished work that might be of relevance.
Journals hand searched	None

5. Methods of review

How will search results be managed & documented? ie which reference management software, how duplicates dealt with	Titles and abstracts will be imported into bibliographic management software (Endnote) and duplicates will be identified and removed. Initial screening of search results by title will be done by HT. The database of de-duplicated and title-eligible studies will be uploaded into 'Covidence' and this software will be used to support the rest of the review process. Abstracts will then be screened independently by HT and one other member of a team of reviewers. Records will be kept to enable a clear flowchart to be developed demonstrating numbers of articles identified and excluded at each stage.
Selection process Number of reviewers, how agreements to be reached and disagreements dealt with, etc.	Selection will be a two-step process. In the first step, two investigators (HT and one other) will independently screen abstracts of the articles. Full reports of those selected will then be read and evaluated by HT and one other of a team of investigators against the inclusion and exclusion criteria to determine which studies to include in the review. Any disagreement will be resolved by discussion and if needed, consensus with a third investigator (CM or SH).

Quality assessment Tools or checklists used with references or URLs, was this piloted? Is it to be carried out at same time as data extraction?	A modified Quality in Prognosis Studies (QUIPS)[6] tool will be used for quality assessment. Key markers of quality relevant to our study question include how well defined the study population is in relation to PMR diagnosis, whether the setting from which they are recruited is well described and whether the outcome measure is well described enough (either in the article or via references) that it can be evaluated later using the OMERACT Filter 2.1 Instrument Selection Algorithm. Quality assessment will be carried out at the same time as data extraction.
How is data to be extracted? What information is to be collected on each included study? If databases or forms on Word or Excel are used, were these piloted and how is this recorded and by how many reviewers?	Data will be extracted into a standardised Excel data collection database. This database will be developed specifically for this review and piloted and amended as needed prior to use. Data will be extracted independently by two reviewers and any discrepancies resolved through discussion.

Outcomes to be extracted & hierarchy/priority of measures ie which measure is preferred and if that is not available which is next in order of preference?	Information extracted will include: • journal information • lead author • publication year • study design / study type • setting (primary / secondary care or mixed) • criteria used to define PMR • age range of participants • gender mix of participants • range of PMR duration of participants • sample size • intervention (if any) • duration of follow up • outcomes measured and instruments used to assess PMR, with categorisation into relevant domains as outlined by the OMERACT PMR Working Group in 2016 (laboratory markers of systemic inflammation, pain, stiffness, physical function). Supplementary literature searches will be carried out pertaining to key instruments identified for each domain to establish whether they have been validated for use in PMR and if not, in which conditions they have been validated (content and face validity, feasibility/practicality, longitudinal and cross-sectional construct validity and test-retest reliability). This will inform future work by the OMERACT PMR working group to develop recommendations on instruments to be used for each domain.

Narrative synthesis Details of what methods, how synthesis will be done and by whom. Is the Narrative Synthesis Framework to be used?	A systematic narrative synthesis will be provided with information presented in the text and tables to summarise and explain the characteristics and findings of included studies. A table of the outcome measures identified, the domains they map to and the associated instruments used will also be presented and discussed.
Meta-analysis Details of what and how analysis and testing will be done. If no meta-analysis is to be conducted, please give reason.	N/A
Will the overall strength of evidence be assessed? If so, how? ie GRADE?	No, this isn't applicable to our study question. We will consider the quality of individual studies to inform judgements about the relevance of outcomes measured and instruments used but as we are not evaluating a specific intervention or treatment we do not intend to make any assessment of overall strength of evidence.

6. Presentation of results	
Outputs from review Papers and target journals, conference presentations, reports, etc	This systematic review will be submitted for publication in a peer reviewed journal and also used to inform the work of the OMERACT PMR working group ahead of the 2018 international meeting. It will also contribute to HT's PhD thesis.

Timeline for review – when do you aim to complete each stage of the review	
Protocol	By October 2017
Literature searching	By October 2017
Quality appraisal	By January 2017
Data extraction	By January 2017
Synthesis	By February 2018

Writing up	By March 2018

References

1. Hunder GG. The early history of giant cell arteritis and polymyalgia rheumatica: first descriptions to 1970. Mayo Clinic Proceedings. 2006;81(8):1071-83.
2. Weyand CM, Fulbright JW, Evans JM, Hunder GG, Goronzy JJ. Corticosteroid requirements in polymyalgia rheumatica. Archives of Internal Medicine. 1999;159(6):577-84.
3. Dejaco C, Singh YP, Perel P, et al. 2015 Recommendations for the management of polymyalgia rheumatica: a European League Against Rheumatism/American College of Rheumatology collaborative initiative. Annals of the Rheumatic Diseases. 2015;74(10):1799-807.
4. Hoes JN, Jacobs JW, Verstappen SM, Bijlsma JW, Van der Heijden GJ. Adverse events of low- to medium-dose oral glucocorticoids in inflammatory diseases: a meta-analysis. Annals of the Rheumatic Diseases. 2009;68(12):1833-8.
5. Duarte C, Ferreira RJ, Mackie SL, et al. Outcome Measures in Polymyalgia Rheumatica. A Systematic Review. Journal of Rheumatology. 2015;42(12):2503-11.
6. Hayden JA, van der Windt DA, Cartwright JL, Cote P, Bombardier C. Assessing bias in studies of prognostic factors. Annals of Internal Medicine. 2013;158:280-286.

Appendix 4.2: Outcomes in PMR systematic review search terms

Searches run 10/10/17

Medline via Ovid SP

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

Cochrane library via Wiley Online

1. polymyalgia rheumatica:ti,ab,kw
2. MeSH descriptor: [Polymyalgia Rheumatica] explode all trees
3. polymyalgia
4. rheumatic polymyalgia
5. polymyalgia arteritica
6. forestier certonciny syndrome
7. rheumatic myalgia
8. rhizomelic pseudopolyarthritis
9. polymyalgi
10. senile gout
11. 1-10 combined with OR

Web of Science

1. polymyalgia rheumatica
2. rheumatic polymyalgia
3. polymyalgia arteritica
4. forestier certonciny syndrome
5. rheumatic myalgia
6. rhizomelic pseudopolyarthritis
7. senile gout
8. polymyalgi*
9. 1-8 combined with OR

Embase via HDAS

1. polymyalgia rheumatica
2. “RHEUMATIC POLYMYALGIA”
3. polymyalgi*
4. polymyalgia arteritica
5. Forestier certonciny syndrome
6. pseudopolyarthritis rhizomelica
7. senile gout
8. 1-7 combined with OR

CINAHL via EBSCO

1. polymyalgia rheumatica
2. MH (“Polymyalgia rheumatica”)
3. rheumatic polymyalgia
4. polymyalgia arteritica
5. forestier certonciny syndrome
6. rheumatic myalgia
7. rhizomelic pseudopolyarthritis
8. polymyalgi*
9. senile gout
10. 1-9 combined with OR

Totals

References found and imported: 16222
Duplicates removed: 3028
References left: 13194

First round title screening 819 kept.

Second round title screening 629 kept.

Imported into Covidence – 609 (20 duplicates found and removed)

Appendix 4.3: Studies for which full text was not available

Reference	Reason
Eghtedari, 1976. Circulating immunoblasts in PMR. (Eghtedari et al., 1976).	Full text not available due to age of article
Feinberg, 1995. Steroid treatment of PMR (Feinberg et al., 1995)	Unable to access full text
Fu, 2007. Clinical observation on effect of modified Yanghe Decoction combined with glucocorticoid for treatment of polymyalgia rheumatica (Fu, 2007)	Full text in Chinese and not available on-line
Lund, 1987. Establishment of the relative anti-inflammatory potency of deflazacort and pred in PMR. (Lund et al., 1987)	Unable to access full text
Miyake, 2014. Analysis of the relationship between PMR and matrix metalloproteinase-3 levels during the first medical exam and during treatment (Miyake & Katsuyama, 2014)	Unable to access full text
Nazarinia, 2012. Efficacy of methotrexate in PMR. (Nazarinia et al., 2012)	Full text only available in Arabic.
Otteva, 2008. Activity index in rheumatic polymyalgias. (Otteva & Kocherova, 2008)	Full text only available in Russian.
Seyfarth, 1996. Thrombocytosis in PMR. (Seyfarth et al., 1996)	Full text in German and not available on-line
Wagener, 1995. Treatment of rheumatic polymyalgia with PMR. (Wagener, 1995)	Full text in German and not available on-line
Yamane, 2003. Serum amyloid A as a useful index of disease activity in polymyalgia rheumatica. (Yamane et al., 2003)	Full text in Russian and not available on-line

Appendix 4.4: Data extraction spreadsheet

Study ID (Year, lead author)	Journal	Study design	Setting	Criteria used to define PMR	Sample size	Age range / mean age	Gender ratio	Intervention	Duration of follow up	Outcomes measured	Instruments used	Instrument validity described / referenced	Instrument validated in PMR	Key findings
Alvarez-Rodriguez 2010	Ann rheum dis	Prospective case control study	Not stated	Chuang et al Patients with PMR and ESR< 40 mm/1 h but who satisfied other clinical	34 patients, 17 controls	Mean age cases 72.8.	61.8% F	All received initial dose 10mg pred then tapering dose determined individually	Post treatment sample taken after mean treatment duration of 9.7 weeks	Soluble cytokines (inc IL6)	CBA assay and ELISA	Yes	N/A	IL6 is the more detectable proimflammatory circulating cytokine in PMR. Levels of IL6 an TNFa reduce with CS. Cytokine profile didn't allow distinction between patients with high or low ESR.
										Intracellular cytokines	Flow cytometry analysis	Yes	N/A	
										Cytokines in supernatants	ELISA	Yes	N/A	
Benucci 2015	European Review for Medical and Pharmacologic al Sciences	Non-randomised prospective observational study	Not stated	2012 EULAR / ACR criteria	81 GC naïve patients, 38 treated with 6-MP and 43 treated with MR-P	MR-P mean age 73.9. 6-MP mean age 74	MR-P group 53% F. 6-MP group 63% F	6-methylprednis olone vs. modified-release prednisolone	12m	ESR				Changes in inflammatory markers were similar with either treatment. During the 1st month, MR-P significantly decreased IL-6 levels. Tapering was more rapid with MR-P than with 6-MP, and more MR-P patients could discontinue GC treatment altogether.
										CRP				
										Fibrinogen				
										IL-6	ELISA			
										Cortisol				
										TNFa				
										Tapering of GC dose				
										Time to discontinue treatment				
Binard 2008	Arthritis & rheumatism	Prospective cohort	Secondary care	PMR diagnosed by a rheumatologist in the absence of other diseases that might mimic PMR. Treated with GCs.	89	Mean age 74.6	Not given	Standard treatment	Variable	PMR-AS	PMR-AS	Yes	Has been evaluated	PMR-AS values 9.35 were 96.6% sensitive and 90.7% specific for flare diagnosis. An increase 6.6 in PMR-AS between 2 visits showed even better diagnostic performance.
										Presence of synovitis				
										Limb girdle pain				
										Fever				
										Nocturnal awakenings due to PMR				
										Weight loss				
										Flare diagnosed?				
										GC dose change?				
										VAS confidence in diagnosis of a flare				
Boiardi 2006	J Rheum	Prospective case control study	Secondary care	The diagnosis of PMR was confirmed if all the following	120 patients, 112 controls	Mean age of cases 72	73.2% F	Standard steroid treatment, same schedule	2 years	ESR				The 174 G/C promoter IL-6 polymorphism does not appear to be implicated in the susceptibility to developing isolated PMR. We observed in control subjects an association between CC genotype and
										CRP				
										IL-6	ELISA	Yes		
										Genotyping	PCR technique	Yes		
Caporali 2004	Ann Int Med	RCT	Secondary care	Chuang et al	72	Mean age 72	69% F in treatment group, 64% in placebo group	Methotrexate 10mg po vs placebo for 48w. All received folic acid and pred 25mg starting dose, tapered within 24w if controlled.	76 weeks	Number of patients no longer taking pred at 24, 48 and 72 weeks				Proportion of patients free of pred at 76w was higher in the mtx group than in pred + placebo group (effect only seen after a year of treatment). Fewer flares occurred in the mtx group.
										Relapses and recurrences	No instrument (defined flare as aching and stiffness at shoulder, hip or both, plus ESR>30mm/h or CRP >5mg/L			
										Cumulative pred dose				
										Duration of pred therapy				
										ESR				
										CRP				

Catanosos 2007	Arthritis & rheumatism	Pilot efficacy study	Secondary care	Healey	6 patients with longstanding PMR	Mean age 75	83% F	Etanercept twice weekly for 24w. Pred dose tapered at 4w to 5mg if remission attained then 2.5mg at 12w if still in remission	9 months	VAS pain				The results of this open pilot study suggest that etaner- cept could be useful as a steroid-sparing agent in PMR.
										Duration of morning stiffness				
										VAS patient global				
										VAS medical global				
										HAQ DI Italian version		Yes		
										Shoulder abduction scale				
										ESR				
										CRP				
										FBC, U+E, LFT				
										% response according to EULAR response criteria	Calculated from other measures	Yes		
										PMR-AS	Calculated from other measures	Yes		
										Shoulder USS		Yes		
Cawley 2017	Clin rheum	Inception cohort baseline data (cross sectional study)	Primary Care	Diagnosed by GP (read code recorded within the study time frame)	652	Mean age 73	62% F	None	N/A	Pain severity	NRS	Yes		in the earliest stages of PMR, over half of patients have some degree of functional limitation. Those PMR patients with high pain or stiffness severity, a high number of painful or stiff body sites and those with limitations in the shoulders report significantly poorer functional status.
										Stiffness severity	NRS	Yes		
										Pain site	Manikins			
										Stiffness site	Manikins			
										Ability to raise arms above head	Yes/No/Don't know question			
										Function	mHAQ	Yes		
										Anxiety	GAD	Yes		
										Depression	PHQ-8	Yes		
Cimmino 2006	Ann N Y Acad Sci	Prospective cohort	Secondary care (tertiary referral centres)	Chuang et al	80	Mean age 68 in F and 71 in M	65% F	Standard treatment (starting dose range 5-40mg, tapered by approx 20% monthly)	Mean 14.9 months	Physical examination	Fever, peripheral arhritis, TA, wt loss			The amount of inflammation didn't differ between F and M. However, F needed higher doses of pred and had more relapses.
										Intensity of pain	VAS			
										Duration of morning stiffness				
										Questionnaire to assess symptoms and health status		No		
										ESR				
										CRP				
										Hb				
Cimmino 2011	BMC MSK disorders	Prospective cohort	Secondary care (tertiary referral centres)	Bird et al	60	Mean age 71	58% F	12.5mg pred	1 month (chart r/v after 6m in addition to assess rate of exacerbations)	Standardised clinical examination		Described		78% responded to 12.5mg pred within 1m of treatment. Mean interval between initation of treatment and remission in responders was 6.6 days. The only factor predicting response to steriods was low weight.
										Pain intensity	NRS			
										Fatigue intensity	NRS			
										ESR				
										CRP				
										RF				
										USS shoulders (at baseline only - not really an outcome)				

Cimmino 2008	Clin exp rheum	5 year follow up after an RCT (see Caporali 2004)	Secondary care	Chuang et al	57	Mean age 78	47% F in mtx group, 44% F in controls	None	One off assessment 5 years post original RCT	Duration and dose of pred treatment				MTX-treated patients had less flare-ups of PMR and lower indexes of inflammation. How- ever, no difference in the incidence of steroid-related side effects was found between MTX-treated patients and controlsTwenty patients (35.1%) were still on steroids after a mean interval of 6.5 years from the beginning of the study. The patients who were still in need of steroid treatment purely for their PMR were 15 (26.3%) (one patient was excluded from this count because the original diagnosis was changed to rheu- matoid arthritis, another because he had developed temporal arteritis, and three other patients because
										Number of flares		Described		
										Pain	VAS			
										VAS physician				
										Function	HAQ	Yes	Has been evaluated	
										Adverse events	From a list			
										CRP				
										ESR				
Cleuziou 2012	J Rheum	Prospective cohort study - 89 patients attended for 149 visits therefore some cases cross-sectional, others	Secondary care	Rheumatologist diagnosed PMR in absence of other conditions mimicking PMR	89	Mean age 74.6	Not given	Standard treatment	No info - each visit treated as an episode with no linked data	PMR-AS	PMR-AS	Yes	Yes	PMR-AS may be helpful to tailor GC dosage in paitents with PMR in everyday practice. Dose tapering should be started when the PMR-AS falls below 10.
										Synovitis				
										Limb girdle pain				
										Fever				
										Nocturnal awakenings due to PMR				
										Weight loss				
										Physican global				
										Flare diagnosed?				
										Dose changed?				
Corrao 2009	Clin rheum	Case series	Secondary care	Chuang et al	9	76.7	89% F	Etanercept 25mg twice weekly with rapid tapering of pred	6m tight follow up, ongoing since until paper written	ESR				Prednisolone was withdrawn in all after the 6m treatment period and all were still in remission at 1y.
										CRP				
										Fasting glu				
										Aching at shoulders and hips	Pain VAS			
										Functional status	HAQ			
Cutolo 2017	RMD Open	RCT	Secondary care	Extended 2012 EULAR / ACR criteria including PMR VAS score >50 and CRP>2xULN	62	69	66% F	15mg MR or IR pred + placebo for 4w (bd dosing)	4w (outcomes measured twice daily)	PMR VAS				Non-inferiority wasn't proven. Even with low numbers though the results show a trend towards MR pred causing higher rates of complete response.
										Overall pain	VAS			
										Shoulder pain	VAS			
										Fatigue	VAS			
										Duration of morning stiffness				
										Time of medication intake				
										IL-6 (baseline, week 1 and week 4)				
										CRP (baseline, week 1 and week 4)				
										ESR (baseline, week 1 and week 4)				
Dasgupta 1998	Br J Rheum	RCT - 12w double blind placebo controlled, followed by an	Secondary care	Untreated PMR'. Patients were entered into the study if they had all	60 - 30 enrolled in each group	72 for PO PNL, 69 for IM MP	73% F PO PNL, 71% for IM MP	120mg i.m MP every 3w for 12w then reducing regime or OP	96w	Early morning stiffness duration				Remission rates were similar in both groups but cumulative pred dose in the MP group was 56% less than in the OP group. Weight gain and fracture rate was lower in the MP group (though numbers small). No sig difference in other side effects between the
										Pain	VAS			
										FBC				
										ESR				

Dasgupta 1991	Ann rheum dis	Prospective cohort	Secondary care	Jones and Hazelman	16	No info	No info	120mg im methylpred every 3w for 12w then reducing regime	12m	Morning stiffness duration				All participants achieved remission quickly. 3 patients had to revert to oral pred but of the others, the cumulative steroid dose over 12m was lower than with standard treatment and reported side effects were minimal. The hypothalamic pit axis wasn't suppressed after 12w of treatment.
										Pain	VAS			
										FBC				
										ESR				
										Immunoglobulins				
										Lymphcyte subsets				
										Hypothalamic pituitary axis	Diurnal cortisol rhythm and the metyrapone test			
DiMunno 1995	J Rheum	Open cross over design for comparison of daily vs alt daily regimens. Double blind RCT for comparing therapeutic effects of the different steroids	Secondary care	Pain and stiffness in prox muscle groups for 1-3m, age >50, ESR >40, no evidence of temporal arteritis / malignancy / infection, RF -ve.	31 - 15 in one group, 16 in the other	62 for DFZ, 67 for 6MP	67% F for DFZ, 71% F for 6MP	Randomly allocated to either deflazocort or methylpred for 12w. Half received alt daily dose initially then swapped to daily in a cross over design. Fixed dose of each for 2w then dose titrated to	12w	Shoulder and hip girdle p	VAS	Yes		Daily and alt daily doses of either deflazacort or methylpred were clinically equivalent in average response and trend of response though some people experienced more pain on the off day. The 2 drugs were clinically equivalent with either dosing regimen. 6-MP is more potent that deflazocort (ratio 1.78)
										Morning stiffness duration		No		
										Function (measured at baseline but not again, not reported)	American Rheumatism Association functional class	No		
										FBC				
										ESR				
										CRP				
										Plasma proteins, LFTs, fibrinogen, LDH, U+E, glucose				
										Urinalysis				
Diamantopoulos 2013	Biomed research int	Retrospective case series	Secondary care	EULAR / ACR cri	12	67.4	Not given	10mg leflunamide,	Up to 3 years	Dose of steroids				Mean duration of treatment was 10m. CRP reduced by 6 and mean pred dose by 3mg. Authors claim
										CRP				
Feinberg 1996	J Rheum	Prospective cohort	Secondary care	Age >50, ESR >40, Pain in 2/3 of shoulder girdle, hip girdle or paracervical musculature, No other signs of connective tissue disease	43	69.9	74% F	Mtx 7.5mg weekly, increased to 10mg then 12.5mg if no reponse. Pred dose was kept constant.	9m	ESR				Mtx didn't induce remission in any patient in this group and no-one was able to reduce they're pred dose (NB. Unusual sample)
										FBC				
										LFTs, glu, LDH				
										Clinical symptoms	Mentioned in results but no indication of how they were measured	No		
Ferraccioli 1996	J Rheum	Open RCT	Secondary care	Age >50, shoulder and / or pelvic girdle, pain and / or stiffness, morning stiffness >1h, no weakness on exam, symptoms for >4w, raised	24 - 12 in each group	67.4	83% F	10mg im MTX weekly + 25mg then reducing course pred vs 15mg then reducing course pred alone. Treated for 6m then pred stopped	12m	ESR				All patients were in remission at 12m. 50% of those in the MTX group stopped pred vs none in the pred group. Total pred dose was significantly less in the MTX group
										CRP				
										Bone G1a protein				
										Urine calcium and OH-Pro/creatinine ratio	Standard assays			
										BMD	DEXA			
										FBC				
										LFTs				
										Urinalysis				
										Mean daily pred dose				

Hutchings 2007	Arthritis care and research	Prospective cohort	Secondary care	Modified Jones and Hazelman criteria	129	70.9	59.7% F	Standard steroid treatment	12m	ESR				1/3 still had symptoms / raised inflammatory markers at 3w. Many had relapses and 81% had >1 AE. QoL was lower than popn norms and improved over the year. Higher levels of ESR, stiffness and pain were individually associated with lower physical QOL. Higher levels of all disease activity markers were individually associated with lower mental QOL. Higher levels of stiffness and pain were individually associated with worse HAQ scores, with proximal pain having the strongest association. Agreement on diagnosis was strong but there was a group for whom agreement was less consistent.
										CRP				
										QoL	mHAQ, SF-36 (mental and physical components)	Referenced	HAQ yes	
										Pain	VAS			
										Stiffness duration				
										Likelihood of PMR being correct diagnosis judged at 12m	Expert judgement, 2 raters			
Izumi 2015	RMD open	Retrospective case series	Secondary care	Bird's criteria and EULAR / ACR provisional classification criteria	13	74	84.6% F	Tocilizumab infusion every 4w in addition to prednisolone +/- methotrexate at whatever dose they had been taking	Up to 96w	Pain	VAS			Tocalizimab was well tolerated by all in this study. At 12weeks, we observed a significant glucocorticoid-sparing effect. No relapse was observed during the study period, and 8 of the 13 patients could discontinue prednisolone by the last follow-up. How best to taper tocizumab is not yet known.
										Patient global assessment	VAS			
										Physician global assessment	VAS			
										Function	HAQ-DI			
										ESR				
										CRP				
										Duration of morning stiffness				
Jimenez-Palop 2010	Annals of the rheumatic diseases	Prospective cohort	Secondary care	Clinical diagnosis by a rheumatologist (based on >1m pain and stiffness in neck and shoulders +/- in hips, with rapid response to 10-20mg	53	74	66% F	Standard steroid treatment	12w	Pain	VAS			69% had US inflammation in at least one bilateral site at baseline. No difference in US inflammatory findings was found between those with normal ESR and those with highly elevated ESR. Clinical, laboratory and US parameters decreased during F/U but there was no sig correlation i.e US findings seem to be an independent measure of disease activity in this study. The responsiveness of the US findings was better than that of clinical or lab markers. Intra and interobserver reliability was high.
										Morning stiffness				
										ESR				
										CRP				
										USS assessment	Standardised scanning protocol	Being tested		
Kalke 2000	Rheum	Prospective cohort	Secondary care	Jones and Hazelman	18	68.5	78% F	Some were on oral pred, some on methylpred. Standard	24w	Early morning stiffness				HAQ reduced at each time point and it's SRM values wer better than those of other parameters of disease activity. Sections on dressing, grooming and rising were the most responsive.
										Pain	VAS			
										CRP				
										Function	HAQ	Being tested		
										FBC				
Kreiner 2010	Arthritis research and therapy	Single centre double blind RCT	Secondary care	Chuang et al	40 - 20 patients, 20 controls. 10 treated, 10 placebo in each group	72 (patients)	65% F (patients)	4 doses of etanercept over 14 days or iv saline as placebo	14 days	PMR-AS				Effects of etanercept were modest - statisically sig reduction in PMR-AS but remained higher than in controls. HAQ didn't improve and amount of tramadol used didn't reduce significantly.
										ESR				
										TNF-alpha				
										IL-6				
										Function	HAQ	Referenced		
										Tramadol intake.				
Krogsgaard 1995	J Rheum	Double blind prospective study	Secondary care	Bird et al	30	75	63% F	20mg pred vs 24mg deflazacort, dose reduced according to response	12m	Muscle pain	Graded 0-3			5mg pred was equipotent to 7.5mg deflazocort at 6m and 8mg at 12m. Starting dose of 20mg pred induced remission in 15/16 patients.
										Muscle tenderness	Graded 0-3			
										Morning stiffness	Graded 0-3			
										ESR				
										Fibrinogen				
Lally 2016	Arthritis and rheum	Single centre open lable study (pilot efficacy study)	Secondary care	Healy criteria	10	68	50% F		15m	ESR				All 9 were able to come off pred by 6m with no relapses or recurrences throughout the F/U period
										CRP				
										PMR-AS		Referenced		
										HAQ-DI		Referenced		

Leeb 2003	Ann rheum dis	Prospective cohort	Secondary care	Bird / Wood criteria	Cohort 1 - 76 Cohort 2 - 24	Cohort 1 - 68.7 Cohort 2 - 71	Cohort 1 - 91% F Cohort 2 - 71% F	Standard steroid treatment	24w	ESR				All parameters other than serum Fe reduced over the study time period. 11 patients had 15 episodes of relapse. All those that changed were sig correlated with VAS pain and PGA. Multiple regression analysis with VAS pain as dependent variable showed only EUL and myalgia correlated with VAS pain. Core response set proposed. Each item in core set affects the response rate though EUL has the least influence. 50% improvement in each criteria agreed clinically meaningful by consensus.
										CRP				
										Alpha globulin and serum Fe				
										Pain	VAS			
										Physician global	VAS			
										Morning stiffness	Minutes			
										Muscle tenderness	0-3 scale			
										Self reported myalgia	0-3 scale			
										Ability to raise arms	0-3 scale			
Leeb 2004	Ann rheum dis	Prospective cohort followed by cross sectional validation study	Secondary care	All assessed against the 4 existing diagnostic criteria sets and then were diagnosed by an experienced clinician	Cohort 1 - 57, Cohort 2 - 24 for prospetive study. 53 patients for validation study.	Cohort 1 - 69.73 Cohort 2 - 71. Validation study - 67.3	Cohort 1 - 89% F Cohort 2 - 71% F. Validation study - 68% F	Standard steroid treatment	24w	PMR-AS				Authors conclude that they have demonstrated that PMR-AS is a valid and reliable tool. Better than PMR response criteria in that it doesn't rely on knowing baseline situation.
										Patient satisfaction	1-5 scale			
										Patient global asessment	VAS			
										ESR				
Leeb 2007	Arthritis & rheumatism	Cross sectional evaluation followed by a longitudinal study (classed as cohort)	Secondary care or private rheum clinic	Bird et al	Step 1 - 78 Step 2 - 39	Step 1 - mean age 66 Step 2 - mean age 68	Step 1 - 64% F Step 2 - 62% F	Standard treatment (2 received mtx)	Step 1 - one off assessment Step 2 - at least 2 assessments during 17m	PMR-AS	PMR-AS		Yes	Authors state that the study shows the PMR-AS is as reliable as VAS global and ESR for expressing disease activity and propose a cut off of 1.5 to suggest remission. Statistical methods questionable however.
										Patient satisfaction with disease status	Austrian school marking system	Reference given		
										ESR				
										Assessment of general health	VAS global			
Littman 1995	J Rheum	Double blind RCT	Not stated	Morning stiffness >30m, bilateral pain +/- stiffness in shoulder or hips for >1m and ESR >30, all responsive to steroid treatment	32 - 16 in each group	69.2 (tendinap), 64 (placebo) - significantly different	81% F (tendinap) 44% F (placebo)	120mg tendinap daily vs placebo. All received 10mg pred, reduced by 2.5mg every 3w	15w or until disease flared at which point they dropped out	ESR				Tendinap did have a steroid sparing effect in this group but side effects were high and drop out rate was high making overall sample size small
										CRP				
										FBC, U+E, LFT, urinalysis				
										Patient global	1-5 scale			
										Physician global	1-5 scale			
										Morning stiffness duration	Averaged for the week before assessment			
										Pain	VAS (0-32)			
										Muscle or joint stiffness	VAS (0-32)			
										Time to onset of fatigue for daily chores	Hours			
Macchioni 2009	Rheumatology	Prospective cohort	Secondary care	Healy criteria	57	74	81% F	12.mg pred tapered according to fixed schedule	Mean 41 months	ESR				At least one US sign of inflammation (subacromial / subdeltoid bursitis, long head biceps tenosynovitis or glenohumeral synovitis) was present in 98.2% patients and it was bilateral in 84%. Prevalence of US signs of inflammation was sig reduced by 24w. 59% of those in clinical remission at 24w still had US signs of inflammation. Frequency of relapses / recurrences was the same in those with or without residual inflammation. A postive power doppler signal at diagnosis (vascularity) correlated with increased
										CRP				
										USS				
										Leeb's DAS (PMR-AS)				

Mackie 2015	Ann Rheum Dis	Prospective cohort	Secondary care	Bird criteria (and all fulfilled ACR / EULAR criteria)	22 PMR patients, 16 RA controls	Median age 75 in one group, 78 In other	58% F	15mg pred increased to 20mg at 1m if clinically indicated then reducing regime	Median 2 years	Pain / stiffness location	Mannequins			A subset with characteristic extracapsular pattern of inflammation on MRI was more likely to feel 'back to normal' after GCs. IL-6 correlated with pelvic inflammation on MRI. The non-extracapsular inflammation group were more heterogenous.
										Pain severity	VAS			
										Stiffness severity	VAS			
										Fatigue severity	VAS			
										Function	HAQ-DI			
										PMR-AS				
										Back to normal question	5 point likert scale			
										Whole body MRI	Scored by 2 independent raters, classified as extracapsular or non-extracapsular pattern			
										ESR				
										CRP				
										PV				
										IL-6	ELISA			
Matteson 2012	J Rheum	Prospective cohort	Secondary care	EULAR / ACR classification criteria	85	72.6	60% F	15mg pred initially, standardised tapering regime	26w	CRP				All outcome measures improved from week 1 to week4. Only mHAQ showed sig improvement between weeks 4 and 26. Authors suggest a min set of outome measures consisting of patient reported global pain, hip pain, morning stiffness, physical function (MHAQ), mental functioan and an inflammatory marker should be used in practice and clinical trials. 16% and 10% didn't respond to steroids at 4 and 26w respectively. Test - retest reliability examined in a small subgroup (14 patients) was poor for fatigue VAS, mental function and morning stiffness.
										ESR				
										Physical examination				
										Pain	VAS			
										Fatigue severity	VAS			
										Morning stiffness duration	Minutes			
										Patient global	VAS (how is your PMR affecting you today?)			
										QoL	SF-36			
										Function	mHAQ			
										USS shoulders and hips				
McCarthy 2013	Rheumatology	Prospective cohort	Secondary care	New diagnoses - Jones and Hazelman. Group with existing and stable PMR also included. GCA wasn't	60	71.8	82% F	Newly diagnosed group were commenced on 15mg pred and continued on this for 6w (not standard)	6w	PMR-AS		Referenced		Fibrinogen was more specific than CRP or ESR in confirming active disease and response to treatment in this study.
										CRP				
										ESR				
										Fibrinogen				
McCarthy 2014	J Rheum	Prospective cohort - same cohort as above study	Secondary care	New diagnoses - Jones and Hazelman. Group with existing and stable PMR also included. GCA wasn't excluded but only 1	60	71.8	82% F	Newly diagnosed group were commenced on 15mg pred and continued on this for 6w (not standard)	6w	PMR-AS		Referenced		VASDA and VASQOL were more responsive to changes in disease activity than VAS pain / morning stiffness / PMF-AS or mHAQ. On the mHAQ, items relating to rising, dressing and grooming showed the greatest responsiveness to changes in disease activity. PMR-AS correlated well with all PRO. Of the biomarkers, fibrinogen correlated most strongly with the PRO.
										CRP				
										ESR				
										Fibrinogen				
										Patient assessment of disease activity	VAS	No	No	
										Patient assessment of QoL	VAS	No	No	
										Function	mHAQ			

Migliore 2005	European review for medical and pharmacologic al sciences	Pilot study of efficacy	Secondary care	ACR criteria	7	72	100% F	infliximab infusion at baseline, 15d, 4w and 16w. Then mtx to maintain remission.	Mean 8m	ESR CRP HbA1c Clinical symptoms	 ? (mentioned in results but no info on what was assessed)			Clinical symptoms improved in all, ESR and CRP improved in 5 out of 7 by week 6. HbA1 didn't deteriorate.
Palard-Novello 2016	Eur J Nuc Med Mol Imaging	Prospective open label study	Secondary care	Chuang et al	18	67.8	33% F	TCZ at baseline, wk 4 and wk 8. If no response at wk 8, classified as non-responder and given pred. From week 12, everyone was given pred at dose dependent on PMR-AS.	12w	PMR-AS CRP ESR F-FDG PET / CT	 	Reference Described	 Being tested	Abnormal uptake was found in ischial tuberosities, hips and shoulders in sig majority at baseline. This sig reduced by week 2 and week 12. However, there was no correlation bettween change in uptake and change in PMR-AS/CRP/ESR. Authors conclude that F-FDG PET / CT could reflect disease activity and might be useful in evaluation of response to treatment.
Pulsatelli 2008	Arthritis and rheum	Prospective cohort	Secondary care	Based on Healy criteria 1)persistent pain at last 2 areas 2) ems >1HR 3) Rapid pred response	93 PMR patients, 48 controls	Median 74	74% F	Pred at median starting dose of 17.5mg per day. No further info.	24m	ESR CRP Relapses Physical examination sgp130 sIL-6R Hb	 Not reported			There was no difference in sIL-6R or sgp130 levels between patients and controls at baseline or during follow up. Higher sIL-6R levels correlated with higher numbers of relapses.
Pulsatelli 2010	Clin Exp Rheum	Prospective cohort	Secondary care	Healey	93 PMR patients, 48 controls	74	74% F	Pred at mean starting dose of 17.5mg per day. No further info.	24m	PTX3	ELISA		No	Levels of PTX3 were not significantly different between patients and controls at disease onset or during follow up.
Salvarani 2003	J Rheum	Pilot study	Secondary care	Healey	4	65.8	100% F	Infusion of infliximab at weeks 0, 2 and 6. Pred at 5mg daily for 1st 2w then withdrawn if in remission.	12m	ESR CRP IL-6 Symptoms and signs of PMR	 Not specified			3 of the 4 were symptom free with normal ESR and CRP at 1 year. The acute phase reactants including IL-6 fell by week 2. Infliximab was well tolerated by all.
Salvarani 2005	Arthritis and rheum	Prospective cohort	Secondary care	Healey. Those with ESR <40 but who satisfied all the other criteria were also included. RF +ve and GCA excluded	94	Median - 74	75% F	Pred at mean starting dose of 17.5mg per day, adjusted according to response then tapered according to	Mean 39m	ESR CRP IL-6 Clinical symptoms and signs Relapse / remission	 Little info Symptoms and raised ESR / CRP			CRP was elevated in 98.9% at diagnosis whereas ESR was elevated in 91.5%. Those with raised ESR at diagnosis had higher risk of >1 relapse / recurrence. Those with persistently elevated CRP or IL-6 had higher rates of relapse / recurrence. No patients had persisently elevated ESR.

Salvarani 2000	J Rheum	Double blind placebo controlled RCT	Secondary care	Healey	20 (10 in each group)	70 in treatment group, 71 in control group	80% F in treatment group, 60% F in control group.	Bilateral shoulder injections weekly for 4w - either saline or 6-methylpred	6m	Systenic signs / symptoms (fever, wt loss, anorexia)				All those in the treatmet group responded initially and maintained this for the first 4w whilst having regular injections. Half withdrew within 4w of the last injection due to recurrence of symptoms.
										ESR				
										CRP				
										IL-6				
										Morning stiffness duration				
										Pain	VAS			
										Patient assessment	VAS			
										Physician assessment	VAS			
										Bilateral shoulder MRI (on 5 patients only)				
Salvarani 2007	Ann int med	Double blind placebo controlled RCT	Secondary care	Healey	51	71 in both groups	70% in treatment group, 54% in control group (quite a difference)	Standard pred treatment reducing from 15mg - 0mg over 16w for all. Either infliximab or placebo infusions at 0, 2, 6, 14 and 22w.	52w	CRP				Proportion of patients who were free of relapse or recurrence at 52w didn't differ between groups. Duration of pred therapy and cumulative pred dose didn't differ between groups
										ESR				
										Signs and symptoms of PMR (aching and stiffness at shoulder or hip girdle or both)	Physical examination and a questionnire	No - no details		
										Disability	Italian version of HAQ-DI	Referenced		
										FBC, U+E, LFT, ANA				
										No relapse / recurrence				
Viapiana 2015	Rheum Int	RCT	Secondary care	Jones and Hazelman	52	71 in pred group. 75 in methylpred group	77% F in pred group. 71% F in methylpred group.	25mg pred or 20mg methylpred. Tapered according to fixed dosing	12m	ESR				At 2w, 89% of those in pred group were in remission vs 100% of methylpred group. Difference was accounted for by 3 patients in the pred group who took longer to achieve remission. Changes in ACTH and cortisol were similar between the 2 groups.
										CRP				
										Fibrinogen				
										Cortisol and ACTH				
Weyand 1999	Archives of int med	Prospective cohort	Not stated	Diagnosis of PMR was based on the presence of: (1)morning stiffness of >30 mins; (2)pain in the shoulders and/or arms, hips and/or	30	Not given	Not given	Starting dose of 20mg pred, increased to 30mg if no response then tapered by 2.5mg every 2w.	At least 6m after treatment stopped. Median 32m.	Physical examination				3 subsets of patients were identified by rapidity of response and numbers of flares. Pretreatment ESRs were helpful in identifying those that required low doses of steroids for <1 year and the response pattern of IL-6 to steroids identified those with a chronic relapsing course and those who only had partial response to an initial dose of 20mg pred. Authors suggest stratifying treatment by certain clinical criteria (initial ESR and Il-6 and initial response to treatment).
										Morning stiffness duration				
										Pain severity	VAS			
										Stiffness severity	VAS			
										Patient global	1-5 scale			
										Physician global	1-5 scale			
										ESR				
										IL-6				
										FBC				
Devauchelle-Pensec 2016	Ann Rheum Dis	Prospective longitudinal stuy	Secondary care	Chuang's criteria	20	Median 70	35% F	3 iv infusions of tocilizumab at baseline, 4 then 8w. Then pred from weeks 12-24 in either low dose if PMR-AS <10 of standard dose	24w	PMR-AS				All achieved primary endpoint of PMR-AS<10 by week 12. None required rescue therapy or had a flare in either treatment phase. Sig GC sparing effect noted. Adverse events rate were high - 37 events in 14 patients.
										Fatigue	VAS			
										Global disease activity	VAS			
										Pain	VAS			
										QoL	SF-36			
										MRI shoulders and pelvis				
										FDG PET-CT				
										USS				

Appendix 4.5: Risk of bias assessment spreadsheet

Study	Clearly defined study objective	Appropriate design for study question	QUIPS domain 1 - study participation						RoB domain 1	QUIPS domain 2 - study attrition					RoB domain 2	QUIPs domain 4 - outcome measurement			RoB domain 4	Additional considerations for RCTs			RoB RCTs
			Adequate participation by eligible persons *(response rate / sample size)*	Description of the population of interest *(criteria used for diagnosis)*	Description of the baseline study sample *(number of participants, age range, gender distribution)*	Sampling frame and recruitment	Period and place of recruitment *(primary or secondary care / non-clinical setting)*	Inclusion and exclusion criteria *(to give a representative sample)*		Adequate response rate *(and adequate duration of planned follow up for study question)*	Attempts to collect info from drop outs	Reasons for loss to follow up given	Adequate description of those lost to follow up	No imp differences between those who completed it and those that didn't		Clear definition of the outcome *(and instrument)* is provided	Method of outcome measure used is valid and reliable *(instrument validity described or referenced)*	Method and setting of outcome measurement is the same for all study participants		Randomisation process adequate	Blinding process adequate	Were the groups treated equally throughout	
Alvarez-Rodriguez 2010	**Yes** - to identify the cellular source of circulating cytokines and the state of activation of the different peripheral blood mononuclear cells (PBMC) in patients with PMR. We also analysed the influence of CS treatment on circulating cytokines.	**Yes** - prospective case control study	Process not explained	Clear and appropriate	Controls age matched but gender proportions are different.	No info on how ppts were recruited.	No info.	Appropriate as far as info is given. No info on where controls came from.	**Moderate**	No info about the 34 pts and 17 controls beyond the description of baseline characterisitcs and the 1st dose of pred. Post treatment sample was only taken from 14 patiets after mean duration of 9.7w.	No info	No info	No info	No info	**High**	Yes	Yes	Unclear	**Moderate**	N/A	N/A	N/A	**N/A**
Benucci 2015	**Yes** - to evaluate the changes in inflammation markers and their correlations with cortisol levels after treatment with 6-methylprednisolone (6-MP) or MR-P in patients with "early" PMR.	Yes - non-randomized, prospective observational study. RCT would have been better but reasonale pragmatic alternative	No detail about recruitment	Clear and appropriate	Given clearly in a table - appropriate. Very low baseline CRP.	No info.	No info.	No information given.	**High**	No info about response rate but duration of F/U was reasonable at 12m. However results are presented at specitic time points with no justification for this given.	No info.	No info	No info	No info	**High**	Yes	Yes	Yes	**Low**	N/A	N/A	No - dosing and tapering schedules were different with no explanation of this given.	
Binard 2008	**Yes** - to evaluate the effectiveness of the PMR-AS for diagnosing disease flares	**Yes** - prospective cohort	No info on response rate. Good sample size.	Reasonable - no specific diagnostic criteria utilised - "defined PMR as a diagnosis by a Rheumatologist"	Partial	Consecutive patients who met study criteria	Secondary care	Only info given is diagnosis of PMR	**Moderate**	N/A	N/A	N/A	N/A	N/A	**N/A**	Yes	Yes	Yes though Involved some subjective assessment by one of a group of physicians	**Moderate**	N/A	N/A	N/A	**N/A**
Biordi 2006	**Yes** - assessment of the role of an IL-6 polymorphism in the susceptibility to, and severity of, PMR. Investigating whether the –174 G/C promoter polymorphism of IL-6 might modulate the circulating level of IL-6 and the risk of relapse / recurrence	**Yes** - prospective case-control	Yes, all patietns diagnosed over a 5 yr period	Clear and appropriate	Clear table - appropriate	Clear	Secondary care	All patients with diagnosis of PMR included. Exclusion criteria given.	**Low**	Yes	No info	No info	No info	No info	**High**	Yes	Yes	Yes	**Low**	N/A	N/A	N/A	**N/A**
Caporali 2004	**Yes** - to compare the efficacy and safety of pred plus mtx and pred alone in patients with PMR	**Yes** - RCT	Yes	Clear and appropriate	Clear table - appropriate	Clear	Secondary care	Appropriate	**Low**	Yes	Yes	Yes	Partial	Unclear - 14% discontinued or lost to follow up	**Moderate**	Yes	Yes	Yes	**Low**	Yes	Yes	Yes	**Low**
Catanoso 2007	**Yes** - to investigate whether etanercept has a steroid-sparing effect in the treatment of patients with relapsing polymyalgia rheumatica (PMR).	Pilot study of efficacy	Yes for a pilot study	Specified	Clear table - appropriate	N/A	Secondary care	Inclusion criteria were the following: relapsing PMR, at least 12 months prednisone treatment, presence of corticosteroid adverse events, and inability to reduce prednisone dosage below 7.5 mg/day. Exclusion criteria specified and appropriate	**Low (pilot study)**	Yes	N/A	N/A	None lost to follow up	N/A	**Low**	Yes	Yes	Yes	**Low**	N/A	N/A	N/A	**N/A**

Study	Clearly defined study objective	Appropriate design for study question	QUIPS domain 1 - study participation						RoB domain 1	QUIPS domain 2 - study attrition					RoB domain 2	QUIPs domain 4 - outcome measurement			RoB domain 4	Additional considerations for RCTs			RoB RCTs
			Adequate participation by eligible persons *(response rate / sample size)*	Description of the population of interest *(criteria used for diagnosis)*	Description of the baseline study sample *(number of participants, age range, gender distribution)*	Sampling frame and recruitment	Period and place of recruitment *(primary or secondary care / non-clinical setting)*	Inclusion and exclusion criteria *(to give a representative sample)*		Adequate response rate *(and adequate duration of planned follow up for study question)*	Attempts to collect info from drop outs	Reasons for loss to follow up given	Adequate description of those lost to follow up	No imp differences between those who completed it and those that didn't		Clear definition of the outcome *(and instrument)* is provided	Method of outcome measure used is valid and reliable *(instrument validity described or referenced)*	Method and setting of outcome measurement is the same for all study participants		Randomisation process adequate	Blinding process adequate	Were the groups treated equally throughout	
Alvarez-Rodriguez 2010	**Yes** - to identify the cellular source of circulating cytokines and the state of activation of the different peripheral blood mononuclear cells (PBMC) in patients with PMR. We also analysed the influence of CS treatment on circulating cytokines.	**Yes** - prospective case control study	Process not explained	Clear and appropriate	Controls age matched but gender proportions are different.	No info on how ppts were recruited.	No info.	Appropriate as far as info is given. No info on where controls came from.	**Moderate**	No info about the 34 pts and 17 controls beyond the description of baseline characterisitcs and the 1st dose of pred. Post treatment sample was only taken from 14 patiets after mean duration of 9.7w.	No info	No info	No info	No info	**High**	Yes	Yes	Unclear	**Moderate**	N/A	N/A	N/A	**N/A**
Benucci 2015	**Yes** - to evaluate the changes in inflammation markers and their correlations with cortisol levels after treatment with 6-methylprednisolone (6-MP) or MR-P in patients with "early" PMR.	Yes - non-randomized, prospective observational study. RCT would have been better but reasonale pragmatic alternative	No detail about recruitment	Clear and appropriate	Given clearly in a table - appropriate. Very low baseline CRP.	No info.	No info.	No information given.	**High**	No info about response rate but duration of F/U was reasonable at 12m. However results are presented at specitic time points with no justification for this given.	No info.	No info	No info	No info	**High**	Yes	Yes	Yes	**Low**	N/A	N/A	No - dosing and tapering schedules were different with no explanation of this given.	
Binard 2008	**Yes - to** evaluate the effectiveness of the PMR-AS for diagnosing disease flares	**Yes** - prospective cohort	No info on response rate. Good sample size.	Reasonable - no specific diagnostic criteria utilised - "defined PMR as a diagnosis by a Rheumatologist"	Partial	Consecutive patients who met study criteria	Secondary care	Only info given is diagnosis of PMR	**Moderate**	N/A	N/A	N/A	N/A	N/A	**N/A**	Yes	Yes	Yes though Involved some subjective assessment by one of a group of physicians	**Moderate**	N/A	N/A	N/A	**N/A**
Biordi 2006	**Yes** - assessment of the role of an IL-6 polymorphism in the susceptibility to, and severity of, PMR. Investigating whether the --174 G/C promoter polymorphism of IL-6 might modulate the circulating level of IL-6 and the risk of relapse / recurrence	**Yes** - prospective case-control	Yes, all patietns diagnosed over a 5 yr period	Clear and appropriate	Clear table - appropriate	Clear	Secondary care	All patients with diagnosis of PMR included. Exclusion criteria given.	**Low**	Yes	No info	No info	No info	No info	**High**	Yes	Yes	Yes	**Low**	N/A	N/A	N/A	**N/A**
Caporali 2004	**Yes** - to compare the efficacy and safety of pred plus mtx and pred alone in patients with PMR	**Yes** - RCT	Yes	Clear and appropriate	Clear table - appropriate	Clear	Secondary care	Appropriate	**Low**	Yes	Yes	Yes	Partial	Unclear - 14% discontinued or lost to follow up	**Moderate**	Yes	Yes	Yes	**Low**	Yes	Yes	Yes	**Low**
Catanoso 2007	**Yes** - to investigate whether etanercept has a steroid-sparing effect in the treatment of patients with relapsing polymyalgia rheumatica (PMR).	Pilot study of efficacy	Yes for a pilot study	Specified	Clear table - appropriate	N/A	Secondary care	Inclusion criteria were the following: relapsing PMR, at least 12 months prednisone treatment, presence of corticosteroid adverse events, and inability to reduce prednisone dosage below 7.5 mg/day. Exclusion criteria specified and appropriate	**Low (pilot study)**	Yes	N/A	N/A	None lost to follow up	N/A	**Low**	Yes	Yes	Yes	**Low**	N/A	N/A	N/A	**N/A**

Dasgupta 1998	**Yes** - to compare the efficacy and safety of im methylprednisolone acetate with oral prednisolone	**Yes** - RCT	No info on response rate, sample size reasonable	Described	Clear table - appropriate	Not clear	Several secondary care sites	Patients were entered into the study if they had all of the following: shoulder and pelvic girdle muscular pain in the absence of true muscle weakness, morning stiffness > 30 min, ESR > 30 mm/h, absence of rheumatoid or other inflammatory arthritis or malignant disease, absence of signs of inflammatory muscle disease, normal serum CK and TSH. Exclusion criteria specified and appropriate.	**Moderate**	No info on response rate, duration of follow up reasonable	No info	Yes	Reasons for withdrawal given but patients who withdrew not described / compared to those who didn't. Numbers of withdrawals similar in each group though	Not certain (though numbers same in each group)	**High**	Yes	Yes	Not clear	**Moderate**	Not clear	Yes	Yes	**Moderate**
Dasgupta 1991	**Yes** - to evaluate the effects of im methylpred in newly diagnosed PMR over 12m	Yes - prospective corhot	No info on response rate. Small sample size but reasonable for pilot study of efficacy and tolerability	Described, appropriate	No info	No info	No info	Jones and Hazelman criteria to diagnose PMR, GCA excluded	**High**	No info	No info	Yes	No	Unclear	**High**	Yes	Yes	Not clear	**Moderate**	N/A	N/A	N/A	**N/A**
DiMunno 1995	**Yes** - to compare clinical efficacy and equivalence of daily vs alt daily daflazacort and methylpred and to determine the potency ratio of the 2 steroids	**Yes** - open cross over design for 1st part, double blind RCT for 2nd	No info on response rate. Small sample size.	Described	Described, appropriate	No info	Seconary care	Criteria to diagnose PMR clear. Excluded many comorbidities so not very representative	**Moderate**	No info on response rate. Follow up only 12w.	Data from drop-outs included in safety analysis but not in efficacy analysis	Yes	No	Unclear (and 2 who dropped out early weren't included in the analysis)	**Moderate**	Yes	Yes for those that were measured as outcomes of efficacy	Not stated deinitively	**Moderate**	Yes	Yes	Yes	**Low**
Diamantopoulos 2013	Yes - to explore the role of leflunamide as a steroid sparing agent in GCA and PMR	**Not the best option** - retrospective case series	N/A	Described, appropriate	Described - by definition a 'difficult to treat' group	Appropriate	Secondary care	EULAR / ACR criteria for PMR. Difficult to treat disease or had a flare when reducing steroids or mtx. Had to be on >5mg pred.	**Moderate** (though not really applicable to case series study)	N/A	Yes - data from drop outs included in analysis	Yes (reasons for discontinuing treatment given)	Yes	Unclear	**Low** (as far as applies)	Yes	Yes	Yes	**Low**	N/A	N/A	N/A	N/A
Feinberg 1996	Yes - to examine the efficacy of mtx in treating PMR without GCA	Prospective cohort - reasonable as an exploratory study	Yes, reasonable sample size	Clearly described	Clearly given	Clearly explained, appropriate	Secondary care	Criteria for PMR reasonable. Had to also continue to have DM, glaucoma or osteoporosis and still have symptoms and raised ESR despite 20mg / day of pred (high threshold). Exclusion criteria appropriate.	**Moderate** (unusually difficult to treat sample)	Yes	Yes	Yes	Reasons for drop-outs given but not description of participants who dropped out	Unclear	**Moderate**	No	Yes for ESR, others not described	Unclear	**High**	N/A	N/A	N/A	**N/A**
Ferraccioli 1996	**Yes** - to report the effects of mtx plus pred vs pred alone in PMR	Yes - open RCT	No info on response rate. Small sample size.	Clearly described	Clear table	No info beyond being recruited from an OP rheum clinic	Secondary care	Criteria for PMR reasonable. Excluded if suspicion of GCA. Patients with RA, SLE and paraneoplastic conditions also excluded.	**Moderate**	No info on response rate, reasonable duration of F/U	No drop outs	N/A	N/A	N/A	**Low**	Yes	Yes	Yes	**Low**	Not explained	N/A	No - pred reducing regimes were different with a higher starting dose in the mtx group	**High**

Hutchings 2007	**Yes -** 1) to evaluate the impact of PMR on clinical outcomes and QOL in the first year; 2) to examine the relationship between laboratory measures and clinical outcomes, and changes in QOL; 3) to evaluate agreement between rheumatolo- gists in confirming the initial diagnosis of PMR after 1 year of followup.	**Yes** - prospective cohort	Screened 249, included 129. Reasons for exclusion given.	Clearly described	Clearly described	GPs referred patients with suspected PMR prior to starting steroids.	Secondary care - 8 rheum clinics	Jones and Hazelman criteria for diagnosis of PMR. Symptoms of GCA, infection, advanced OA, abnormal CK / TSH or prior steroid therapy excluded.	**Low**	Yes	Not clear whether info from drop outs was included in the regression analysis	Yes	Yes	Analysis was done with those not considered to have PMR at 12m as well as without this group and little changed.	**Moderate**	Yes	Yes, referenced	Not definitively stated	**Low**	N/A	N/A	N/A	**N/A**
Izumi 2015	**Yes** - to assess the effectivenss and safety of tocilizumab in intractable PMR	**Acceptable** - retrospective case series	N/A	Clearly described	Clearly described	Consecutive patients who had been treated with tocilizumab during the	Secondary care	Diagnosed with PMR and had then been started on tocilizuab for relapse	**Not really applicable here**	Followed up for up to 96w	None discontinued treatment during the observation period	N/A	N/A	N/A	**Low**	Yes	Some referenced	Not stated	**Low**	N/A	N/A	N/A	**N/A**
Jimenez-Palop 2010	**Yes** - to assess the sensitivity to change of US inflammatory findings in patients with PMR treated with steroids	**Yes** - prospective cohort	Reasonable	Clearly described	Described for the 59 recruited but not for the 53 actually analysed	Consecutive patients diagnosed in the partcipating centres	Secondary care	PMR diagnosed by a rheumatologist. No other clinically evident MSK disease.	**Moderate**	6 excluded after initial recruitement - reasons given. Follow up only 12w so only early stage disease evaluated	N/A	Yes	Yes	No	**Low**	Yes	Some referenced	Yes same sonogropher within each centre, standard protocol, intraobserver reliability assesssed	**Low**	N/A	N/A	N/A	**N/A**
Kalke 2000	**Yes** - to evaluate the HAQ in assessment of functional status, responsiveness to change and correlation with conventional disease activity indices in PMR	**Yes** - prospective cohort	No info on response rate, small sample size.	Described	Minimal info	No info	No info	Newly diagnosed (Jones and Hazelman criteria) and untreated PMR. No GCA, malignancy, infection, connective tissue disease or advanced OA.	**Moderate**	No info on response rate, follow up 24 weeks so only tested the HAQ in early - middle phase of the disease	N/A	3 withdrawn due to changes in diagnosis	N/A	N/A	**N/A**	Yes	HAQ referenced	Yes	**Low**	N/A	N/A	N/A	**N/A**
Kreiner 2010	**Yes** - to determine the therapeutic potential of TNF-alpha receptor blockade in PMR	**Yes** - single centre double blind RCT	Small but adequate sample size from their power calculation	Clearly described	Clearly described	Patients with suspected PMR referred to the study clinic. Controls recruited via newspaper ad.	Secondary care	Chuang criteria to diagnose PMR. Excluded if prior use of steroids, GCA, infection, uncontrolled DM / BP / heart failure, other inflammatory diseases, cancer, abnormal TSH or Ca.	**Low**	Very short follow up though justification given	N/A	Yes	N/A	No	**Low**	Yes - though odd to choose tramadol to use and measure as an outcome	Yes	Yes	**Moderate**	Yes	Yes	Yes	**Low**
Krogsgaard 1995	**Yes** - to establish the antiinflammatory equipotency between pred and deflazacort	**Yes** - prospective study, double blind	No info on response rate, reasonable sample size	Clearly described	Clearly described	Consecutive patients with newly diagnosed PMR by Bird et al criteria.	Secondary care	Excluded those with GCA, cander, kidney or liver disease, GI surgery or bone disease.	**Moderate**	No info on response rate, reasonable F/U for the question	Unclear - not clear whether data from dropouts was included in the analysis	Yes	No	Unclear	**Moderate**	Yes	Reasonable - though all clinician assessed rather than patient reported	Yes - same assessor	**Low**	N/A	Yes	N/A	N/A

Lally 2016	**Yes** - to assess the efficacy and safety of tocilizumab in newly diagnosed PMR	**Yes** - single centre open-label study (pilot efficacy study)	No info on response rate, small sample size but reasonable for exploratory study	Clearly described	Described	Consecutive patients meeting the criteria. Those that declined / didn't meet all criteria were used as a comparator group - ?bias in this though baseline characteristics	Secondary care	Healy criteria to diagnose PMR. Enrolled within 1m of diagnosis and had to have had <20mg pred daily. GCA, other inflammatory arthropathy, connctive tissue disease, +ve RF / CCP ab all excluded.	**Moderate**	No info on response rate, F/U reasonable	No	Yes - 1 dropped out due to mild infusion reaction	No	Unclear	**Moderate**	Yes	Yes	Not definitively stated - clinician involved in assessing knew whether treatment or control	**High**	N/A	N/A	N/A	**N/A**
Leeb 2003	Yes - to develop response criteria for PMR for monitoring treatment and comparing different treatment regimens	**Yes** - prospective cohort	Reasonable	Clearly described	Described	213 patients from 8 centres across Europe were enrolled into a diagnostic criteria study. From this 76 were enrolled into the response criteria study.	Pan-European research group study - not clear how recruited	PMR diagnosed by Bird / Wood criteria and available for long term follow up. RA, shoulder OA or spondylitis, spondylosis, SLE, connective tissue disease, myopathies, cancer and PD excluded.	**Moderate**	Reasonable response rate. Fairly short follow up but pragmatic decision in light of drop outs	No	No	No	Unclear	**High**	Yes	Reasonable	Multisite and some subjective measures	**Moderate**	N/A	N/A	N/A	**N/A**
Leeb 2004	Yes - to develop a composite score for measurement of disease activity in PMR and assess its int and ext validity	Reasonable - prospective cohort initially then cross sectional validation study	Reasonable	Clearly described for cohort. Not described for validation study participants.	Described	From the 76 enrolled into the above response criteria study, 57 were followed up for 24w. 2nd cohort of 24 were recruited from an	Pan-European research group study - not clear how recruited.	PMR diagnosed by Bird / Wood criteria and available for long term follow up. RA, shoulder OA or spondylitis, spondylosis, SLE, connective tissue disease, myopathies, cancer and PD excluded.	**Moderate**	Reasonable response rate. Fairly short follow up but pragmatic decision in light of drop outs from previous enrollment	N/A	N/A	N/A	N/A	**N/A**	Yes	Reasonable / being tested	Multisite and some subjective measures	**Moderate**	N/A	N/A	N/A	**N/A**
Leeb 2007	**Yes** - to confirm the reliabilty and applicability of the PMR-AS and establish a threshold for remission	Cross sectional evaluation followed by a longitudinal study (classed as cohort)	Little info	Clear and appopriate	Described and reasonable	Little info	Secondary care / private clinic (?biased sample).	Bird criteria, GCA excluded clinically	**High**	No info	N/A	N/A	No info	No info	**Moderate**	Yes	Outcomes chosen for comparison weren't standard (PATSAT).	No info - ?were administered by treating / study physician	**High**	N/A	N/A	N/A	**N/A**
Littman 1995	**Yes** - to determine whether tendinap has a steroid sparing effect in PMR	**Yes** - double blind placebo controlled RCT	No info on response rate, small sample size	Described	Described - sig difference between groups	Multisite, no other info given.	Multisite, no other info on recruitment given	Aged 50-80 with diagnosis of PMR in the last 6m. Controlled on 10mg pred. Exclusion - evidence of other rheumatic diseases	**High**	No info on response rate, short duration of F/U but reasonable to test initial efficacy.	Yes	Yes	Yes - table of all data given including those who dropped out	Unclear	**Moderate**	Yes	Some were, some unusual outcomes	Not definitively stated	**High**	Yes	Yes	Yes	**Low**
Macchioni 2009	**Yes** - to determine if USS and power doppler is useful in identifying relapsing PMR	**Yes** - prospective cohort	Good response rate, reasonable sample size	Described	Clearly described	All patients diagnosed with PMR over an 18m period in a rheum clinic.	Secondary care, 18m.	Healy PMR criteria, GCA or previous steroid treatement excluded. Anyone who developed RA during course of study also excluded	**Low**	High response rate, reasonable follow up	No drop outs	N/A	N/A	N/A	**Low**	Yes	Yes	Yes - same 2 assessors throughout, blinded to initial findings	**Low**	N/A	N/A	N/A	**N/A**
Mackie 2015	**Yes** - to determine whether whole-body MRI defines clinically relevant subgroups within polymyalgia rheumatica (PMR) including glucocorticoid responsiveness	**Yes** - prospective cohort	No info on response rate, small but reasonable sample size	Described	Described	Consecutively diagnosed patients from one clinic	Secondary care	Bird criteria for diagnosis, previously untreated. At least one of ESR / CRP / PV elevated. -ve RF / antiCCP	**Low**	Not clear	N/A	N/A	N/A	N/A	**N/A**	Yes	Yes	Yes	**Low**	N/A	N/A	N/A	**N/A**

Matteson 2012	**Yes** - to prospectively evaluate the disease course and the performance of clinical, PRO and musculoskeletal ultrasound measures in patients with PMR	**Yes** - prospective cohort (sample of the same cohort used for classification criteria development study)	No info on response rate, reasonable sample size	Described - just the subset of the cohort who met the proposed classification criteria	Described	Not clear how recruited	Recruited at 21 rheum clinics in 10 European countries and the US	New-onset PMR - age ≥ 50, new-onset bilateral shoulder pain, and no corticosteroid treatment (for any condition) within the 12w prior, who fulfilled all the inclusion criteria [i.e., morning stiffness > 45 min, raised markers of CRP and/or ESR) and exclusion criteria at presentation [i.e., no infection, active cancer, GCA, or clinical features of the common PMR mimics] and in accord with expert clinician judgment of the participating investigator that the patient had PMR. Diagnosis re-evaluated at each F/U visit.	**Low**	No info on response rate, reasonable F/U	No	No	No	Unclear	**High**	Yes	Yes	Standardised data collection forms, translated into native language.	**Low**	N/A	N/A	N/A	**N/A**
McCarthy 2013	**Yes** - to establish whether plasma fibrinogen was a superior biomarker of disease activity in active PMR than the standard biomarkers, ESR and CRP.	**Yes** - prospective cohort	No info on response rate, reasonable sample size	Described	Described, appropriate	One rheum clinic, recruitment process not described	Secondary care, one site	Group of newly diagnosed patients (Jones and Hazelman criteria) and group of people with existing, stable and inactive PMR. Patients were excluded if they had either +ve RF and/or anti-CCP, a concomitant diagnosis of another CTD, systemic infection, abnormal levels of CK or TSH or suspected underlying malignancy. People with GCA could be included but only 1 was (hence included in the r/v).	**Moderate**	No info on response rate, short F/U	No drop outs	N/A	N/A	N/A	**Low**	Yes	Measures referenced. ?pain VAS was physician recorded...	Yes	**Moderate**	N/A	N/A	N/A	**N/A**

McCarthy 2014	**Yes** - to prospectively examine the responsiveness of a number of PRO measures in PMR, as well as their relationship to the biomarkers ESR, CRP and plasma fibrinogen.	Yes - prospective cohort (same cohort as above study)	No info on response rate, reasonable sample size	Described	Described, appropriate	One rheum clinic, recruitment process not described	Secondary care, one site	Group of newly diagnosed patients (Jones and Hazelman criteria) and group of people with existing, stable and inactive PMR. Patients were excluded if they had either +ve RF and/or anti-CCP, a concomitant diagnosis of another CTD, systemic infection, abnormal levels of CK or TSH or suspected underlying malignancy. People with GCA could be included but only 1 was (hence included in the r/v).	**Moderate**	No info on response rate, short F/U	No drop outs	N/A	N/A	N/A	**Low**	Yes	Some measures referenced, some being tested	Yes	**Low**	N/A	N/A	N/A	**N/A**
Migliore 2005	To test infliximab as a steroid sparing agent in patients with PMR plus diabetes / osteoporosis	**Acceptable** - pilot study of efficacy	Very small sample but only set up as a pilot study. No info on how recruited.	Described.	Described.	No info	No info	Newly diagnosed PMR (ACR criteria). Either DM or osteorposis in addition. Infection excluded.	**High**	No info	N/A	N/A	N/A	N/A	**N/A**	Yes	Limited measures selected, not clear what 'clinical symptoms' were measured or how	Not clear	**High**	N/A	N/A	N/A	**N/A**
Pallard-Novello 2016	**Yes** - to evaluate the use of F-FDG PET / CT for the assessment of tocilizumab as a first line treatment in PMR	**Yes** - longitudinal prospective clinical trial (part of the TENOR trial)	18 of the 21 enrolled in the TENOR trial had the imaging study	Described.	Described	No info about how recruited to TENOR trial	Secondary care - 2 rheum clinics	Chuang et al criteria for PMR, diagnosis in last 12m, aged 50-80, PMR-AS > 10, ESR >40 or CRP > 10, no other inflammatory rheum or CT disease. GCA. Infection, malignancy and severe OA excluded.	**Moderate**	No info on response rate, reasonable duration of F/U for question	No	No. One withdrew consent, one didn't have imaging ?why. 2 only had baseline scan.	No	Unclear	**High**	Yes	Described and referenced	Clinical exam performed by experienced clinician - ?same for all patients	**Low**	N/A	N/A	N/A	**N/A**
Pulsatelli 2008	**Yes** - to investigate the modulation of systemic levels of soluble interleukin-6 receptor (sIL-6R) and soluble gp130 (sgp130) in untreated and treated polymyalgia rheumatica (PMR) patients in order to evaluate the relationship of these molecules with clinical outcome and their feasibility to provide a prognostic tool in clinical practice.	**Yes** - prospective cohort	No info on response rate, sample size good (93).	Described. No exclusion criteria stated.	Clear table	Consecutive, untreated patients with PMR. No further info and no info on where controls came from.	Secondary care	Diagnosis based on Healy criteria. No exclusion criteria stated.	**Moderate**	No info on response rate, reasonable duration of F/U for question	No info	No info	No info	No info	**High**	Yes	For some outcomes	Yes	**Moderate**	N/A	N/A	N/A	**N/A**
Pulsatelli 2010	**Yes** - to evaluate serum long pentraxin PTX3 feasibility as a prognostic marker in PMR	**Yes** - prospective cohort (same cohort as above)	No info on response rate, sample size good.	Described. No exclusion criteria stated.	Clear table	Consecutive, untreated patients with PMR. No further info and no info on where controls came from.	Not stated	Diagnosis based on Healy criteria. No exclusion criteria stated.	**Moderate**	No info on response rate, reasonable duration of F/U for question	No info	No info	No info	No info	**High**	Yes	For some outcomes	Yes	**Moderate**	N/A	N/A	N/A	**N/A**

Pulsatelli 2010	**Yes** - to evaluate serum long pentraxin PTX3 feasibility as a prognostic marker in PMR	**Yes** - prospective cohort (same cohort as above)	No info on response rate, sample size good.	Described. No exclusion criteria stated.	Clear table	Consecutive, untreated patients with PMR. No further info and no info on where controls came from.	Not stated	Diagnosis based on Healy criteria. No exclusion criteria stated.	**Moderate**	No info on response rate, reasonable duration of F/U for question	No info	No info	No info	No info	**High**	Yes	For some outcomes	Yes	**Moderate**	N/A	N/A	N/A	**N/A**
Salvarani 2003	**Yes** - to investigate if infliximab has a steroid sparing effect in people with PMR who are resistant to steroid therapy and have steroid side effects	Reasonable - pilot study	Very small sample (n=4) - pilot study	Described	Described individually	Selected sample from one clinic	Secondary care	PMR relapsed on reduction of pred to 7.5 - 12.5mg pred and multiple vertebral fractures	**High**	N/A	N/A	N/A	N/A	N/A	**N/A**	No	No	Unknown	**High**	N/A	N/A	N/A	**N/A**
Salvarani 2005	**Yes** - to determine lab parameters that may be useful to identify people with PMR who require long term steroid therapy	**Yes** - prospective cohort	Yes, all diagnosed people from 2 secondary carecinics, reasonable sample size	Described	Clear table	All patients diagnosed with PMR over a 4 yr period in 2 Italian secondary care clinics	Secondary care	Diagnosis by Healey criteria. Excluded if RA / RF+ve / GCA.	**Low**	Yes, all patients. Assessed monthly for 6m then 3 monthly thereafter (NB survival bias)	N/A	No drop outs	N/A	N/A	**Low**	Yes for acute phase reactants but no detail on other assessments	Not for all	Assessed by same rheumatologist, standardized data colleiton form.	**Moderate**	N/A	N/A	N/A	N/A
Salvarani 2000	**Yes** - to determine the efficacy and safety of shoulder steroid injections in PMR	**Yes** - double blind placebo controlled RCT	Response rate not clear. Small sample size.	Diagnostic inclusion criteria described but no further info. Excluding those with pelvic involvement skews the sample	Described	Consecutive patients' - no further info given	Secondary care	Healey's criteria for diagnosis. Excluded if on anticoag treatment, bleeding disorders, pelvic girdle involvement, GCA or peripheral synovitis.	**High**	No info on response rate. Follow up reasonable.	No - no data presented from drop outs i.e the whole control group, other than baseline data.	Yes - non response	No	Unclear	**High**	Yes	Yes	No - only 5 had MRI	**Moderate**	Process not explicitly stated	Yes	No. Non-responders dropped out and this was everyone in the placebo group. Also MRI only done in the first 5 participants	**High**
Salvarani 2007	**Yes** - to compare the efficacy of pred + infliximab with pred + placebo in newly diagnosed PMR	**Yes** - double blind placebo controlloed RCT	No info on response rate, sample size reasonable	Well described	Described - some differences between groups	7 rheum clinics in Italy	Secondary care	Diagnosis by Healey criteria. Exclusions: previous steroids, biological agents, or immunosuppressive agents; GCA, RA, pleuritis, pericarditis, leukopenia, or thrombocytopenia and ANAs indicating SLE / other CTD; presence of myositis, hypothyroidism, and psoriatic arthritis; presence of uncontrolled diabetes or hypertension, infection or neoplasm; and presence of active or inactive TB	**Moderate**	No info on response rate, follow up adequate	All patients who completed follow up included in the analysis. Best and worst case scenarios analysed to assess effects of missing data.	Yes	Reasons given but no other info	Uncertain	**Moderate**	No, not for all	Unclear	PI at each site made the assessment (7 sites)	**Moderate**	Yes	Yes	Yes	**Low**
Viapiana 2015	**Yes** - to compare methylpred to pred in terms of its clinical response and its effect on HPA axis in patients initiating GL treatment for PMR	**Reasonable - randomised trial** but not blinded	Sample size reasonable by their power calculation but not clear what modelling assumption their sample size was for	Diagnostic criteria described. No info on who gets referred to the clinic	Described - large BMI difference between groups	All those referred to the clinic with new onset PMR	Secondary care	Jones and Hazelman criteria for diagnosis. Excluded those who didn't' show prompt response to steroids and those with GCA, previous steroids or CI to steroids.	**Moderate**	No info on response rate. Follow up reasonable.	No - per protocol anaysis	Yes	Reasons given but no other info	Only 2 dropouts and these were before study started.	**Low**	Yes for outcomes stated but mentions symptoms being assessed and no further info on this	Yes but no clinical / functional assessment specified, purely biochemical	Unclear - see previous comment re 'symptoms'.	**High**	Randomly assigned in 1:1 ratio. No futher detail.	N/A	Yes	**High**

Weyand 1999	**Yes** - to determine whether clinical or laboratory parameters in PMR could be identified that allow for stratifying patients into subsets with differences in corticosteroid requirements.	**Yes** - prospective cohort	No info on response rate, small sample size	Diagnostic inclusion criteria described but no further info.	Not given	No info	Not stated	Diagnosis of PMR was based on the presence of: (1)morning stiffness of >30 mins; (2)pain in the shoulders and/or arms, hips and/or thighs, neck, and/ or torso for >1m and (3)an ESR of >40 mm/h. Patients with an ESR <40 mm/h and a typical presentation for PMR were enrolled if the diagnosis was independently confirmed by a second rheumatologist. GCA excluded.	**High**	No info on response rate. Reasonable follow up.	No	Yes	Reasons given but no other info	Different in that a different diagnosis emerged so appropriately excluded	**Moderate**	Yes	Uncertain	No info	**Moderate**	N/A	N/A	N/A	**N/A**
Devauchelle-Pensec 2016	**Yes** - to evaluate the efficacy and safety of first- line tocilizumab in PMR.	**Yes** - open label prospective longtitudinal study (pilot efficacy study)	No info on response rate, small sample size but adequate by their power calculation and exploratory study	Diagnostic criteria described, no other info on recruitment	Described	2 uni hospitals in France	Secondary care, no other info	PMR by Chuang's criteria, onset in past 12m. PMR-AS >10 and either no previous GC treatment or GC for <1m, stopped 7 days prior to study. Aged 50-80, ESR >40 or CRP >10 and no evidence of any other inflammatory rheumatic or CTD. GCA and various other comorbidities excluded	**Moderate**	No info on response rate, follow up short but adequate for question	N/A	N/A	N/A	N/A	**Low**	Yes	Accepted measures chosen	Yes	**Low**	N/A	N/A	N/A	**N/A**

Appendix 4.6: Summary of data extraction and risk of bias assessment of included studies

Study (lead author and year)	Study Aim	Study design	Outcome and instruments by domain				Disease activity / global assessment	Imaging	Other outcomes measured	Risk of bias rating -domain 1 (study participation)	Risk of bias rating - domain 2 (study attrition)	Risk of bias rating - domain 4 (outcomes assessment)	Risk of bias rating - RCTs
			Markers of systemic inflammation	Pain	Stiffness	Physical function							
Alvarez-Rodriguez 2010	To identify the cellular source of circulating cytokines and the state of activation of the different peripheral blood mononuclear cells (PBMC) in patients with PMR. We also analysed the influence of CS treatment on circulating cytokines.	Prospective case control study	IL-6						Other cytokines	Moderate	High	Moderate	N/A
Bennuci 2015	To evaluate the changes in inflammation markers and their correlations with cortisol levels after treatment with 6- methylprednisolone (6-MP) or MR-P in patients with "early" PMR.	Non-randomised prospective observational study	ESR, CRP, fibrinogen, IL-6						Cortisol, TNF-alpha, tapering of GC dose, time to discontinue treatment	High	High	Low	N/A
Binard 2008	To evaluate the effectiveness of the PMR-AS for diagnosing disease flares	Prospective cohort	CRP	VAS	Duration of MS	EUL (0-3 scale)	PMR-AS		Presence of synovitis, fever, nocturnal awakenings due to PMR, weight loss, VAS confidence in diagnosis of a flare	Moderate	N/A	Moderate	N/A
Boiardi 2006	Assessment of the role of an IL-6 polymorphism in the susceptibility to, and severity of, PMR. Investigating whether the –174 G/C promoter polymorphism of IL-6 might modulate the circulating level of IL-6 and the risk of relapse / recurrence	Prospective case control study	ESR, CRP, IL-6						Genotyping	Low	High	Low	N/A
Caporali 2004	To compare the efficacy and safety of pred plus mtx and pred alone in patients with PMR	RCT	ESR, CRP						Questionnaire to assess health status (unspecified), physical examination, relapses and recurrences, cumulative pred dose, duration of treatment	Low	Moderate	Low	Low
Catanoso 2007	To investigate whether etanercept has a steroid-sparing effect in the treatment of patients with relapsing polymyalgia rheumatica (PMR).	Pilot efficacy study	ESR, CRP	VAS	Duration of MS	HAQ DI (Italian version), shoulder abduction scale (0-3)	VAS patient global, VAS physician global, PMR-AS	Shoulder USS	FBC, U+E, LFT	Low (pilot study)	Low	Low	N/A

Cawley 2017	To examine the relationship between different characteristics of pain and stiffness and the functional status of newly diagnosed patients with PMR	Inception cohort baseline data		Severity NRS, site (mannekin)	Severity NRS, site (mannekin)	mHAQ, ability to raise arms above head (Y/N/DK question)			Anxiety (GAD), depression (PHQ-9)	Low	Low	Low	N/A
Cimmino 2006	To assess if there are gender-related differences in PMR activity at presentation and if women are more resistant to treatment than men.	Prospective cohort	ESR, CRP	Intensity VAS	Duration of MS				Physical examination, questionnaire to assess symptoms and health status	Moderate	Moderate	Moderate	N/A
Cimmino 2011	To test if 12.5mg is an adequate starting dose and evaluate the clinical predictors of drug response	Prospective cohort	ESR, CRP	Intensity NRS				Shoulder USS (at baseline only)	Fatigue intensity NRS, RF, standardised clinical examination	Low	Low	Low	N/A
Cimmino 2008	To assess the incidence of long term side effects of mtx + pred vs pred alone treatment for PMR as well as assessing natural history of treated PMR.	5 year follow up after an RCT (see Caporali 2004) (essentially a cohort study)	ESR, CRP	VAS		HAQ	VAS physician		Duration and dose of pred treatment, number of flares, adverse events	Low	Low	High	N/A
Cleuziou 2012	To evaluate the usefulness of the PMR-AS in guiding adjustment of GC dosage	Prospective cohort study - 89 patients attended for 149 visits therefore some cases cross-sectional, others longituidinal "	CRP	VAS	Duration of MS	EUL (0-3 scale)	PMR-AS, physician global VAS		Synovitis, fever, nocturnal awakenings due to PMR, weight loss, flares, dose changes	Moderate	N/A	Moderate	N/A

Corrao 2009	To explore if etanercept has a steroid sparing effect in patients with PMR and decompensated diabetes	Case series	ESR, CRP, IL-6	VAS		HAQ			Fasting glu	Low (though not really applicable to case series study)	N/A	Low	N/A
Cutolo 2017	To evaluate the efficacy and safety of MR vs IR pred in newly diagnosed PMR	RCT		Overall pain VAS, shoulder pain VAS	Duration of MS		PMR VAS		Fatigue VAS, time of medication intake	Moderate	Moderate	Low	Low
Dasgupta 1998	To compare the efficacy and safety of im methylprednisolone acetate with oral prednisolone	RCT	ESR	VAS	Duration of MS				FBC	Moderate	High	Moderate	Moderate
Dasgupta 1991	To evaluate the effects of im methylpred in newly diagnosed PMR over 12m	Prospective cohort	ESR	VAS	Duration of MS				FBC, Immunoglobulins, lymphocyte subsets, hypothalamic-pit axis (diurnal cortisol rhythm and the metyrapone test)	High	High	Moderate	N/A
Devauchelle-Pensec 2016	To evaluate the efficacy and safety of first- line tocilizumab in PMR	Pilot efficacy and safety study	CRP	VAS		SF-36	Global disease activity VAS	MRI shoulders and pelvis, FDG PET-CT, USS	Fatigue VAS	Moderate	Low	Low	N/A
Dimunno 1995	To compare clinical efficacy and equivalence of daily vs alt daily daflazacort and methylpred and to determine the potency ratio of the 2 steroids	Open cross over design for comparison of daily vs alt daily regimens. Double blind RCT for comparing therapeutic effects of the different steroids	ESR, CRP, fibrinogen	VAS	Duration of MS	American Rheumatism Association functional class (only measured at baseline)			FBC, plasma proteins, LFTs, LDH, U+E, glu, urinalysis	Moderate	Moderate	Moderate	Low
Diamantopoulos 2013	To explore the role of leflunamide as a steroid sparing agent in GCA and PMR	Retrospective case series	CRP						Dose of steroids	Moderate (though not really applicable to case series study)	Low (as far as applies)	Low	N/A
Feinberg 1996	To examine the efficacy of mtx in treating PMR without GCA	Prospective cohort	ESR						FBC, LFTs, LDH, glu, clinical symptoms (unclear how assessed)	Moderate	Moderate	High	N/A
Ferraccioli 1996	To report the effects of mtx plus pred vs pred alone in PMR	Open RCT	ESR, CRP						Bone G1a protein, urine calcium and OH-pro/creatinine ratio, BMD, FBC, LFTs, urinalysis, mean daily pred dose	Moderate	Low	Low	High

Hutchings 2007	1) to evaluate the impact of PMR on clinical outcomes and QOL in the first year; 2) to examine the relationship between laboratory measures and clinical outcomes, and changes in QOL; 3) to evaluate agreement between rheumatolo- gists in confirming the initial diagnosis of PMR after 1 year of followup.	Prospective cohort	ESR, CRP	VAS	Stiffness durati	mHAQ, SF-36 (mental and physical components)			Likelihood of PMR being correct diagnosis judged at 12m (expert judgement)	Low	Moderate	Low	N/A
Izumi 2015	To assess the effectiveness and safety of tocilizumab in intractable PMR	Retrospective case series	ESR, CRP	VAS	Duration of MS	HAQ-DI	Patient and physician global assessment VAS			Not really applicable here	Low	Low	N/A
Jimenez-Palop 2010	To assess the sensitivity to change of US inflammatory findings in patients with PMR treated with steroids	Prospective cohort	ESR, CRP	VAS	Duration of MS			USS assessment (standard scanning protocol - being tested)		Moderate	Low	Low	N/A
Kalke 2000	To evaluate the HAQ in assessment of functional status, responsiveness to change and correlation with conventional disease activity indices in PMR	Prospective cohort	CRP	VAS	Duration of MS	HAQ			FBC	Moderate	N/A	Low	N/A
Kreiner 2010	To determine the therapeutic potential of TNF-alpha receptor blockade in PMR	Single centre double blind RCT	CRP, ESR, TNF-alpha, IL-6	VAS	Duration of MS	HAQ, EUL 0-3	PMR-AS, physician global		Tramadol intake	Low	Low	Moderate	Low
Krogsgaard 1995	To establish the antiinflammatory equipotency between pred and deflazacort	Prospective observational study, double blind	ESR, fibrinogen	Muscle pain (physician graded 0-3)	Morning stiffness (physician graded 0-3)				Muscle tenderness (physician graded 0-3)	Moderate	Moderate	Low	N/A
Lally 2016	To assess the efficacy and safety of tocilizumab in newly diagnosed PMR	Pilot efficacy and safety study	ESR, CRP	VAS	Duration of MS	HAQ-DI, EUL 0-3	PMR-AS, physician global VAS			Moderate	Moderate	High	N/A
Leeb 2003	To develop response criteria for PMR for monitoring treatment and comparing different treatment regimens	Prospective cohort	ESR, CRP	Pain VAS and self reported myalgia (0-3 scale)	Duration of MS	Ability to raise arms (0-3 scale)	Physician global VAS		Alpha-globulin, serum iron, muscle tenderness (graded 0-3)	Moderate	High	Moderate	N/A
Leeb 2004	To develop a composite score for measurement of disease activity in PMR and assess its int and ext validity	Prospective cohort followed by cross sectional validation study	CRP, ESR	VAS	Duration of MS	EUL 0-3	PMR-AS, physician global VAS, patient global assessment VAS		Patient satisfaction (0-5 s	Moderate	N/A	Moderate	N/A
Leeb 2007	To confirm the reliabilty and applicability of the PMR-AS and establish a threshold for remission	Cross sectional evaluation followed by a cohort study	CRP, ESR	VAS	Duration of MS	EUL (0-3 scale)	PMR-AS. assessment of general health (VAS global)		Patient satisfaction with disease staus (Austrian school marking system)	High	Moderate	High	N/A
Littman 1995	To determine whether tendinap has a steroid sparing effect in PMR	Double blind RCT	ESR, CRP	VAS (0-32)	Duration of MS, severity of stiffness VAS (0-32)		Patient global (1-5 scale), physician global (1-5 scale)		FBC, U+E, LFT, urinalysis, time to onset of fatigue for daily chores	High	Moderate	High	Low
Macchioni 2009	To determine if USS and power doppler is useful in identifying relapsing PMR	Prospective cohort	ESR, CRP	VAS	Duration of MS	EUL (0-3 scale)	PMR-AS, physician global	USS		Low	Low	Low	N/A
Mackie 2015	To determine whether whole-body MRI defines clinically relevant subgroups within polymyalgia rheumatica (PMR) including glucocorticoid responsiveness	Prospective cohort	ESR, CRP, PV, IL-6	VAS, location (mannequins)	Duration of MS, VAS, location (mannequins)	HAQ-DI, EUL (0-3 scale)	PMR-AS, physician global	Whole body MRI	Back to normal question (5 point likert scale)	Low	N/A	Low	N/A

Matteson 2012	To prospectively evaluate the disease course and the performance of clinical, PRO and musculoskeletal ultrasound measures in patients with PMR	Prospective cohort	ESR, CRP	VAS	Duration of MS	SF-36, mHAQ	Patient global VAS (how is your PMR affecting you today?)	USS shoulders and hips	Physical examination, fatigue severity VAS	Low	High	Low	N/A
McCarthy 2013	To establish whether plasma fibrinogen was a superior biomarker of disease activity in active PMR than the standard biomarkers, ESR and CRP	Prospective cohort	ESR, CRP, fibrinogen	VAS	Duration of MS	EUL (0-3 scale)	PMR-AS, physician global			Moderate	Low	Moderate	N/A
McCarthy 2014	To prospectively examine the responsiveness of a number of PRO measures in PMR, as well as their relationship to the biomarkers ESR, CRP and plasma fibrinogen	Prospective cohort	ESR, CRP, fibrinogen	VAS	Duration of MS	mHAQ, EUL (0-3 scale)	PMR-AS, physician global, patient assessment of disease activity VAS		Patient assessment of QoI	Moderate	Low	Low	N/A
Migliore 2005	To test infliximab as a steroid sparing agent in patients with PMR plus diabetes / osteoporosis	Pilot efficacy study	ESR, CRP						HbA1c, clinical symptoms (not defined)	High	N/A	High	N/A
Palard-Novello 2016	To evaluate the use of F-FDG PET / CT for the assessment of tocilizumab as a first line treatment in PMR	Prospective open label study	ESR, CRP	VAS	Duration of MS	EUL (0-3 scale)	PMR-AS, physician global	F-FCG PET / CT		Moderate	High	Low	N/A
Pulsatelli 2008	To investigate the modulation of systemic levels of soluble interleukin-6 receptor (sIL-6R) and soluble gp130 (sgp130) in untreated and treated polymyalgia rheumatica (PMR) patients in order to evaluate the relationship of these molecules with clinical outcome and their feasibility to provide a prognostic tool in clinical practice	Prospective cohort	ESR, CRP						Physical examination, relapses, sgp130, sIL-6R, Hb	Moderate	High	Moderate	N/A
Pulsatelli 2010	To evaluate serum long pentraxin PTX3 feasibility as a prognostic marker in PMR	Prospective cohort							PTX3	Moderate	High	Moderate	N/A
Salvarani 2003	To investigate if infliximab has a steroid sparing effect in people with PMR who are resistant to steroid therapy and have steroid side effects	Pilot study	ESR, CRP, IL-6						Symptoms and signs of PMR (not specified)	High	N/A	High	N/A
Salvarani 2005	To determine lab parameters that may be useful to identify people with PMR who require long term steroid therapy	Prospective cohort	ESR, CRP, IL-6						Symptoms and signs of PMR (not specified), relapse and remission	Low	Low	Moderate	N/A
Salvarani 2000	To determine the efficacy and safety of shoulder steroid injections in PMR	Double blind placebo controlled RCT	ESR, CRP, IL-6	VAS	Duration of MS		Patient and physician global assessment VAS	Bilateral shoulder MRI (5 patients only)	Systemic symptoms / signs (fever, wt loss anorexia)	High	High	Moderate	High
Salvarani 2007	To compare the efficacy of pred + infliximab with pred + placebo in newly diagnosed PMR	Double blind placebo controlled RCT	ESR, CRP			HAQ-DI (Italian version)			Relapse / recurrence, signs and symptoms of PMR (aching and stiffness at shoulder or hip girdle or both) - physical examination and a questionnaire, FBC, U+E, LFT, ANA	Moderate	Moderate	Moderate	Low

Viapiana 2015	To compare methylpred to pred in terms of its clinical response and its effect on HPA axis in patients initiating GL treatment for PMR	RCT	ESR, CRP, fibrinogen						Cortisol and ACTH	Moderate	Low	High	High
Weyand 1999	To determine whether clinical or laboratory parameters in PMR could be identified that allow for stratifying patients into subsets with differences in corticosteroid requirements	Prospective cohort	ESR, IL-6	VAS	Duration of MS and severity VAS		Physician global (1-5 scale), patient global (1-5 scale)		Physical examination, FBC	High	Moderate	Moderate	N/A

Appendix 5.1: Search strategies for evaluation of evidence regarding measurement properties of candidate instruments

Searches carried out in Medline via OVID in June 2018

Search strategy for PMR and VAS / NRS and duration of morning stiffness

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
11	((numeric* and rating and (scale or score)) or numeric scale or nrs or nprs).mp.
12	((visual and analogue and (scale or score)) or visual scale or VAS).af.
13	duration of morning stiffness.mp.
14	morning stiffness duration.mp.
15	11 OR 12 OR 13 OR 14
16	10 AND 15

Search strategy for the HAQ

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
11	“health assessment questionnaire”.mp
12	HAQ.mp
13	mHAQ.mp
14	HAQ-DI.mp
15	11 OR 12 OR 13 OR 14
16	10 AND 15

Search strategy for ESR and CRP

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
11	“erythrocyte sedimentation rate”.mp
12	ESR.mp
13	“c-reactive protein”.mp
14	CRP.mp
15	11 OR 12 OR 13 OR 14
16	10 AND 15

Appendix 7.1: Published papers on development work carried out prior to this PhD

"I suddenly felt I'd aged": A Qualitative Study of Patient Experiences of Polymyalgia Rheumatica (PMR)

Helen Twohig, Caroline Mitchell, Christian Mallen, Adewale Adebajo, Nigel Mathers.

ABSTRACT

Objectives To explore patient experiences of living with, and receiving treatment for, PMR.
Methods Semi-structured qualitative interviews, with 22 patients with PMR recruited from general practices in South Yorkshire. Thematic analysis using a constant comparative method, ran concurrently with the interviews and was used to derive a conceptual framework.
Results 5 key themes emerged highlighting the importance of: 1) pain, stiffness and weakness, 2) disability, 3) treatment and disease course, 4) experience of care, 5) psychological impact of PMR. Patients emphasised the profound disability experienced that was often associated with fear and vulnerability, highlighting how this was often not recognised by health care professionals. Patients' experiences also challenge medical convention, particularly around the concept of 'weakness' as a symptom, the use of morning stiffness as a measure of disease activity and the myth of full resolution of symptoms with steroid treatment. Treatment decisions were complex, with patients balancing glucocorticoid side effects against persistent symptoms.
Conclusions Patients often described their experience of PMR in terms of disability rather than focussing on localised symptoms. The associated psychological impact was significant.
Practice implications Recognising this is key to achieving shared understanding, reaching the correct diagnosis promptly, and formulating a patient-centred management plan.

1. INTRODUCTION

Polymyalgia rheumatica (PMR) is the most common inflammatory rheumatic condition in people aged over 50 with an incidence of 1 in 1000 in this age group and a lifetime risk of 2.4% for women and 1.7% for men [1,2]. It is characterised by pain and stiffness in the hips and shoulders, raised inflammatory markers and response to glucocorticosteroids, although atypical presentations can occur in up to 20% of those affected [3,4]. PMR has a major impact on quality of life [5] and treatment with corticosteroids is associated with a high rate of adverse effects [6]. Despite this, it remains an under-researched and poorly understood condition with the lack of primary care research particularly notable considering that the majority of PMR is diagnosed and managed in primary care [7].

Patients with PMR require frequent, comprehensive clinical assessments. At each consultation assessment of disease activity and response to treatment is needed, as well as evaluation of treatment side effects and assessment for complications [8]. Exploring and understanding the patient experience of PMR as an 'illness' is crucial in order to facilitate shared decisions about treatment, balancing symptom control and functional enablement against adverse effects of steroid therapy. Much of the research into PMR to date however focuses on a biomedical model of 'the disease' and current clinical assessment therefore tends to be set in this paradigm.

There is increasing emphasis in many areas of health care on patient reported outcome measures (PROMS) as one tool to help in the drive to achieve the goal of person-centred care. Only by exploring patient experiences can the outcomes which are meaningful to patients be identified. For example, in rheumatoid arthritis, an appreciation of the significance of fatigue was first identified through qualitative exploration [9,10] and it is now recommended that fatigue is measured in addition to the core outcome set in all clinical trials of the condition [11].

There is work being done towards agreeing a core set of outcome measures for use in clinical trials of PMR [12]. However, there are no measures available which assess outcomes directly from the perspective of a patient with the condition. A PROM developed specifically for PMR would contribute greatly to a comprehensive assessment of the condition. The first step in developing a PROM is to determine the conceptual framework through qualitative studies of the target population [13].

We therefore set out to explore patient experiences of living with, and receiving treatment for, PMR with the dual aims of enhancing understanding of the condition from the patient perspective and allowing derivation of a conceptual framework for future development of a PROM.

2. METHODS

Ethical approval for this study was obtained from the Dyfed Powys Research Ethics Committee (REC 12/WA/0344, 15/11/12).
Participants were recruited from 10 general practices from South Yorkshire. A purposive sampling strategy was used to recruit practices which were diverse according to their Index of Multiple Deprivation score, list size and training status.
Patients aged 50 years and over with a Read coded PMR diagnosis and classical PMR symptoms (documented in the electronic medical record as having bilateral shoulder and

/ or pelvic girdle pain and stiffness for at least 2 weeks, and evidence of an acute phase response (raised ESR / CRP)) were included.
Patients with atypical features (e.g. normal ESR / CRP), were eligible if their diagnosis had been made by a rheumatologist. Patients were excluded if they had significant dementia or memory impairment, a primary diagnosis of giant cell arteritis, a concomitant inflammatory arthropathy, active cancer or if the GP decided that participation wasn't appropriate (e.g. other terminal illness).
An invitation letter and study information sheet were sent to suitable patients and if they wished to participate they replied directly to the research team. Reminder letters were sent 2 weeks later to those that had not replied to the initial invitation.
A topic guide (see appendix 1) was developed, informed by discussion with members of a PMR patient support group, a literature review and consultation with the study multidisciplinary advisory group. Topics included in the initial guide were onset of the condition, symptoms and functional effects, diagnosis, flares and relapses, starting and stopping treatment, resolution of the condition and information provision. An open questioning style was used with minimal prompts to allow themes to emerge naturally [14]. Interviews were conducted by either HT or CaM, in participants' homes or in the Academic Unit of Primary Medical Care (University of Sheffield) according to participant preference. After the interviews, patients' notes were reviewed by HT to gather data on comorbidities, ESR / CRP results and steroid dose regimes.
Interviews were taped, independently transcribed and then systematically analysed using a constant comparative method to establish themes grounded in the data [15]. NVivo10 software was used to manage the data. Analytic codes and categories were developed through an iterative, thematic and self-conscious process, beginning in parallel with the data collection and informing subsequent interviews as concepts and themes emerged. The process of constant comparison continued until theoretical saturation was reached and no new themes were emerging.
Two researchers (HT and CaM) analysed the data independently and any differences were considered and discussed until agreement was reached. A third researcher (NM) moderated a selection of interviews to ensure comprehensiveness and consistency of identified themes.
10 practices took part in recruitment, with 7 of these identifying patients suitable for inclusion. Recruitment ranged from 0-7 patients per practice.
43 patients were invited to participate. There were 18 non-responders and 3 patients (all male) who agreed to take part but weren't required for interview as data saturation had been reached.
12 men and 10 women were interviewed. 2 patients were excluded post-interview (one had his diagnosis revised to inflammatory arthritis during the course of his illness and one had extensive co-morbidities and could not distinguish the effects of PMR from other conditions). The age range of participants was 53-81 years and the range of time from diagnosis to interview was 5 months to 2 years 3 months. 3 had been referred to secondary care at some stage in the course of their condition and the rest had been managed entirely in primary care. (see appendix 2 for table of participant details).

3. RESULTS

5 key themes were identified which were all interlinked and related. A conceptual framework was developed which reflected the relationship between the themes and subthemes (see appendix 3).

Theme 1: Pain, stiffness and weakness

"I could hardly move in bed, it was aching all down my back and I just felt, I suddenly felt I'd aged, like I were about 80 year old, that's what it felt like. And very stiff, very achy like when you turned over in bed it was painful." UPN 16

There was significant heterogeneity in symptoms described by participants. Some described severe pain whilst others described muscle ache, likened to that caused by flu or vigorous exercise. In others, stiffness predominated and pain was mentioned secondarily to this.

Although weakness is not a widely accepted symptom of PMR, and is not part of the recent classification criteria,[1] several patients used the term. In most cases, with greater elaboration, it became clear that the term 'weakness' was being used to describe limited function due to pain or stiffness. However, a few participants were certain that they were experiencing true weakness.

The majority of participants experienced variation in their symptoms through the day though there were a few who said that their pain and stiffness was constant. Some did describe a classical morning stiffness pattern but most painted a more nuanced picture of diurnal variation with worsening of symptoms after periods of rest or after any significant activity.

Box 1 – Pain, stiffness and weakness

"And I really screamed in pain. You know, to get dressed. Or even to lift my arms up. The pain was terrible." *UPN 18*

"Well it's not pain, it were more of a bad ache and I couldn't do much, you know." *UPN 13*

"..the shoulders and the biceps... they felt very weak… they weren't painful, just wouldn't work" *UPN 5*

"When you first have PMR, it used to take me til about tea-time to actually come round. And even when I started on the prednisolone, I didn't sort of come round straight away as I've told you. But that's when I noticed the prednisolone was working, that the pain was -, I was freer much earlier in the day." *UPN 2*

"I would say my best time is 10 o'clock while 3 and then I seem to get really tired. I think it's when you've done most of what you want to do and then you sit down and then I kind of seize up." *UPN 3*

Theme 2: Disability

"I couldn't put my coat on, couldn't get up the stairs, couldn't get in and out of the car and I noticed - I've got an allotment and I were in the greenhouse and on my knees and I couldn't get up, I'd got to crawl on my knees to get something to pull me up with." UPN 11

Many participants described profound disability which came on over a relatively short time period of time (typically days to weeks). Often these were people who, despite their age, had previously been active and suddenly suffered a life-changing reduction in their ability to carry out many activities of daily living. It was notable that participants often described their experience of PMR in terms of what the condition stopped them doing, rather than detailing specific symptoms.

One repeated observation by patients was that they became so stiff that they couldn't turn over in bed. A range of other activities were affected including getting dressed, toileting, managing stairs and getting in and out of a bath or the car.

Box 2: Disability

"I went on holiday in the September and on the holiday, I thought it was the travelling that had done it, I couldn't turn over from front to back in bed. And I couldn't get my hips down onto the loo. I was fine then for a few more days and then still on that holiday, I had that same thing again. I woke up and I was on my front, I couldn't get over in bed. And I developed strange pains across the top of my shoulders. I came back from that holiday….and I just went down within about a week of not being able to get out of bed, not being able to turn over in bed. And my husband was actually swinging my legs out, getting my arms and pulling me up out of bed." *UPN 2*

"I didn't know how to get in the car because my legs wouldn't bend, my arms wouldn't bend, she had to put one of her little one's booster seats on the front seat so I didn't have to lower myself quite so low and it had got to the stage where I couldn't lift my arms to comb my hair… really struggling with everything, walking upstairs and everything." *UPN 14*

"…it got to such a stage where I were laying in bed and quite frankly I could hardly move in bed and at the top of my arms – certainly from the elbow up to the top of the shoulder here and here I was sort of getting these cramp type pains." *UPN 21*

Theme 3: Experience of care

"I mean if they'd have given me steroids for like 24, 48 hours and it had the effect it did, they would have known long before." UPN 6

The path to diagnosis was very variable. Whilst some patients were diagnosed early on in the course of their illness, many felt that, with hindsight, a diagnosis could have been made earlier. Some expressed significant frustration about this. Several patients saw doctors multiple times and were tried on a range of treatments including analgesia, stopping statins, physiotherapy and in one case even antidepressants, prior to a diagnosis being made.

Participants tended to feel that their condition was poorly understood by the medical profession. Many had been given patient information leaflets (PILS) and some found these useful in that they validated their experiences and gave them confidence. Others however, were frustrated that the PILs portrayed the condition as mild and resolving within 2 years when this wasn't their experience. Patients and their relatives frequently

sought information from the internet but despite this, were often left with a sense of uncertainty about PMR and its management.

Box 3: Experience of care

"And I went to the doctors, well they were telling me to take paracetamols like and then they were no good, he increased it to some stronger stuff and I went back again, I said 'they weren't doing us any good' and then I suggested to him could it be this Simvastatin that I were on. And he said he'd thought of that and stopped it for about a month I think. And that didn't have any effect and so I went, I had a blood test and went for results of the blood test and he more or less knew what it were then straightaway." *UPN 11*

"I think I'd been up to see her when it first started, 'cause I could hardly walk... She kept sending me for these blood tests and the last time they wanted another blood test off me, my husband went up; he says, 'Look, my wife can't get out of bed this morning.' And they sent a doctor down to take it. And then he says to her, 'I think you ought to send her in hospital. She needs treating. She's not getting anywhere.' And that's what she did then, you see – she sent me to hospital." *UPN 18*

"I did actually have a month on a, what do they call it, you know the antidepressants, because I was going with all these pains and I wasn't getting anywhere at all. But I knew as soon as I started on the antidep-, it wasn't for me and that was it, after the month I came off them and I thought well, you know, I'm just going to see this through and I'm just going to have to see what's going to happen. And I'm going to have to create eventually and ask to see a specialist or something, because when you get to my age and you've been fit, you do know your own body, you know if there's something right or wrong." *UPN 2*

"When they fetched me back in and told me what I'd got and she printed so many sheets out and she said, the doctor, 'this is exactly you' and it was, that you can't get out of bed and you can't do this and you can't do the other….I mean it was all about it and it was me, definitely me." *UPN 17*

"When they said what I'd got, I was very pleased when they gave me the medication and it started to work so well. I was very happy about that but when they said that there's no cure for it because we don't know what it is, that was a bit upsetting." *UPN 5*

"I then looked online. There's quite a bit online actually but it all says the same thing – they don't know." *UPN 6*

Theme 4: Treatment and course of the condition

"I was smiling again because I'd got the power and I'd got the strength back. I got the walking back, I could go out." UPN 5

Prednisolone treatment brought about rapid resolution of symptoms in the majority of patients and many reported being amazed and relieved at how quickly they were able to resume normal activities. However, the burden of side effects from steroid treatment was also a strong theme. Weight gain, hyperactivity and irritability were the most frequently mentioned but there was a wide range of symptoms which patients attributed to the prednisolone. For some patients it reached a point where they felt the side effects were worse than the symptoms of PMR itself, though others viewed the side effects as 'a

small price to pay'. Several patients also commented on the additional tablet burden associated with being on long term prednisolone treatment as most were also prescribed calcium and vitamin D supplements, a bisphosphonate and a proton pump inhibitor.
The rate and pattern of reduction of prednisolone dose varied considerably between participants, as did the degree to which patients took charge of this themselves versus being guided by their doctor. Many described being aware of a slight worsening in symptoms with each dose reduction but that this would settle after a few days. Several patients had had more significant relapses at points during the disease course necessitating increasing their prednisolone dose. In most cases this was experienced as a resurgence of their original symptoms though at a less severe intensity.
Some of the participants were interviewed at a stage in their condition where they had reached very low doses of prednisolone or had even had a trial of stopping treatment altogether. In some cases participants described balancing the negative effects of being on low dose prednisolone with the, by then mild, PMR symptoms to achieve their desired quality of life. In general however there was a sense of not quite being back to the level of health that they had enjoyed prior to developing PMR. Some commented on the fact that they had aged during the disease course and become less fit due to reduced activity and the weight gain associated with treatment. This combination of factors resulted in them not feeling that they were able to recover fully to their pre-morbid state.

Box 4: Treatment and course of the condition

"He put me on these Prednisolone and it was like magic, it was just so good, you know, that I had no pain and I went back again to let him know how I was going on and I says 'thank you, you know, I can't say to you what a difference that's made to me'" *UPN 12*

"I just didn't feel like me, you know, it was almost like somebody else was living inside. I became tense, sometimes, or a bit ratty. And I really didn't like the weight gain and I think I put on over a stone in the first few months, you know, I went up a whole size of clothing and everything, which was not nice really." *UPN 19*

"Well I can put up with it, I can live with it, it's affected me all these aches and pains, aching and that, it's not as much of a sharp pain, it's just, you know, like, nagging ache. I can put up with that, but it's just, I think it's these side effects what I'm getting with the tablets what's worse. I feel as though this is worse now than the actual bad aching." *UPN 13*

"Yeah, I have put a bit of weight on with it and I've noticed that my stomach gets -, I never had a stomach but it gets really swollen, more so when I've had something to eat kind of thing and my face looks really bloated some days, you know, yeah, but I just think back to when I first started with it, you know, and I think to me it's a small price to pay, you know." *UPN 14*

"Every time he dropped the dose for a week, I could tell that it had dropped dose and I weren't very well, but I carried on and it like worked itself off, I worked through it sort of thing" *UPN 11*

Theme 5: Psychological impact

"But, well, I thought worst, you know, I thought I were like, what these illnesses where you just finish paralysed, I don't know what they call them but I felt it were going to be something like that, because I were getting worse." UPN 11

This was a striking and recurrent theme which linked closely with all of the other themes but particularly with that of disability. The pain and disability itself clearly impacted on patients' mood but many also described feeling fearful about the possible diagnosis and prognosis. Several patients specifically mentioned fearing that they had developed motor neurone disease, multiple sclerosis or some form of terminal muscle wasting illness. Previously fit people suddenly felt vulnerable and lost confidence and independence.
As a consequence there was frequently a significant sense of relief when a diagnosis of PMR was made. The importance of having a label to validate their experience and symptoms was apparent and the relief was even greater because a diagnosis of PMR meant that they could immediately receive an effective treatment.
After diagnosis, the focus of the psychological impact was different but it was still present. Many then reported anxiety about disease trajectory and adverse effects of medication, as well as experiencing a sense of loss for the life they had prior to the condition developing.

Box 5: Psychological impact

"At one time he was in the bathroom and he'd been trying to perform with a towel and couldn't and he started crying and he broke down and I went to him and I've never known him cry like that before and he says 'I'm bloody useless'."
"Well, I thought me life, I wouldn't say me life had come to an end but I was so – "
"You thought your life, as it had been, had finished." *UPN 1 and his wife*
"Not being able, as I say, those 3 days I didn't take any Ibuprofen before I saw the doctor, I really had to depend on my daughter, you know, yeah. and frightened, really, being in the house on my own, like going upstairs, you know, because although there's a rail, as I say, my legs just wouldn't bend to go up and once I got to the top of the stairs when the rail finished, I didn't know how I were going to go any further, it were a real ordeal, you know, yeah. And there were certain things that I daren't to when I were on my own, I wouldn't have dared got in the bath, you know, unless somebody was in the house and I just kept the phone on me all the time, because I really thought I were going to fall at some point." *UPN 14*
"But as I say, it were just – it got to a stage as I say when I went to the doctors – it got to a stage when I were literally struggling to turn over in bed – that were quite frightening because you'd lay on your back and all of a sudden you're thinking 'well, it's almost like being locked in your body in a way'. Yeah – you hear about – and I forget what the name of these – some of these things – but these wasting away diseases – I forget – I can't remember what the name is – it's on the tip of my tongue now – but you think 'well, if it's something to do with the muscles or if it's something like that have I got something like multiple sclerosis coming on or something like that' and not being a doctor I wouldn't know what the symptoms are. It frightened me quite frankly and it knocked me off balance in a way because – so it's made me feel more vulnerable." *UPN 21*
"I was just glad to get a diagnosis, you know, and I was euphoric, you won't believe this! But the day they told me I'd got PMR, I was euphoric, I was picking the phone up to my sister and said 'I've got an answer now, I've got this' because to me I were then going to get the cure and get better." *UPN 2*
"Because I hadn't heard of it at all. I really did think oh thank God somebody's listening to me. I thought I was imagining it." *UPN 3*

4. DISCUSSION AND CONCLUSION

4.1 Discussion

This is the first qualitative study to explore the effect of PMR on patients' lives. Studies of other chronic rheumatological conditions have contributed to a wealth of models describing the effects of long term conditions on patients and their families e.g. Bury's 'Chronic illness as biographical disruption' [16] and Weiner's 'Strategies for tolerating uncertainty' [17], and many of the themes identified in this study correlate well with these existing models. Eisenberg's concept of the distinction between 'diseases' (which doctors diagnose and treat) and 'illnesses' (which patients' experience) [18] is also highly relevant to PMR. Given that PMR is a heterogenous condition, affects older age groups (who will have a huge range of comorbidities, life experiences and coping strategies), causes pain and disability and is treated with medication capable of causing significant harm, the importance of assessing 'illness' rather than focussing on 'disease' is particularly pronounced. The risk otherwise is of significant under- or over-treatment with associated harms. A patient reported outcome measure for PMR could significantly contribute to a holistic assessment, acting as a bridge between 'disease' and 'illness' and thus between doctor and patient.

The results from this study support previous findings of the heterogeneity of PMR which contributes to the complexity of diagnosis and assessment of disease activity [3,19]. The terminology used when discussing symptoms is important in achieving a shared understanding between doctor and patient and enabling a correct diagnosis. Recognising therefore that patients may describe weakness as a feature of PMR is important, whether or not it is truly a separate construct from pain or stiffness.

Morning stiffness is a characteristic feature of inflammatory musculoskeletal conditions and is part of the diagnostic criteria for PMR. However, there have been questions raised over the usefulness of this concept in this condition [12,20] and the findings from this patient group echo the suggestion that stiffness often persists through the day and is worse after any period of rest. It may be more appropriate to discuss stiffness rather than 'morning stiffness' in PMR and the use of concepts such as duration of morning stiffness as outcome measures may be unhelpful.

Participants in this study tended to describe the impact of PMR in terms of 'disability' rather than detailing localised symptoms. Difficulty with carrying out a wide range of daily activities was described and this significantly affected quality of life. Key limitations mentioned on many occasions were the inability to turn over in bed and the inability to get up after bending down. These particular difficulties contributed to a sense of helplessness and vulnerability and exemplify the overlap between the themes of disability and the psychological impact of the condition. The relatively rapid change in people's ability to carry out every day activities was associated with disruption of normal roles, 'loss of self' and a sense of uncertainty as has been described in studies of other long term conditions [16,17,21].

Another striking emergent theme from these interviews was the profound psychological impact of the symptoms of PMR prior to diagnosis. Many patients had significant anxiety about a wide range of potentially serious neurological and malignant conditions. In cases where there was a perceived delay in diagnosis, this anxiety was exacerbated. The symptoms of PMR have a wide differential diagnosis and controversy still exists as to the defining characteristics of the condition [19]. The diagnosis is often made over a series of consultations forming a process which may include an initial trial of treatment [8].

Understanding and acknowledging patient anxiety and addressing their specific fears during this process could improve patient experience.

4.2 Strengths and limitations

This study is unusual in that it recruited patients from primary care, the setting where the majority of patients with PMR are managed [7]. This is a true strength of this study and allows a wider transferability of the findings. Whilst patients may not have received a 'gold standard' diagnosis from a rheumatology specialist, we only included those with a PMR diagnosis and evidence of meeting the classification criteria [3] for PMR – namely bilateral pain and stiffness in the hips and shoulders and elevated inflammatory markers. It was surprising that more men than women were recruited given that the quoted incidence ratio is 2:1 female to male [1] but, whilst we acknowledge that there are gender differences in the way patients experience chronic illnesses eg. in stress and relationships, we don't believe that this pattern of recruitment detracts from the transferability of the main findings.

Both of the researchers carrying out the interviews and analysis in this study were GPs. Their prior understanding of the condition therefore arose from this background and will no doubt have shaped the research process to some degree. However, as researchers they also had training in qualitative interviewing and reflexive analytical skills and were systematic and self-conscious in their approach. The participants were aware of the researchers' profession and this may have affected the way they discussed their experiences. However the setting of the interviews and naturalistic style (as opposed to general practice consultation in a surgery) will hopefully have mitigated this to some degree.

4.3 Conclusions

This qualitative primary care study has broadened our understanding of PMR and its effects on patients' lives. The discussions around pain and stiffness and the course of the condition were anticipated, but the severity and impact of the disability, the associated fear and vulnerability and the often less than ideal experience of care were all surprisingly strong themes. The systematic analytical approach used in this study allowed these themes to emerge and be tested through constant comparison, ensuring that the resulting concepts are truly grounded in the patient experience.

4.4 Practice implications

This study highlights several important aspects of patients' experiences of PMR which may not necessarily be recognised or considered by health care professionals in our traditional understanding of the condition. Through greater professional understanding of the ways in which the condition affects patients, patient care may be improved.

In addition to the 5 main themes identified from the interview data we have derived an itemised list of functional activities that participants reported being limited by their PMR. We plan to use this, set in the context of the conceptual framework developed from the rich interview data, to design a patient reported outcome measure specific to PMR. It is hoped that this will have both research and clinical utility by contributing to a standardised assessment of the condition.

References

1. Smeeth L, Cook C, Hall AJ. Incidence of diagnosed polymyalgia rheumatica and temporal arteritis in the United Kingdom, 1990-2001. *Ann Rheum Disease* 2006;65:1093-8
2. Crowson CS, Matteson EL, Myasoedova E, et al. The lifetime risk of adult-onset rheumatoid arthritis and other inflammatory autoimmune rheumatic diseases. *Arthritis Rheum* 2011;63:633-39.
3. Dasgupta B, Cimmino MA, Maradit-Kremers H, et al. 2012 provisional classification criteria for polymyalgia rheumatica: a European League Against Rheumatism/American College of Rheumatology collaborative initiative. *Annals Rheum Disease* 2012;71:484-92.
4. Helfgott SM, Kieval RI. Polymyalgia rheumatica in patients with a normal erythrocyte sedimentation rate. *Arthritis Rheum* 1996;39:304-07
5. Hutchings A, Hollywood J, Lamping DL, et al. Clinical outcomes, quality of life, and diagnostic uncertainty in the first year of polymyalgia rheumatica. *Arthritis Rheum* 2007;57:803-9
6. Gabriel SE, Sunku J, Salvarani C, et al. Adverse outcomes of antiinflammatory therapy among patients with polymyalgia rheumatica. *Arthritis Rheum* 1997;40:1873-8
7. Barraclough K, Liddell WG, du Toit J, et al. Polymyalgia rheumatica in primary care: a cohort study of the diagnostic criteria and outcome. *Fam Prac* 2008;25:328-33
8. Dasgupta B, Borg FA, Hassan N, et al. BSR and BHPR guidelines for the management of polymyalgia rheumatica. *Rheumatology (Oxford)* 2010;49:186-90
9. Hewlett S, Cockshott Z, Byron M, et al. Patients' perceptions of fatigue in rheumatoid arthritis: Overwhelming, uncontrollable, ignored. *Arthritis Care Res* 2005;53:697-702
10. Ahlmén M, Nordenskiöld U, Archenholtz B, et al. Rheumatology outcomes: the patient's perspective. A multicentre focus group interview study of Swedish rheumatoid arthritis patients. *Rheumatology (Oxford)* 2005;44:105-10
11. Kirwan JR, Minnock P, Adebajo A, et al. Patient perspective: fatigue as a recommended patient centered outcome measure in rheumatoid arthritis. *J Rheumatol* 2007;34:1174-77
12. Mackie SL, Arat S, da Silva J, et al. Polymyalgia Rheumatica (PMR) Special Interest Group at OMERACT 11: outcomes of importance for patients with PMR. *J Rheumatol* 2014;41:819-23
13. US Department of Health and Human Services Food and Drug Administration. Guidance for industry: patient-reported outcome measures: use in medical product development to support labelling claims. 2009. www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf.
14. Lincoln YS, Guba EG, Pilotta JJ. Naturalistic inquiry: Beverly Hills, CA: Sage Publications, 1985.
15. Glaser BG, Strauss AL. The discovery of grounded theory : strategies for qualitative research. Hawthorne, N.Y.: Aldine de Gruyter, 1967.
16. Bury M. Chronic illness as biographical disruption. *Sociology of health and illness* 1982;4:167-182
17. Weiner C. The burden of rheumatoid arthritis: tolerating the uncertainty. *Soc. Sci & Med* 1975;9:97-104

18. Eisenberg L. Disease and illness: distinctions between professional and popular ideas of sickness. *Culture, Medicine and Psychiatry* 1977;1:9-23
19. Weyand CM, Fulbright JW, Evans JM, et al. Corticosteroid requirements in polymyalgia rheumatica. *Arch Intern Med* 1999;159:577-84
20. Dasgupta B, Salvarani C, Schirmer M, et al. Developing classification criteria for polymyalgia rheumatica: comparison of views from an expert panel and wider survey. *J Rheumatol* 2008;35:270-7
21. Charmaz K. Loss of self: a fundamental form of suffering in the chronically ill. *Sociology of health and illness* 1983;5:168-92

Assessment of the face validity, feasibility and utility of a patient completed questionnaire for polymyalgia rheumatica: a postal survey using the QQ-10 questionnaire

Helen Twohig, Georgina Jones, Sarah Mackie, Christian Mallen, Dr Caroline Mitchell

Abstract

Background The development of a patient-reported outcome measure (PROM) for polymyalgia rheumatica (PMR), a condition that causes pain, stiffness and disability, is necessary as there is no current validated disease-specific measure. Initial literature synthesis and qualitative research established a conceptual framework for the condition along with a list of symptoms and effects of PMR that patients felt were important to them. These findings were used to derive the candidate items for a patient-completed questionnaire. We aim to establish the face validity of this initial 'long form' of a PROM.

Methods People with a current or previous diagnosis of PMR were recruited both from the community and from rheumatology clinics. They were asked to complete the PMR questionnaire along with the QQ-10 questionnaire, which is a measure used to assess the face validity, feasibility and utility of patient healthcare questionnaires.

Results A total of 28 participants with an age range of 59-85 years and a length of time since diagnosis from 4 months to 18 years completed the QQ-10. The overall mean 'value'

score was 79% (SD 12) and the mean 'burden' score was 21% (SD 18). The free text comments were analysed thematically and were found to focus on layout, content, where in the clinical pathway the questionnaire would be most beneficial, specific items missing and other areas for consideration.

Conclusions The high mean value score and low burden score indicate that the questionnaire has good face validity and is acceptable to patients. The questionnaire now needs to undergo further psychometric evaluation and refinement to develop the final tool for use in clinical practice and research.

Keywords:

Polymyalgia rheumatica

Patient perspective

Outcomes research

Patient reported outcome measures

Questionnaire validity and utility assessment

Background

Polymyalgia rheumatica (PMR) causes significant pain, stiffness and disability in older adults and is treated with systemic glucocorticoids which can themselves can cause significant additional morbidity.[1,2] A recent systematic review of outcome measures used in PMR research studies[3] identified a wide variety of instruments in use, none of which were specifically developed to measure symptom burden in patients with PMR. Less than 10% of studies measured physical function, quality of life or fatigue, despite these being important to patients.[4] Further work, including a Delphi survey conducted by the OMERACT PMR Working Group, highlighted the need for a disease-specific outcome measure that would cover domains of life impact relevant to patients.[5] Applying international guidelines for PROM development,[6] we carried out initial qualitative work with patients with PMR to better understand patient experience of the condition and establish a conceptual framework for a future PMR PROM.[7] Here we report the next steps taken to derive a "long-list" of candidate items for a PMR-specific PROM and the assessment of face validity, feasibility and utility both in the participants of

the original study and in a separate group of patients, using a validated method, the QQ-10 questionnaire.[8] The QQ-10 is a measure developed to collect standardised information on important aspects of a questionnaire's qualities from the patient's perspective and is used to assess the face validity, feasibility and utility of patient healthcare questionnaires. This part of the iterative PROM development process is particularly important to ensure that the instrument is acceptable and contains the content that is relevant to patients with PMR, and completion burden of the final questionnaire is minimised thus improving follow up and questionnaire completion rates in clinical trials.[9] These data are vital for development of a PROM that is valid for either research studies or clinical practice.

Methods

A long list of candidate items for the PROM was developed from the data obtained from a previous qualitative study carried out by our group.[7] These items were categorised into main symptoms / duration (4 items), function (24 items), emotional and psychological well-being (11 items), steroid side effects (10 items) and overall well-being (1 item). Respondent validation of the themes and items was carried out with the original interview participants. All 20 interviewees from the qualitative study were sent a copy of the long-list and asked to return it with any comments that they had on the content. 10 of these individuals agreed to a structured telephone interview to discuss the long-list in more detail and the information gathered from this was used to make changes to it at that stage to form the draft questionnaire (Table 1).

To assess face validity and utility in an independent group of patients, patients with a diagnosis of PMR were recruited through two routes: (1) community-based: through a patient-led patient support group, PMR&GCAUK North East Support and (2) hospital-based: through rheumatology clinics at Leeds Teaching Hospitals NHS Trust. This recruitment strategy was designed to sample from across the spectrum of patients with PMR as it is a condition managed in both primary and secondary care. Ethical approval was received from the National Institute for Social Care and Health Research Research Ethics Service, Wales REC 7 (Ref 12/WA/0344) and all participants provided informed consent. Patients received the study materials by post (community-based recruitment) or

in a sealed envelope from their treating rheumatologist (hospital-based recruitment), completed the study materials at home in their own time, and returned the completed forms to the lead researcher (HT), who was based at a different institution and in a different city to their rheumatology clinic.

The data from the QQ-10 questionnaire were analysed both quantitatively and qualitatively. For quantitative analysis we used the QQ-10 scoring method.[8] Likert ratings from strongly disagree to strongly agree (coded as 0-4) were summed separately for the first six questions comprising the value score (helped me communicate about my condition, relevant to my condition, easy to complete, included all the aspects of my condition I am concerned about, was enjoyable, would be happy to complete as part of routine care), and from the last four questions comprising the burden score (too long, embarrassing, complicated, upset me).

Qualitative thematic analysis[10] was performed on comments received in response to the three free-text questions at the end of the QQ-10:

1. Do you have any comments or suggestions on how the questionnaire you used could be improved (e.g. its structure, appearance or design)
2. Were any of your important symptoms, problems or concerns missed out by the questionnaire you used?
3. Do you feel that any areas or problems in the questionnaire you used were over-represented?

Results

Twenty eight patients took part (20 female and 8 male; age range 59-85 years; duration since PMR diagnosis 4 months to 18 years; apart from a single participant who was 18 years post diagnosis, no patient was more than 5 years post diagnosis). 18 of the participants were still on steroid treatment for their PMR.

The overall mean value score was 79% (SD 12) and the mean burden score was 21% (SD 18). The median of each domain making up the value score was >2 (range 0-4) and the median of each domain making up the burden score was <1.5 (range 0-4). Charts 1 and 2

show the median scores for each question and Chart 3 shows the spread of responses for each question.

The five emergent themes from the free text comments (Table 2) were 1) layout, 2) content, 3) where in the clinical pathway the questionnaire would be most beneficial, 4) specific items not covered and 5) other areas for consideration for inclusion. The content theme encompassed sub-themes of depth and detail, specificity to PMR and heterogeneity of the condition.

Discussion

This study represents a distinct, patient-orientated phase within a stepped standardised methodological approach to developing a PROM for PMR for research and clinical practice. The use of the validated QQ-10 measure is an example of how to embed patient perspectives at all stages within the PROM development process, moving beyond the paradigm of clinician-orientated outcome measures.

The high value and low burden scores are encouraging as regards face validity and feasibility of the questionnaire. They are similar to scores obtained when the QQ-10 has been used in other studies, including evaluation of the King's Health Questionnaire[11] and evaluation of use of a bladder diary,[12] and both of these tools were judged as having been proven to be useful based on these results. The free text comments provided added richness to the response data and their analysis highlights some important required amendments to the structure and layout of the questionnaire as well as suggesting some additional points for inclusion. The comments related to content echo some of the findings from our earlier qualitative work[7] and some of the known challenges of outcome measurement in PMR; chiefly the heterogeneity of the condition, difficulty in assessment in the presence of co-morbidities and overlap with other conditions.[1,4,13]

The strengths of this study include the use of a validated instrument (QQ-10) to assess face validity, feasibility and utility of instruments designed to assess the life impact of particular disease states. The QQ-10 itself was rigorously developed using standard psychometric methods and has been demonstrated to have high internal consistency and item correlation and to be acceptable and understandable to patients.[8] It also has the

advantages of being quick and cheap to administer and allowing comparison of different versions of a measure at several stages during development. At this stage of questionnaire development we made the decision not to employ interview-based qualitative methods, such as cognitive interviewing, as the postal method of responding did not exclude patients too frail to participate in cognitive interviews, and allowed patients to respond honestly, in private, and without any concern that the responses might be seen by their treating clinicians. A further strength of this study was the use of a combination of recruitment methods which identified patients with a range of disease durations similar to that described in previous literature.[14,15]

The main limitation of this study was the small number of participants and opportunistic (and thus not statistically representative) sampling method. However, this sample size was appropriate for this stage of PROM development.[16,17] A further limitation of this study was that the community-based method of recruitment required patients to self-identify as being diagnosed with PMR. We did not attempt to validate diagnosis by means of classification or other criteria designed to select patients with PMR for research studies, because we wanted to assess this questionnaire in a real-life setting for clinical practice, not just for use in research studies.

Conclusions

The long-form of our PMR questionnaire was shown to have face validity for patients and be acceptable for use in their care. The work reported here represents an essential step in the PROM development process, paving the way for further work using a larger sample size that will allow formal psychometric validation, item reduction and ultimately generation of a fully-validated patient-reported outcome measure for PMR.

List of abbreviations

PROM – patient reported outcome measure

PMR – polymyalgia rheumatica

OMERACT – outcome measures in rheumatology (www.omeract.org). An independent initiative of international health professionals interested in outcome measures in rheumatology.

Declarations

Ethics

Ethical approval was received from the National Institute for Social Care and Health Research Research Ethics Service, Wales REC 7 (Ref 12/WA/0344) and all participants provided informed consent.

Acknowledgements

We would like to thank all of the study participants and the trustees and members of PMR&GCA UK North East Support for their continued support and cooperation.

References

1. Mackie SL, Mallen CDM. Polymyalgia rheumatica. *BMJ* 2013;347:f6937.
2. Dejaco C, Singh YP, Perel P, et al. 2015 Recommendations for the Management of Polymyalgia Rheumatica. A European League Against Rheumatisim / American College of Rheumatology Collaborative Initiative. *Ann Rheum Dis* 2015;74:1799-1807.
3. Duarte C, Ferreira JO, Mackie SL, et al. Outcome measures in polymyalgia rheumatica. A systematic review. *J Rheumatol* 2015;42(12):2503-11.
4. Mackie SL, Arat S, da Silva J, et al. Polymyalgia rheumatica (PMR) special interest group at OMERACT 11: outcomes of importance for patients with PMR. *J Rheumatol* 2014;41(4):819-23.
5. Helliwell T, Brouwer E, Pease CT, 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(1):182-6.
6. U.S. Dept of Health and Human Services Food and Drug Administration. Guidance for industry: patient-reported outcomes measures: use in medicinal product development to support labelling claims. 2009. http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/UCM19328 2.pdf

7. Twohig HJ, Mitchell CM, Mallen CDM et al. "I suddenly felt I'd aged": A qualitative study of patient experiences of polymyalgia rheumatica (PMR). *Patient Educ Couns* 2015;98:645-50.
8. Moores KL, Jones GL, Radley SC. Development of an instrument to measure face validity, feasibility and utility of patient questionnaire use during health care: the QQ-10. *Int J Qual Health Care* 2012;24:517-24.
9. Prescott RJ, Counsell CE, Gillespie WJ, et al. Factors that limit the quality, number and progress of randomised controlled trials. *Health Technol Assess* 1993;3(20):1-143.
10. Boyatzis RE. Transforming qualitative information: Thematic analysis and code development. Thousand Oaks, London, & New Delhi: SAGE Publications. 1998
11. Vij M, Srikrishna S, Robinson D, et al. Quality assurance in quality of life assessment – measuring the validity of the King's Health Questionnaire. *Int Urogynaecol J* 2014;25(8):1133-5.
12. Vella M, Robinson D, Cardozo L, et al. The bladder diary: do women perceive it as a useful investigation? *Eur J Obstet Tynecol Reprod Biol* 2012;162(2):221-3.
13. Dasgupta B, Cimmino MA, Maradit-Kremers H, 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.
14. Bahlas S, Ramos-Remus C, Davis P. Clinical outcome of 149 patients with polymyalgia rheumatica and giant cell arteritis. *J Rheumatol* 1998;25(1):99-104.
15. Mackie SL, Hensor EM, Haugeberg G, et al. Can the prognosis of polymyalgia rheumatic be predicted at disease onset? Results from a 5-year prospective study. *Rheumatology* 2009;49(4):716-722.
16. Jones G, Kennedy S, Barnard A, et al. Development of an endometriosis quality-of-life instrument: The Endometriosis Health Profile-30. *Obstet Gynecol* 2001:98(2), 258-264.
17. Mirzaee Rabor F, Taghipour A, Mirzaee M et al. Developing a questionnaire for Iranian women's attitude on medical ethics in vaginal childbirth. *Nurs Midwifery Stud* 2015 Dec;4(4):e29004. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4733499/ (accessed 1 July, 2016)

Appendix 7.2: List of items derived from the interview data

Symptoms and effects

- Pain
- Stiffness
- Weakness
- Fatigue
- Sweats
- Sleep disturbance
- Low mood
- Anxiety
- Fear about diagnosis / prognosis
- Loss of confidence
- Feeling vulnerable
- Fear of falling
- Needing help to look after yourself
- Not wanting to go out
- Unable to do usual hobbies / activities

Functional limitations

- Bending down and getting up
- Getting in and out of a car
- Driving
- Getting in and out of bed
- Turning over in bed
- Getting in and out of a chair
- Getting in and out of the bath
- Washing and drying fully
- Dressing (including coat and shoes and socks)
- Combing or blow drying hair

- Toileting – getting on and off and wiping
- Managing stairs
- Walking up stairs
- Carrying and lifting things
- Opening jars / tins
- Gardening
- Housework
- Inability to sit for a prolonged time (>30 mins)
- Walking

Side effects of prednisolone

- Weight gain
- Change in appearance (fatter face, saggy skin)
- Irritability
- Low mood
- Euphoria
- Hyperactivity
- Easy bruising
- Indigestion
- Hair loss

Appendix 7.3: PMR-PROM Version 1

Pain / stiffness

1. How severe has the pain from your PMR been during the last 2 weeks?

   Likert / VAS / 0-10

2. How severe has the stiffness caused by your PMR been during the last 2 weeks?

   Likert / VAS / 0-10

3. On average, how much of each day has the pain / stiffness from your PMR been present for during the last 2 weeks?

   All day *About half the day* *Around 2-3 hours* *<2 hours*

Function

4. Over the last 2 weeks, compared to what you can normally do, to what extent has your PMR affected your ability to do the following activities?

	Unaffected	**Caused some difficulty**	**Caused much difficulty**	**Caused me to be unable to do**	**Not relevant**
Bend down					
Get up after bending down					
Get in or out of a car					
Drive a car					
Get in or out of bed					
Turn over in bed					

Get in or out of a chair					
Get in or out of the bath					
Wash yourself fully					
Dry yourself fully after a shower / bath					
Get dressed					
Comb or blow dry your hair					
Get on or off the toilet					
Wipe yourself after going to the toilet					
Walk up stairs					
Walk up hills					
Walk on the flat					
Carry or lift things					
Grip objects					
Do housework					
Do gardening					
Sit for more than 30 mins at a time					

Emotional and psychological well being

5. In the last 2 weeks have your PMR symptoms….

	Never	Rarely	Sometimes	Often	Always
Caused you to feel low in mood?					
Caused you to feel anxious?					
Caused you to feel vulnerable?					
Lowered your self-confidence?					
Made you worried that you might fall over?					
Caused you to need more help with looking after yourself?					
Made you less inclined to go out?					
Stopped you doing hobbies that you used to do?					
Made you worry about the future?					
Affected your sleep?					
Made you feel more tired than usual?					

Steroid side effects

6. How much have you been affected by side effects from your medication in the last 2 weeks?

Likert scale / VAS / 0-10

7. In the last 2 weeks, have you been bothered by any of the following side effects of your steroid medication?

	Yes	**No**
Weight gain		
Change in appearance (fatter face, saggy skin)		
Irritability		
Low mood		
Euphoria		
Hyperactivity		
Easy bruising		
Indigestion		
Hair loss		

8. Do you feel back to the level of health you were at before you first experienced PMR symptoms?

   Yes
   No

Appendix 7.4: PMR-PROM Version 2

Name: Today's date:
Date of birth:

Pain / stiffness / weakness

1. How severe has the pain from your PMR been during the last 2 weeks? Please mark it on the scale below where 0 = no pain and 10 = the worst pain you've ever felt.

   0 1 2 3 4 5 6 7 8 9 10

2. How severe has the stiffness caused by your PMR been during the last 2 weeks? Please mark it on the scale below where 0 = no stiffness and 10 = the worst stiffness you've ever felt.

   0 1 2 3 4 5 6 7 8 9 10

3. How severe has the weakness caused by your PMR been during the last 2 weeks? Please mark it on the scale below where 0 = no weakness and 10 = the worst weakness you've ever felt.

   0 1 2 3 4 5 6 7 8 9 10

4. On average, how much of each day has the pain / stiffness from your PMR been present for during the last 2 weeks?

   All day *About half the day* *Around 2-3 hours* *<2 hours*

Function

5. Over the last 2 weeks, compared to what you can normally do, to what extent has PMR affected your ability to do the following activities?

	Unaffected	**Caused some difficulty**	**Caused much difficulty**	**Caused me to be unable to do**	**Not relevant**
Bend down					
Get up after bending down					
Get in or out of a car					
Drive a car					
Get in or out of bed					
Turn over in bed					
Get in or out of a chair					
Get in or out of the bath					
Wash yourself fully					
Dry yourself fully after a shower / bath					
Get dressed					
Take your coat on or off					
Put on or take off your socks and shoes					

Comb or blow dry your hair					
Get on or off the toilet					
Wipe yourself after going to the toilet					
Walk up stairs					
Walk up hills					
Walk on the flat					
Carry or lift things					
Grip objects					
Do housework					
Do gardening					
Sit for more than 30 minutes at a time					

Emotional and psychological well-being

6. In the last 2 weeks have your PMR symptoms….

	Never	Rarely	Sometimes	Often	Always
Caused you to feel low in mood?					
Caused you to feel anxious?					
Caused you to feel vulnerable?					
Lowered your self-confidence?					
Made you worried that you might fall over?					
Caused you to need more help with looking after yourself?					
Made you less inclined to go out?					
Stopped you doing hobbies that you used to do?					
Made you worry about the future?					
Affected your sleep?					
Made you feel more tired than usual?					

Steroid side effects

7. How much have you been affected by side effects from your medication in the last 2 weeks?
   Please mark it on the scale below where 0 = unaffected and 10 = severely affected

   0 1 2 3 4 5 6 7 8 9 10

8. In the last 2 weeks, have you been bothered by any of the following side effects of your steroid medication?

	Yes	**No**
Weight gain		
Change in appearance (fatter face, saggy skin)		
Irritability		
Low mood		
Euphoria		
Hyperactivity		
Easy bruising		
Indigestion		
Hair loss		

Overall well-being

9. Do you feel back to the level of health you were at before you first experienced PMR symptoms?

   Yes
   No

Appendix 7.5: PMR-PROM Version 3

Today's date:

Patient Details:

Name:

Date of birth:

Gender:

Month and year of PMR diagnosis:

Current dose of prednisolone:

Other current medication or therapies:

The following questions ask about your symptoms of polymyalgia rheumatica and the way in which it is affecting you at the moment.
If you are unsure about how to answer a question, please give the best answer you can.
Thankyou.

1. How severe has the pain from your PMR been during the last 2 weeks?
   Please mark it on the scale below where 0 = no pain and 10 = the worst pain you've ever felt.

   0 1 2 3 4 5 6 7 8 9 10

2. How severe has the stiffness caused by your PMR been during the last 2 weeks?
   Please mark it on the scale below where 0 = no stiffness and 10 = the worst stiffness you've ever felt.

   0 1 2 3 4 5 6 7 8 9 10

3. How severe has the weakness caused by your PMR been during the last 2 weeks?
   Please mark it on the scale below where 0 = no weakness and 10 = the worst weakness you've ever felt.

   0 1 2 3 4 5 6 7 8 9 10

4. On average, how much of each day has the pain / stiffness from your PMR been present for during the last 2 weeks?
   Please circle the answer that most closely applies to you.

   All day *About half the day* *Around 1-3 hours* *<1 hour*

Function

5. Over the last 2 weeks, compared to what you can normally do, has PMR limited your ability to do the following activities?

	No, not limited at all	**Yes, limited a little**	**Yes, limited a lot**	**Not relevant**
Bend down				
Get up after bending down				
Get in or out of a car				

Drive a car				
Get in or out of bed				
Turn over in bed				
Get in or out of a chair				
Get in or out of the bath				
Wash yourself fully				
Dry yourself fully after a shower / bath				
Get dressed				
Take your coat on or off				
Put on or take off your socks and shoes				
Comb or blow dry your hair				
Get on or off the toilet				
Wipe yourself after going to the toilet				
Engage in intimate / sexual activity				
Walk up stairs				
Walk up hills				

Walk on the flat				
Carry or lift things				
Reach above your head for things				
Grip objects				
Do housework				
Do gardening				
Sit for more than 30 minutes at a time				
Participate in sports				

Emotional and psychological well-being

6. In the last 2 weeks have your PMR symptoms….

	None of the time	A little of the time	Some of the time	Most of the time	All of the time
Caused you to feel low in mood?					
Caused you to feel anxious?					
Caused you to feel vulnerable?					
Lowered your self-confidence?					
Made you worried that you might fall over?					
Caused you to need more help with looking after yourself?					
Made you less inclined to go out?					
Stopped you doing hobbies that you used to do?					
Made you worry about the future?					
Affected your sleep?					
Made you feel more tired than usual?					

Treatment side effects

7. How much have you been affected by side effects from your medication in the last 2 weeks?
   Please mark it on the scale below where 0 = unaffected and 10 = severely affected

   0 1 2 3 4 5 6 7 8 9 10

8. In the last 2 weeks, have you been bothered by any of the following side effects of your steroid medication?

	Yes	**No**
Weight gain		
Change in appearance (fatter face, saggy skin)		
Irritability		
Low mood		
Euphoria		
Hyperactivity		
Easy bruising		
Indigestion		
Insomnia		
Hair loss		

Overall well-being

9. Do you feel back to the level of health you were at before you first experienced PMR symptoms?

   Yes / No

Appendix 7.6: QQ-10 Questionnaire

Please circle the answers below each of the following 10 statements that best fit your feelings about the PMR questionnaire that you recently completed. Please use the space at the bottom of the next page to make additional comments.

1. The questionnaire helped me to communicate about my condition

Strongly agree Mostly agree Neither agree or disagree Mostly disagree Strongly disagree

2. The questionnaire was relevant to my condition

Strongly agree Mostly agree Neither agree or disagree Mostly disagree Strongly disagree

3. The questionnaire was easy to complete

Strongly agree Mostly agree Neither agree or disagree Mostly disagree Strongly disagree

4. The questionnaire included all the aspects of my condition that I am concerned about

Strongly agree Mostly agree Neither agree or disagree Mostly disagree Strongly disagree

5. I enjoyed filling in the questionnaire

Strongly agree Mostly agree Neither agree or disagree Mostly disagree Strongly disagree

6. I would be happy to complete the questionnaire again in the future as part of my routine care

Strongly agree Mostly agree Neither agree or disagree Mostly disagree Strongly disagree

7. The questionnaire was too long

Strongly agree Mostly agree Neither agree or disagree Mostly disagree Strongly disagree

8. The questionnaire was too embarrassing

Strongly agree Mostly agree Neither agree or disagree Mostly disagree Strongly disagree

9. The questionnaire was too complicated

Strongly agree Mostly agree Neither agree or disagree Mostly disagree Strongly disagree

10. The questionnaire upset me

Strongly agree Mostly agree Neither agree or disagree Mostly disagree Strongly disagree

Do you have any comments or suggestions on how the questionnaire you used could be improved (e.g. its structure, appearance or design)?

. . . . . .
. . . . . .
. . . . . .

Were any of your important symptoms, problems or concerns missed out by the questionnaire you used?

. . . . . .
. . . . . .
. . . . . .

Do you feel that any areas or problems in the questionnaire you used were over-represented?

. . . . . .
. . . . . .
. . . . . .

Thank you for taking the time to complete this questionnaire.

Appendix 7.7: PMR-PROM Version 4

Today's date:

Personal Details:

Name:

Date of birth:

Gender:

Month and year of PMR diagnosis:

Have you been referred to a rheumatologist about your PMR?:

Current dose of prednisolone:

Other current medication or therapies:

The following questions ask about your symptoms of polymyalgia rheumatica and the way in which it is affecting you at the moment.
If you are unsure about how to answer a question, please give the best answer you can.
Thankyou.

1. Physical Symptoms

a. How severe has the pain from your PMR been during the last 3 days?
Please mark it on the scale below where 0 = no pain and 10 = the worst pain you've ever felt.

0 1 2 3 4 5 6 7 8 9 10

b. On average, how much of each day has the pain from your PMR been present for during the last 3 days
Please circle the answer that most closely applies to you.

All day About half the day Around 1-3 hours Less than 1 hour Less than 30 mins

c. How severe has the stiffness caused by your PMR been during the last 3 days?
Please mark it on the scale below where 0 = no stiffness and 10 = the worst stiffness you've ever felt.

0 1 2 3 4 5 6 7 8 9 10

d. On average, how much of each day has the stiffness from your PMR been present for during the last 3 days?
Please circle the answer that most closely applies to you.

All day About half the day Around 1-3 hours Less than 1 hour Less than 30 mins

e. How severe has the weakness caused by your PMR been during the last 3 days?
Please mark it on the scale below where 0 = no weakness and 10 = the worst weakness you've ever felt.

0 1 2 3 4 5 6 7 8 9 10

f. On average, how much of each day has the weakness from your PMR been present for during the last 3 days?
Please circle the answer that most closely applies to you.

All day About half the day Around 1-3 hours Less than 1 hour Less than 30 mins

2. **Function**

Over the last 3 days, compared to what you can normally do, has PMR limited your ability to do the following activities?

	No, not limited at all	**Yes, limited a little**	**Yes, limited a lot**	**Not relevant**
Bend down				
Get up after bending down				
Get in or out of a car				
Drive a car				
Get in or out of bed				
Turn over in bed				
Get in or out of a chair				
Get in or out of the bath				
Wash yourself fully				
Dry yourself fully after a shower / bath				
Get dressed				
Take your coat on or off				
Put on or take off your socks and shoes				
Comb or blow dry your hair				

Get on or off the toilet				
Wipe yourself after going to the toilet				
Engage in intimate / sexual activity				
Walk up stairs				
Walk up hills				
Walk on the flat				
Carry or lift things				
Reach above your head for things				
Grip objects				
Do housework				
Do gardening				
Sit for more than 30 minutes at a time				
Participate in sports				

3. Emotional and psychological well-being

In the last 3 days have your PMR symptoms….

	None of the time	A little of the time	Some of the time	Most of the time	All of the time
Caused you to feel low in mood?					
Caused you to feel anxious?					
Caused you to feel vulnerable?					
Lowered your self-confidence?					
Made you worried that you might fall over?					
Caused you to need more help with looking after yourself?					
Made you less inclined to go out?					
Stopped you doing hobbies that you used to do?					
Made you worry about the future?					
Affected your sleep?					
Made you feel more tired than usual?					

5. Treatment side effects

a. How much have you been affected by side effects from your medication in the last 3 days?

Please mark it on the scale below where 0 = unaffected and 10 = severely affected

0 1 2 3 4 5 6 7 8 9 10

b. In the last 3 days, have you been bothered by any of the following side effects of your steroid medication?

	No, I'm not affected by this	**Yes but I'm not bothered by it**	**Yes and it's affected me a little**	**Yes and it's affected me a lot**
Weight gain				
Change in appearance (fatter face, saggy skin)				
Irritability				
Low mood				
Euphoria				
Hyperactivity				
Easy bruising				
Indigestion				
Insomnia				
Hair loss				

6. Overall well-being

Do you feel back to the level of health you were at before you first experienced PMR symptoms?

Yes / No

Appendix 7.8: PMR-PROM Version 5

Today's date:

Personal Details:

Name:

Date of birth:

Gender: Male / Female

Month and year of PMR diagnosis:

Have you ever been referred to a rheumatologist about your PMR?:

Current dose of prednisolone (if applicable):

Other current medication or therapies:

The following questions ask about your symptoms of polymyalgia rheumatica and the way in which it is affecting you at the moment.
If you are unsure about how to answer a question, please give the best answer you can.
Thankyou.

1. Physical Symptoms

Pain

a. How severe has the pain **from your PMR** been during the last 3 days?
Please circle the answer below where 0 = no pain and 10 = the worst pain you've ever felt.

0 1 2 3 4 5 6 7 8 9 10

b. On average, how much of each day has the pain **from your PMR** been present for during the last 3 days
Please circle the answer that most closely applies to you.

All day About half the day Around 1-3 hours Less than 1 hour Less than 30 mins

Stiffness

c. How severe has the stiffness **caused by your PMR** been during the last 3 days?
Please circle the answer below where 0 = no stiffness and 10 = the worst stiffness you've ever felt.

0 1 2 3 4 5 6 7 8 9 10

d. On average, how much of each day has the stiffness **from your PMR** been present for during the last 3 days?
Please circle the answer that most closely applies to you.

All day About half the day Around 1-3 hours Less than 1 hour Less than 30 mins

Weakness

e. How severe has the weakness **caused by your PMR** been during the last 3 days?
Please circle the answer below where 0 = no weakness and 10 = the worst weakness you've ever felt.

0 1 2 3 4 5 6 7 8 9 10

f. On average, how much of each day has the weakness **from your PMR** been present for during the last 3 days?
Please circle the answer that most closely applies to you.

All day About half the day Around 1-3 hours Less than 1 hour Less than 30 mins

2. Function

Over the last 3 days, compared to what you can normally do, has PMR limited your ability to do the following activities?

	No, not limited at all	Yes, limited a little	Yes, limited a lot	Not relevant
Bend down				
Get up after bending down				
Get in or out of a car				
Drive a car				
Get in or out of bed				
Turn over in bed				
Get in or out of a chair				
Get in or out of the bath				
Wash yourself fully				
Dry yourself fully after a shower / bath				
Get dressed				
Take your coat on or off				
Put on or take off your socks and shoes				

	No, not limited at all	**Yes, limited a little**	**Yes, limited a lot**	**Not relevant**
Comb or blow dry your hair				
Get on or off the toilet				
Wipe yourself after going to the toilet				
Engage in intimate / sexual activity				
Walk up or down stairs				
Walk up hills				
Walk on the flat				
Carry or lift things				
Reach above your head for things				
Grip objects				
Do housework				
Do gardening				
Sit for more than 30 minutes at a time				
Participate in sports or other hobbies				

3. Emotional and psychological well-being

In the last 3 days have your PMR symptoms….

	No, not at all	A little of the time	Some of the time	Most of the time	All of the time
Caused you to feel low in mood?					
Caused you to feel anxious?					
Caused you to feel vulnerable?					
Lowered your self-confidence?					
Made you worried that you might fall over?					
Caused you to feel you need more help with looking after yourself?					
Made you feel less inclined to go out?					
Made you less interested in doing hobbies that you used to do?					
Made you worry about the future?					
Affected your sleep?					
Made you feel more tired than usual?					

5. Treatment side effects

a. Are you still taking prednisolone (steroid) tablets? Yes / No

If not, please move onto question 6.

b. What dose of prednisolone are you currently taking?

c. How much have you been affected by side effects from your prednisolone medication in the last 3 days?

Please circle the answer below where 0 = unaffected and 10 = severely affected

0 1 2 3 4 5 6 7 8 9 10

d. In the last 3 days, have you been bothered by any of the following side effects of your prednisolone medication?

	No, I'm not affected by this	**Yes but I'm not bothered by it**	**Yes and it's affected me a little**	**Yes and it's affected me a lot**
Weight gain				
Change in appearance (fatter face, saggy skin)				
Sleep disturbance				
Stomach upset or heartburn				
Mood disturbance				
Increased appetite				
Weakness of muscles				
Thin skin or easy bruising				

	No, I'm not affected by this	Yes but I'm not bothered by it	Yes and it's affected me a little	Yes and it's affected me a lot
Swelling of the feet or ankles				
High blood pressure				
High blood sugars				
Cataracts of the eyes				
Hair loss				

6. Overall well-being

Do you feel back to the level of health you were at before you first experienced PMR symptoms?
Please circle your response below

Yes, completely Yes, partially No, not at all

Thankyou very much for completing this questionnairre

Appendix 8.1: Practice invitation letter

Practice Address

Research Institute of Primary Care and Health Sciences,
David Weatherall Building
Keele University
Keele
Stoke on Trent
Stafforshire ST5 5BG

Telephone:
Fax:
Email: h.j.twohig@keele.ac.uk

Date

Dear Practice Manager,

Research project: Development of a questionnaire to assess polymyalgia rheumatica

I am writing to see if your practice would be willing to participate in the above research project, which is being led by Dr Helen Twohig, a GP and researcher from the Research Institute of Primary Care and Health Sciences, Keele University with the support of the West Midlands Clinical Research Network.

The overall aim of this project is to develop and refine a patient reported outcome measure (questionnaire) for polymyalgia rheumatica, a condition that can be difficult to assess and manage. We have developed an initial questionnaire that now needs testing and modifying. We envisage that the final questionnaire will be used in future research studies of PMR to provide better evidence for managing the condition and may also be used directly by patients and doctors in everyday practice.

If you agree to take part in the study, we will ask you to carry out a search for patients diagnosed with PMR within the last 2 years. A GP or research nurse from your practice will need to screen the notes of those patients who meet the inclusion criteria to ensure they are suitable to participate. Those that are identified as suitable will then need to be sent an invitation pack, which includes a patient information leaflet, the questionnaires and a pre-paid envelope addressed to the research office. No further involvement is needed from your practice.

If you would like to discuss the project further or confirm willingness to participate, please email me at h.j.twohig@keele.ac.uk

Yours sincerely

Appendix 8.2: Participant invitation letter

[General Practice name]
[Address line 1]
[Address line 2]
[Address line 3]
[Postcode]

[Date]

[Name]
[Address line 1]
[Address line 2]
[Address line 3]
[Postcode]

Dear [Mrs/Ms/Mr Name]

Research Project: Development of a questionnaire to assess polymyalgia rheumatica (PMR)

I am writing to invite you to take part in a research project, which aims to develop a questionnaire to assess the effects of PMR.

A research team from Keele University have developed a questionnaire. They now need to get lots of people with PMR to complete the questionnaire so that they can look at the responses and make changes to improve it.

An information sheet about the project is enclosed. It tells you about the project in more detail as well as how to contact the researchers if you have any questions or would like some more information.

If, after reading this information, you feel happy to take part, please complete the questionnaires and send them back to the research team in the envelope provided.

Thank you for taking the time to read this.

Yours sincerely

[Insert GP name]

Appendix 8.3: Participant information sheet

Development of a questionnaire to assess polymyalgia rheumatica

What is this study about?

People with polymyalgia rheumatica (PMR) can have many different symptoms which can get better or worse at different times during the course of the illness. People with it often have to stay on medication (usually steroid tablets) for around 2 years and sometimes longer. It is important to be able to fully understand how PMR is affecting a person at any point in time to know if treatments are working and to help patients and their doctors make decisions about the best management.

We want to develop a questionnaire that assesses all aspects of how PMR is affecting someone. This questionnaire will be used in research studies about PMR and directly by people with the condition and their doctors. We have developed an initial questionnaire from information from interviews with people with PMR. This now needs to be tested to allow it to be shortened and improved.

Why have I been invited?

We asked your general practitioner to identify suitable patients who developed PMR in the last 2 years.

They have sent this information to you, but have not given us any information about you.

Do I have to take part?

No, you do not have to take part in the study if you do not want to.

If you decide not to take part, you don't have to give a reason and the service you receive will not be affected.

If you return the questionnaires, we will take it that you are agreeing for the information you provide to be used for our study. All the information will be anonymous.

What do I have to do if I decide to take part?

If you agree to take part, please complete **2 copies** of the questionnaire:

- Fill in the **blue** one answering about how you feel at the moment.

- Fill in the **yellow** one answering about how you felt at the time your doctor told you you had PMR, before you started treatment. This may be some time ago now but try to remember how you felt then and answer the questions as well as you can.

Please put **both questionnaires** in the envelope provided and post it back to the study team.

How will the information from this study be used?

The answers to the questions will be analysed and used to make changes to the questionnaire to improve it.

Will there be any direct benefit from taking part in the study?

There is no direct benefit to your medical care to taking part in the study but you will be helping researchers and doctors to better understand experiences of people in your situation and this may improve care for others in the future.

Will there be any harm from taking part in the study?

It is very unlikely. If you find any of the questions upsetting you do not have to answer them.

Who will have access to my personal information?

No identifiable personal details will be collected. The questionnaire responses will be looked at by the research team.

Sometimes study information is checked by the NHS or Keele University to ensure that the research is being conducted properly.

Who has reviewed the study?

The North East - York Research Ethics Committee has reviewed this study (IRAS ID 241085).

Who is doing the research?

This study is being carried out by Dr Helen Twohig, Dr Sara Muller, Dr Caroline Mitchell and Professor Christian Mallen who are all GPs and / or researchers linked to Keele University. It is funded by a Wellcome Trust Doctoral Fellowship awarded to Helen Twohig.

Who can I contact for further information?

If you have any questions or would like more information about this study before deciding whether to take part, please contact the research team using the contact details below.

Who can I contact if I have any complaints about the study?

If you wish to voice concerns or complain about the research in any way please contact the research governance team at Keele University using the contact details below, or your local NHS complaints department.

RESEARCH TEAM CONTACT

Dr Helen Twohig,
Research Institute of Primary Care and Health Sciences
David Weatherall Building,
Keele University, Staffordshire
ST5 5BG
Tel: 01782 733371
h.j.twohig@keele.ac.uk

INDEPENDENT CONTACT AT KEELE UNIVERSITY

Dr Clark Crawford
Head of Research Integrity, Directorate of Research,
Innovation and Engagement,
Innovation Centre 2,
Keele University, Staffordshire, ST5 5NH
01782 733371
research.governance@keele.ac.uk

Appendix 8.4: Confirmation of ethical approval

Health Research Authority

North East - York Research Ethics Committee

NHSBT Newcastle Blood Donor Centre
Holland Drive
Newcastle upon Tyne
NE2 4NQ

Tel: 0207 104 8082

19 April 2018

Dr Helen Twohig
Research Institute for Primary Care and Health Sciences
Keele University
Keele
ST55BG

Dear Dr Twohig

Study title:	**Development of a patient reported outcome measure for polymyalgia rheumatica: Stage 1 psychometric evaluation - item reduction and formation of questionnaire structure**
REC reference:	**18/NE/0140**
Protocol number:	**RG-0272-18-IPCHS**
IRAS project ID:	**241085**

The Proportionate Review Sub-committee of the North East - York Research Ethics Committee reviewed the above application on 20 April 2018.

We plan to publish your research summary wording for the above study on the HRA website, together with your contact details. Publication will be no earlier than three months from the date of this favourable opinion letter. The expectation is that this information will be published for all studies that receive an ethical opinion but should you wish to provide a substitute contact point, wish to make a request to

defer, or require further information, please contact hra.studyregistration@nhs.net outlining the reasons for your request. Under very limited circumstances (e.g. for student research which has received an unfavourable opinion), it may be possible to grant an exemption to the publication of the study.

Ethical opinion

On behalf of the Committee, the sub-committee gave a favourable ethical opinion of the above research on the basis described in the application form, protocol and supporting documentation, subject to the conditions specified below.

Conditions of the favourable opinion

The REC favourable opinion is subject to the following conditions being met prior to the start of the study.

Management permission must be obtained from each host organisation prior to the start of the study at the site concerned.

Management permission should be sought from all NHS organisations involved in the study in accordance with NHS research governance arrangements. Each NHS organisation must confirm through the signing of agreements and/or other documents that it has given permission for the research to proceed (except where explicitly specified otherwise).

Guidance on applying for HRA Approval (England)/ NHS permission for research is available in the Integrated Research Application System, www.hra.nhs.uk or at http://www.rdforum.nhs.uk.

Where a NHS organisation's role in the study is limited to identifying and referring potential participants to research sites ("participant identification centre"), guidance should be sought from the R&D office on the information it requires to give permission for this activity.

For non-NHS sites, site management permission should be obtained in accordance with the procedures of the relevant host organisation.

Sponsors are not required to notify the Committee of management permissions from host organisations.

Registration of Clinical Trials

All clinical trials (defined as the first four categories on the IRAS filter page) must be registered on a publically accessible database. This should be before the first participant is recruited but no later than 6 weeks after recruitment of the first participant.

There is no requirement to separately notify the REC but you should do so at the earliest opportunity e.g. when submitting an amendment. We will audit the registration details as part of the annual progress reporting process.

To ensure transparency in research, we strongly recommend that all research is registered but for non-clinical trials this is not currently mandatory.

If a sponsor wishes to request a deferral for study registration within the required timeframe, they should contact hra.studyregistration@nhs.net. The expectation is that all clinical trials will be registered, however, in exceptional circumstances non registration may be permissible with prior agreement from the HRA. Guidance on where to register is provided on the HRA website.

It is the responsibility of the sponsor to ensure that all the conditions are complied with before the start of the study or its initiation at a particular site (as applicable).

Ethical review of research sites

The favourable opinion applies to all NHS sites taking part in the study, subject to management permission being obtained from the NHS/HSC R&D office prior to the start of the study (see
"Conditions of the favourable opinion").

Approved documents

The documents reviewed and approved were:

Document	*Version*	*Date*
Covering letter on headed paper [Covering letter]	1	09 February 2018
Evidence of Sponsor insurance or indemnity (non NHS Sponsors only) [Indemnity certificate]		31 July 2017
GP/consultant information sheets or letters [Practice information letter]	1	09 February 2018
HRA Schedule of Events [Schedule of events]	1	13 February 2018
HRA Statement of Activities [Statement of activities]	1	05 March 2018
IRAS Application Form [IRAS_Form_03042018]		03 April 2018
IRAS Application Form XML file [IRAS_Form_03042018]		03 April 2018
Letter from sponsor [Sponsor letter]		16 March 2018
Letters of invitation to participant [Covering letter to patients]	1	09 February 2018
Non-validated questionnaire [PMR PROM]	5	09 February 2018
Other [CV for Christian Mallen]	1	01 July 2016
Other [CV for Caroline Mitchell]		15 August 2017
Participant information sheet (PIS) [Participant information sheet]	1	05 March 2018
Research protocol or project proposal [Protocol]	1	22 March 2018

Summary CV for Chief Investigator (CI) [CV - Helen Twohig]		
Summary CV for supervisor (student research) [Supervisor CV - Sara Muller]	1	05 March 2018

Membership of the Proportionate Review Sub-Committee

The members of the Sub-Committee who took part in the review are listed on the attached sheet.

Statement of compliance

The Committee is constituted in accordance with the Governance Arrangements for Research Ethics Committees and complies fully with the Standard Operating Procedures for Research Ethics Committees in the UK.

After ethical review

Reporting requirements

The attached document "After ethical review – guidance for researchers" gives detailed guidance on reporting requirements for studies with a favourable opinion, including:

- Notifying substantial amendments
- Adding new sites and investigators
- Notification of serious breaches of the protocol
- Progress and safety reports
- Notifying the end of the study

The HRA website also provides guidance on these topics, which is updated in the light of changes in reporting requirements or procedures.

User Feedback

The Health Research Authority is continually striving to provide a high quality service to all applicants and sponsors. You are invited to give your view of the service you have received and the application procedure. If you wish to make your views known please use the feedback form available on the HRA website:
http://www.hra.nhs.uk/about-the-hra/governance/quality-assurance/

HRA Training

We are pleased to welcome researchers and R&D staff at our training days – see details at http://www.hra.nhs.uk/hra-training/

With the Committee's best wishes for the success of this project.

18/NE/0140	**Please quote this number on all correspondence**

Yours sincerely

pp

Mr Chris Turnock Chair

Email: nrescommittee.northeast-york@nhs.net

Enclosures: *List of names and professions of members who took part in the review*

"After ethical review – guidance for researchers"

Copy to: *Dr Clark Crawford, Keele University*
Primary Care Research Support Team , NIHR CRN West Midlands

North East - York Research Ethics Committee

Attendance at PRS Sub-Committee of the REC meeting on 19 April 2018 via correspondence.

Committee Members:

Name	*Profession*	*Present*	*Notes*
Ms Linda Chadd	Library assistant and archivist	Yes	
Dr Jocelyn Hudson	General Practitioner	Yes	
Mr Chris Turnock (Chair)	Head of Technology Enhanced Learning	Yes	

Also in attendance:

Name	*Position (or reason for attending)*
Miss Kerry Dunbar	REC Manager

Appendix 8.5: Results tables of distribution of item responses

Percentage distribution of responses to question on pain, stiffness and weakness severity at diagnosis

	Missing	0	1	2	3	4	5	6	7	8	9	10
Pain severity	0.8	5.2	2.0	0.79	2.4	4.8	5.6	5.2	13.9	20.2	19.0	20.2
Stiffness severity	1.2	5.2	1.6	2.4	2.4	2.4	6.7	7.5	13.5	20.6	15.5	21.0
Weakness severity	1.2	4.4	1.2	2.0	3.5	4.8	9.5	8.7	13.1	22.2	13.4	15.9

Question asked: how severe has the (pain / stiffness / weakness) from your PMR been during the last 3 days?

Responses options: visual analogue scale (VAS), scored from 0-10 where 0 = no pain and 10 = the worst pain you have ever felt.

Percentage distribution of responses to question on pain, stiffness and weakness duration at diagnosis

	Missing	<30 mins	<1hr	1-3 hrs	About half the day	All day
Pain duration	5.6	6.3	1.6	8.3	22.2	59.5
Stiffness duration	2.4	5.6	4.0	8.7	19.4	60.0
Weakness duration	3.6	3.6	3.6	6.7	20.6	62.0

Question asked: on average much of each day has the (pain / stiffness / weakness) from your PMR been present for during the last 3 days?

Response options: 1 = less than 30 mins, 2 = less than 1 hour, 3 = around 1-3 hours, 4 = about half the day, 5 = all day

Percentage distribution of responses to question on pain, stiffness and weakness severity now

	Missing	0	1	2	3	4	5	6	7	8	9	10
Pain severity	7.4	19.1	10.9	12.1	12.9	9.4	8.2	4.3	5.5	4.7	1.6	3.9

Stiffness severity	7.0	16.8	12.9	9.4	11.7	8.6	11.3	5.9	6.3	3.1	3.5	3.5
Weakness severity	3.1	19.5	9.0	9.4	14.9	9.4	9.0	5.1	5.1	7.4	1.2	3.1

Question asked: how severe has the (pain / stiffness / weakness) from your PMR been during the last 3 days?

Responses options: visual analogue scale (VAS), scored from 0-10 where 0 = no pain and 10 = the worst pain you have ever felt.

Percentage distribution of responses to question on pain, stiffness and weakness duration now

	Missing	<30 mins	<1hr	1-3 hrs	About half the day	All day
Pain duration	10.1	27.0	8.2	18.0	17.6	19.1
Stiffness duration	10.5	25.4	9.0	18.0	16.0	21.1
Weakness duration	9.4	23.7	7.8	17.6	18.8	22.7

Question asked: on average much of each day has the (pain / stiffness / weakness) from your PMR been present for during the last 3 days?

Response options: 1 = less than 30 mins, 2 = less than 1 hour, 3 = around 1-3 hours, 4 = about half the day, 5 = all day

Percentage distribution of responses to functional activity items at diagnosis

Item	Not limited	Limited a little	Limited a lot	Not relevant	Missing
Bend down	11.1	26.6	57.9	2.8	1.6
Get up after bending down	9.1	22.2	63.1	3.2	2.4
Get in or out of a car	9.5	31.8	55.2	2.0	1.6
Drive a car	23.8	24.2	23.8	23.4	4.8
Get in or out of bed	8.3	25.4	62.7	1.6	2.0
Turn over in bed	9.1	23.8	63.9	2.0	1.2

Get in or out of a chair	8.3	31.8	54.4	2.4	3.2
Get in or out of the bath	7.1	12.7	48.8	28.6	2.8
Wash yourself fully	20.6	34.5	40.5	1.6	2.8
Dry yourself fully	17.1	41.7	38.1	2.0	1.2
Get dressed	11.1	44.1	41.2	1.6	1.6
Take your coat on or off	11.5	41.7	43.6	1.6	1.6
Put on or take off your socks	8.7	23.8	64.7	1.6	1.2
Comb or blow dry your hair	16.7	34.5	42.5	4.8	1.6
Get on or off the toilet	21.4	39.7	35.3	1.2	2.4
Wipe yourself after going to the toilet	28.2	34.1	32.9	2.4	2.4
Intimate activity	8.3	14.2	17.9	51.2	8.3
Walk up or down stairs	9.1	28.2	51.2	8.3	3.2
Walk up hills	8.3	24.6	56.0	9.1	2.0
Walk on the flat	18.3	44.5	33.3	1.2	2.4
Carry or lift things	7.5	29.8	60.3	1.2	1.2
Reach above your head for things	8.3	20.6	68.3	1.6	1.2
Grip objects	17.9	42.5	36.5	1.2	1.2
Do housework	11.1	33.7	47.2	6.4	1.6
Do gardening	5.6	20.6	61.5	9.5	2.3
Sit for >30mins at a time	27.4	32.9	35.7	1.6	2.4

Participate in sports / hobbies	6.0	14.7	52.8	22.6	4.0

Question asked: Over the last 3 days, compared to what you can normally do, has PMR limited your ability to do the following activities?

Percentage distribution of responses to functional activity items now

Item	Not limited	Limited a little	Limited a lot	Not relevant	Missing
Bend down	34.4	38.3	18.4	2.7	6.3
Get up after bending down	26.6	39.5	24.6	3.1	6.3
Get in or out of a car	34.8	41.8	14.8	3.1	5.5
Drive a car	55.0	14.3	3.4	19.9	7.2
Get in or out of bed	40.2	41.8	10.9	2.3	4.7
Turn over in bed	39.2	38.4	14.9	2.4	5.1
Get in or out of a chair	39.5	41.0	11.7	2.7	5.1
Get in or out of the bath	22.9	20.6	24.1	25.7	6.7
Wash yourself fully	55.5	30.5	5.9	3.1	5.1
Dry yourself fully	51.6	35.2	5.9	2.7	4.7
Get dressed	43.0	43.8	6.3	2.0	5.1
Take your coat on or off	46.9	35.9	9.4	2.3	5.5
Put on or take off your socks	35.9	35.2	21.1	2.7	5.1
Comb or blow dry your hair	51.2	31.6	8.2	5.5	3.5
Get on or off the toilet	58.2	29.7	5.5	2.7	3.9
Wipe yourself	61.3	23.8	8.2	2.7	3.9

after going to the toilet					
Intimate activity	24.0	14.1	5.1	48.8	7.9
Walk up or down stairs	30.5	38.7	18.4	7.8	4.7
Walk up hills	23.8	36.7	27.7	7.8	3.9
Walk on the flat	47.3	38.3	7.8	2.0	4.7
Carry or lift things	23.4	47.3	24.6	1.6	3.1
Reach above your head for things	34.0	39.5	21.5	1.6	3.5
Grip objects	43.8	33.6	17.2	1.6	3.9
Do housework	34.4	40.1	14.8	6.3	3.9
Do gardening	22.3	39.1	25.0	9.4	4.3
Sit for >30mins at a time	57.8	25.4	10.6	2.0	4.3
Participate in sports / hobbies	22.3	23.8	22.7	25.8	5.5

Question asked: Over the last 3 days, compared to what you can normally do, has PMR limited your ability to do the following activities?

Percentage distribution of responses to emotional and psychological well-being items at diagnosis

Item	Not at all	A little of the time	Some of the time	Most of the time	All of the time	Missing
Low in mood	9.9	21.0	31.0	26.6	9.5	1.98
Anxious	17.9	19.1	25.8	25.4	8.3	3.6
Vulnerable	23.0	21.0	21.0	23.8	7.9	3.2
Low self confidence	20.6	17.1	26.2	23.8	8.3	4.0
Worry about falling over	25.0	18.3	23.8	21.0	9.1	2.8

More help with self care	32.9	23.8	19.4	15.1	4.8	4
Less inclined to go out	26.2	20.6	23.0	19.8	7.9	2.4
Less interested in hobbies	17.9	17.5	23.4	27.0	10.7	3.6
Worry about the future	18.3	19.8	22.6	21.4	15.1	2.8
Affected sleep	32.7	22.8	20.5	13.0	7.5	3.5
More tired than usual	22.0	25.1	17.3	21.2	11.0	3.5

Question asked: Over the last 3 days, have your PMR symptoms….?

Percentage distribution of responses to psychological and emotional well-being items now

Item	Not at all	A little of the time	Some of the time	Most of the time	All of the time	Missing
Low in mood	32.4	30.1	23.4	7.4	2.0	4.7
Anxious	43.4	25.8	16.8	6.3	2.3	5.5
Vulnerable	46.9	24.2	15.6	5.5	2.7	5.1
Low self confidence	41.4	27.7	13.7	8.6	3.1	5.5
Worry about falling over	40.2	27.7	13.3	10.6	3.9	4.3
More help with self care	62.1	19.5	8.2	3.9	1.6	4.7
Less inclined to go out	56.9	15.7	14.5	4.7	3.1	5.1
Less interested in hobbies	43.5	18.0	16.5	10.6	5.1	6.3

Worry about the future	39.2	23.9	14.1	12.9	4.7	5.1
Affected sleep	32.7	22.8	20.5	13.0	7.5	3.5
More tired than usual	22.0	25.1	17.3	21.2	11.0	3.5

Question asked: Over the last 3 days, have your PMR symptoms….?

Appendix 8.6: PMR-PROM Version 6

Keele UNIVERSITY

A Patient Reported Outcome Measure for Polymyalgia Rheumatica (PMR)

The following questionnaire asks about your symptoms of polymyalgia rheumatica and the way in which it is affecting you at the moment.
If you are unsure about how to answer a question, please give the best answer you can.
Thank you.

Today's date:

Personal Details:

Age:

Gender: Male Female

Length of time since PMR diagnosis:

Have you been referred to a rheumatologist about your PMR?:

Current dose of prednisolone if taking:

Other current medication or therapies:

1. Symptoms

Pain

a. How severe has the pain **from your PMR** been during the last 3 days?
Please circle the answer below where 0 = no pain and 10 = the worst pain you've ever felt.

0 1 2 3 4 5 6 7 8 9 10

b. On average, how much of each day has the pain **from your PMR** been present for during the last 3 days
Please circle the answer that most closely applies to you.

None *Less than 1 hour* *Around 1-3 hours* *About half the day* *All day*

Stiffness

c. How severe has the stiffness **caused by your PMR** been during the last 3 days?
Please circle the answer below where 0 = no stiffness and 10 = the worst stiffness you've ever felt.

0 1 2 3 4 5 6 7 8 9 10

d. On average, how much of each day has the stiffness **from your PMR** been present for during the last 3 days?
Please circle the answer that most closely applies to you.

None *Less than 1 hour* *Around 1-3 hours* *About half the day* *All day*

Weakness

e. How severe has the weakness **caused by your PMR** been during the last 3 days?
Please circle the answer below where 0 = no weakness and 10 = the worst weakness you've ever felt.

0 1 2 3 4 5 6 7 8 9 10

f. On average, how much of each day has the weakness **from your PMR** been present for during the last 3 days?
Please circle the answer that most closely applies to you.

None *Less than 1 hour* *Around 1-3 hours* *About half the day* *All day*

Fatigue

g. How severe has the fatigue **caused by your PMR** been during the last 3 days?
Please circle the answer below where 0 = no fatigue and 10 = the worst fatigue you've ever felt.

0 1 2 3 4 5 6 7 8 9 10

h. On average, how much of each day has the fatigue **from your PMR** been present for during the last 3 days?
Please circle the answer that most closely applies to you.

None *Less than 1 hour* *Around 1-3 hours* *About half the day* *All day*

2. Function

Over the last 3 days, compared to what you can normally do, has PMR limited your ability to do the following activities?

	No, not limited at all	**Yes, limited a little**	**Yes, limited a lot**
Get in or out of a car			
Get in or out of bed			
Turnover in bed			
Wash yourself fully			
Put on or take off your socks and shoes			
Get on or off the toilet			
Walk up or down stairs			
Carry or lift things			
Reach above your head for things			

3. Emotional and psychological well-being

In the last 3 days have your PMR symptoms….

	No, not at all	**A little of the time**	**Some of the time**	**Most of the time**	**All of the time**
Caused you to feel low in mood?					
Caused you to feel anxious?					
Caused you to feel vulnerable?					
Lowered your self-confidence?					

5. Treatment side effects

In the last 3 days, have you been bothered the following side effects of your prednisolone medication?

	No, I'm not affected by this	**Yes but I'm not bothered by it**	**Yes and it's affected me a little**	**Yes and it's affected me a lot**
Weight gain				
Change in appearance (fatter face, saggy skin)				
Sleep disturbance				
Stomach upset or heartburn				
Mood disturbance				
Increased appetite				
Weakness of muscles				
Thin skin or easy bruising				
Swelling of the feet or ankles				
High blood pressure				
High blood sugars				
Cataracts of the eyes				
Hair loss				

6. Overall well-being

Do you feel back to the level of health you were at before you first experienced PMR symptoms?
Please circle your response below

Yes, completely Yes, partially No, not at all

Thank you very much for completing this questionnaire

Appendix 8.7: PMR-PROM Version 7

A Patient Reported Outcome Measure for Polymyalgia Rheumatica (PMR)

The following questionnaire asks about your symptoms of polymyalgia rheumatica and the way in which it is affecting you at the moment.
If you are unsure about how to answer a question, please give the best answer you can.
Thank you.

Today's date:

Your Personal Details:

Name:

Age:

Gender: **Male** ☐
Female ☐

Length of time since PMR diagnosis:

Have you been referred to a rheumatologist about your PMR?: **Yes** ☐
No ☐

Current dose of prednisolone if taking:

Other current medication:

Other therapy you are having e.g. physio:

1. Symptoms

PAIN

a. How severe has the pain **from your PMR** been during the last 3 days?
Please circle the answer below where 0 = no pain and 10 = the worst pain you've ever felt.

0 1 2 3 4 5 6 7 8 9 10

No pain Severe pain

b. On average, how much of each day has the pain **from your PMR** been present for during the last 3 days
Please circle the answer that most closely applies to you.

None *Less than 1 hour* *Around 1-3 hours* *About half the day* *All day*

STIFFNESS

c. How severe has the stiffness **caused by your PMR** been during the last 3 days?
Please circle the answer below where 0 = no stiffness and 10 = the worst stiffness you've ever felt.

0 1 2 3 4 5 6 7 8 9 10

No stiffness Severe

d. On average, how much of each day has the stiffness **from your PMR** been present for during the last 3 days?
Please circle the answer that most closely applies to you.

None *Less than 1 hour* *Around 1-3 hours* *About half the day* *All day*

WEAKNESS

e. How severe has the weakness **caused by your PMR** been during the last 3 days?
Please circle the answer below where 0 = no weakness and 10 = severe weakness.

0 1 2 3 4 5 6 7 8 9 10

No weakness Severe

f. On average, how much of each day has the weakness **from your PMR** been present for during the last 3 days?
Please circle the answer that most closely applies to you.

None *Less than 1 hour* *Around 1-3 hours* *About half the day* *All day*

FATIGUE

g. How severe has the fatigue **caused by your PMR** been during the last 3 days?
Please circle the answer below where 0 = no fatigue and 10 = severe weakness

0 1 2 3 4 5 6 7 8 9 10

No fatigue Severe

h. On average, how much of each day has the fatigue **from your PMR** been present for during the last 3 days?
Please circle the answer that most closely applies to you.

None *Less than 1 hour* *Around 1-3 hours* *About half the day* *All day*

2. **Function**

Over the last 3 days, compared to what you can normally do, has **your PMR** limited your ability to do the following activities?

	No, not limited at all	**Yes, limited a little**	**Yes, limited a lot**
Get in or out of a car			
Get in or out of bed			
Turnover in bed			
Wash yourself fully			
Put on or take off your socks and shoes			
Get on or off the toilet			
Walk up or down stairs			
Carry or lift things			
Reach above your head for things			

3. Emotional and psychological well-being

In the last 3 days have **your PMR** symptoms….

	No, not at all	**A little of the time**	**Some of the time**	**Most of the time**	**All of the time**
Caused you to feel low in mood?					
Caused you to feel anxious?					
Caused you to feel vulnerable?					
Lowered your self-confidence?					

4. Treatment side effects

In the last 3 days, have you been bothered by any of the following side effects of **your prednisolone** medication?

	No, I'm not affected by this	**Yes, but I'm not bothered by it**	**Yes and it's affected me a little**	**Yes and it's affected me a lot**
Weight gain				
Change in appearance (fatter face, saggy skin)				
Sleep disturbance				
Stomach upset or heartburn				
Mood disturbance				
Increased appetite				
Muscle weakness				
Easily bruised or thin skin				
Swelling of the feet or ankles				
High blood pressure				
High blood sugars				
Cataracts of the eyes				
Hair loss				

5. Overall well-being

Do you feel back to the level of health you were at before you first experienced PMR symptoms?
Please circle your response below

Yes, completely Yes, partially No, not at all

Thank you very much for completing this questionnaire

Appendix 8.8: PMR-PROM Version 8

Keele UNIVERSITY

A Patient Reported Outcome Measure for Polymyalgia Rheumatica (PMR)

The following questionnaire asks about your symptoms of polymyalgia rheumatica and the way in which it is affecting you at the moment.
If you are unsure about how to answer a question, please give the best answer you can.
Thank you.

Today's date:

Your Personal Details:

Name:

Age:

Gender: **Male** ☐
Female ☐

Length of time since PMR diagnosis:

Have you been referred to a rheumatologist about your PMR?: **Yes** ☐
No ☐

Current dose of prednisolone if taking:

Other current medication:

Other therapy you are having e.g. physio:

1. Symptoms

Thinking about how your symptoms **from your PMR** have affected you **during the last three days**, please answer the following questions by circling the response that applies most closely to you

Pain

How bad has your pain been?

0 1 2 3 4 5 6 7 8 9 10

No pain Severe pain

On average, how much of each day has the pain been present for?

None *Less than 1 hour* *Around 1-3 hours* *About half the day* *All day*

Stiffness

How bad has your stiffness been?

0 1 2 3 4 5 6 7 8 9 10

No stiffness Severe

On average, how much of each day has the stiffness been present for?

None *Less than 1 hour* *Around 1-3 hours* *About half the day* *All day*

Weakness

How much weakness have you experienced?

0 1 2 3 4 5 6 7 8 9 10

No weakness Severe

On average, how much of each day has any weakness been present for?

None *Less than 1 hour* *Around 1-3 hours* *About half the day* *All day*

Fatigue

How fatigued have you felt?

0 1 2 3 4 5 6 7 8 9 10

No fatigue Severe

On average, how much of each day have you felt fatigued?

None *Less than 1 hour* *Around 1-3 hours* *About half the day* *All day*

2. Function

Over the last 3 days, has **your PMR** limited your ability to do the following activities?

	No, not limited at all	**Yes, limited a little**	**Yes, limited a lot**
Get in or out of a car			
Get in or out of bed			
Turnover in bed			
Wash yourself fully			
Put on or take off your socks and shoes			
Get on or off the toilet			
Walk up or down stairs			
Carry or lift things			
Reach above your head for things			

3. Emotional and psychological well-being

In the last 3 days have **your PMR** symptoms….

	No, not at all	**A little of the time**	**Some of the time**	**Most of the time**	**All of the time**
Caused you to feel low in mood?					
Caused you to feel anxious?					
Caused you to feel vulnerable?					
Lowered your self-confidence?					

4. Treatment side effects

In the last 3 days, have you been bothered by any of the following side effects of **your prednisolone** medication?

	No, I'm not affected by this	**Yes, but I'm not bothered by it**	**Yes and it's affected me a little**	**Yes and it's affected me a lot**
Weight gain				
Change in appearance (fatter face, saggy skin)				
Sleep disturbance				
Stomach upset or heartburn				
Mood disturbance				
Increased appetite				
Muscle weakness				
Easily bruised or thin skin				
Swelling of the feet or ankles				
Hair loss				

5. Overall well-being

Do you feel back to the level of health you were at before you first experienced PMR symptoms?
Please circle your response below

Yes, completely Yes, partially No, not at all

Thank you very much for completing this questionnaire

Appendix 8.9: PMR-PROM Version 9

Keele UNIVERSITY

The Polymyalgia Rheumatica Impact Scale (PMR-IS)

The following questionnaire asks about your symptoms of polymyalgia rheumatica and the way in which it is affecting you at the moment.
If you are unsure about how to answer a question, please give the best answer you can.
Thank you.

Today's date	
PERSONAL DETAILS	
Name	
Age	
Gender	**Male** ☐ **Female** ☐
Length of time since PMR diagnosis	
Have you been referred to a rheumatologist about your PMR?	**Yes** ☐ **No** ☐
Current dose of prednisolone if taking:	
Other current medication	
Other therapy you are having e.g. physio:	

1. Symptoms

Thinking about how your symptoms **from your PMR** have affected you

during the last three days, please answer the following questions by circling

the response that applies most closely to you.

Pain

How bad has your pain **caused by your PMR** been during the last three days?

1	2	3	4	5	6	7	8	9	10
No Pain									**Severe Pain**

On average, how much of each day has the pain been present for?

None	***Less than 1 hour***	***Around 1-3 hours***	***About half the day***	***All day***

Stiffness

How bad has your stiffness **caused by your PMR** been during the last three days?

1	2	3	4	5	6	7	8	9	10
No Stiffness									**Severe Stiffness**

On average, how much of each day has the stiffness been present for?

None	***Less than 1 hour***	***Around 1-3 hours***	***About half the day***	***All day***

Weakness

How much weakness **caused by your PMR** have you experienced during the last three days?

1	2	3	4	5	6	7	8	9	10
No Weakness									**Severe Weakness**

On average, how much of each day has any weakness been present for?

None	***Less than 1 hour***	***Around 1-3 hours***	***About half the day***	***All day***

Fatigue

How severe has the fatigue **caused by your PMR** been in the last three days?

1	2	3	4	5	6	7	8	9	10
No Fatigue									**Severe Fatigue**

On average, how much of each day have you felt fatigued?

None	***Less than 1 hour***	***Around 1-3 hours***	***About half the day***	***All day***

2. Function

Over the last 3 days, has **your PMR** limited your ability to do the following activities?

	Not limited at all	Moderately Limited	Severely Limited
Get in or out of a car			
Get in or out of bed			
Turnover in bed			
Wash yourself fully			
Put on or take off your socks and shoes			
Get on or off the toilet			
Walk up or down stairs			
Carry or lift things			
Reach above your head for things			

3. Emotional and psychological well-being

In the last 3 days have **your PMR** symptoms….

	No, not at all	**A little of the time**	**Some of the time**	**Most of the time**	**All of the time**
Caused you to feel low in mood?					
Caused you to feel anxious?					
Caused you to feel vulnerable?					
Lowered your self-confidence?					

4. Treatment side effects

In the last 3 days, have you been bothered by any of the following side effects of **your prednisolone** medication?

	No, I'm not affected by this	**Yes, but I'm not bothered by it**	**Yes and it's affected me a little**	**Yes and it's affected me a lot**
Weight gain				
Change in appearance (fatter face, saggy skin)				
Sleep disturbance				
Stomach upset or heartburn				
Mood disturbance				
Increased appetite				
Muscle weakness				
Easily bruised or thin skin				
Swelling of the feet or ankles				
Hair loss				

5. Overall well-being

Do you feel back to the level of health you were at before you first experienced PMR symptoms?

Please circle your response below

Yes, completely	***Yes, partially***	***No, not at all***

Thank you very much for completing this questionnaire

Appendix 8.10: PMR-PROM Version 10

Keele UNIVERSITY

The Polymyalgia Rheumatica Impact Scale (PMR-IS)

The following questionnaire asks about your symptoms of polymyalgia rheumatica (PMR) and the way in which it is affecting you at the moment.
If you are unsure about how to answer a question, please give the best answer you can.
Thank you.

Today's date:

PERSONAL DETAILS	
Name	
Date of birth	
Gender	**Male** ☐ **Female** ☐
Length of time since PMR diagnosis	
Have you been referred to a rheumatologist about your PMR?	**Yes** ☐ **No** ☐
Are you taking prednisolone?	**Yes** ☐ **No** ☐
Current dose of prednisolone if taking:	**mg**
Other current medication	
Other therapy you are having for your PMR e.g. physiotherapy:	

1. Symptoms

Thinking about how your **PMR** has affected you **during the last week**, please answer the following questions by circling one response

Pain

How bad has the pain **caused by your PMR** been during the last week?

0	1	2	3	4	5	6	7	8	9	10
No pain										**Severe pain**

On average, how much of each day has the pain been present for?

None	***Less than 1 hour***	***Around 1-3 hours***	***About half the day***	***All day***

Stiffness

How bad has the stiffness **caused by your PMR** been during the last week?

0	1	2	3	4	5	6	7	8	9	10
No stiffness										**Severe stiffness**

On average, how much of each day has the stiffness been present for?

None	***Less than 1 hour***	***Around 1-3 hours***	***About half the day***	***All day***

Weakness

How much weakness **caused by your PMR** have you experienced during the last week?

0	1	2	3	4	5	6	7	8	9	10
No weakness										**Severe weakness**

On average, how much of each day has any weakness been present for?

None	***Less than 1 hour***	***Around 1-3 hours***	***About half the day***	***All day***

Fatigue

How bad has the fatigue **caused by your PMR** been in the last week?

0	1	2	3	4	5	6	7	8	9	10
No fatigue										**Severe fatigue**

On average, how much of each day have you felt fatigued?

None	***Less than 1 hour***	***Around 1-3 hours***	***About half the day***	***All day***

2. Function

Over the last week, has **your PMR** limited your ability to do the following activities?

Please put a cross in one box on each line

	Not limited at all	Moderately Limited	Severely Limited
Get in or out of a car			
Get in or out of bed			
Turnover in bed			
Wash yourself fully			
Put on or take off your socks and shoes			
Get on or off the toilet			
Walk up or down stairs			
Carry or lift things			
Reach above your head for things			

3. Emotional and psychological well-being

In the last week have **your PMR** symptoms caused any of the following feelings?

Please put a cross in one box on each line

	No, not at all	A little of the time	Some of the time	Most of the time	All of the time
Caused you to feel low in mood?					
Caused you to feel anxious?					
Caused you to feel vulnerable?					
Lowered your self-confidence?					

4. Treatment side effects

In the last week, have you had any of the following side effects from **your prednisolone** medication?

Please put a cross in one box on each line.

	No, I'm not affected by this	**Yes, but I'm not bothered by it**	**Yes and it's affected me a little**	**Yes and it's affected me a lot**
Weight gain				
Change in appearance (fatter face, saggy skin)				
Sleep disturbance				
Stomach upset or heartburn				
Mood disturbance				
Increased appetite				
Muscle weakness				
Easily bruised or thin skin				
Swelling of the feet or ankles				
Hair loss				

5. Overall well-being

Do you feel back to the level of health you were at before you first experienced PMR symptoms?

Please circle one response below

Yes, completely	***Yes, partially***	***No, not at all***

Thank you very much for completing this questionnaire

Appendix 9.1: Practice invitation letter

Practice Address

School of Primary, Community and Social Care,
David Weatherall Building
Keele University
Keele
Stoke on Trent
Staffordshire, ST5 5BG

Telephone:
Email: h.j.twohig@keele.ac.uk

Date

Dear Practice Manager,

Research project: Evaluation of the polymyalgia rheumatica impact scale (PMR-IS)

I am writing to see if your practice would be willing to participate in identifying patients for the above research project, which is being led by Dr Helen Twohig, a GP and researcher from the School of Primary, Community and Social Care, Keele University with the support of the West Midlands Clinical Research Network.

We have developed a questionnaire (the PMR-IS) to assess PMR-related quality of life. PMR is a condition that can be difficult to assess and manage and we envisage that the questionnaire will be used in future research studies of PMR to provide better evidence for managing the condition and may also be used directly by patients and doctors in everyday practice. Before it is ready to use though, we need to evaluate the reliability, validity and responsiveness of the questionnaire. This is the aim of this current study.

If you agree to take part in the study, we will ask you to carry out a search for patients diagnosed with PMR within the last 2 years. A GP or nurse from your practice will need to screen the records of those patients who meet the inclusion criteria to ensure they are suitable to participate. Those that are identified as suitable will be sent an invitation pack via Docmail which includes a patient information leaflet, the questionnaires and a pre-paid envelope addressed to the research office. No further involvement is needed from your practice.

If you would like to discuss the project further or confirm willingness to participate, please email me at h.j.twohig@keele.ac.uk

Yours sincerely

Instructions for Identifying Participants

The practice computer system can generate a list of people who have been diagnosed with PMR in the last two years but we'd also like to check that the diagnosis is secure and that it is appropriate to contact the people on the list and invite them to complete the questionnaire. To do this we ask that the list of the patients be screened by a GP or research nurse to ensure that no one is distressed or confused by receiving the invitation, e.g. people who have been admitted to hospital, have left the practice or who currently have other significant problems in their life.

Names and address of identified potential participants will need to be securely uploaded to Docmail by a member of practice team and a study pack will then be sent out. The pack includes a pre-paid enveloped to return the questionnaire booklet to the research team directly if a person is willing to take part.

The letter and accompanying paperwork ask that potential participants contact the research team and not the practice with any questions so after the letters are sent, any contact about the study would be between the research team and the participants.

Inclusion criteria

Diagnosis of PMR **within the previous 2 years**.

We'd like you to check this diagnosis is secure and hasn't changed since it was initially made. As a guide, the diagnosis should be supported by the following features:

- Age > 50 years.
- Bilateral shoulder or pelvic girdle aching or both for at least 2 weeks.
- Morning stiffness
- Evidence of an acute phase response (raised ESR / CRP).
- If a patient has atypical features (e.g. Normal ESR / CRP) but the diagnosis has been made by a rheumatologist, they can also be included.

Exclusion criteria

Diagnosis of GCA

Inability to read / write English well enough to understand the instructions and complete the questionnaire.

Comorbidities that make invitation to participate in the study inappropriate in the view of the participant's GP (dementia, significant anxiety / depression, receiving end of life care etc.).

People who have been referred to the Haywood Hospital, Staffordshire, for their PMR (as we will be recruiting from there directly).

Appendix 9.2: Participant invitation letter

[General Practice name]
[Address line 1]
[Address line 2]
[Address line 3]
[Postcode]

[Date]

[Name]
[Address line 1]
[Address line 2]
[Address line 3]
[Postcode]

Dear [Mrs/Ms/Mr Name]

Research Project: Evaluation of the polymyalgia rheumatica impact scale (PMR-IS)

I am writing to invite you to test a questionnaire about the effects of PMR, the PMR-Impact Scale.

The questionnaire has been developed by a research team from Keele University. They now want to test it to make sure that it is reliable and that a person's score changes if their PMR improves or worsens. They also want to compare it to other similar questionnaires to see if it is better for assessing PMR.

An information sheet about the project is enclosed. It tells you more about the project and how to contact the team if you have any questions or would like more information.

If you feel happy to take part, please complete the questionnaire booklet, including the consent page, and return it in the envelope provided.

If you do not wish to take part, you do not need to reply. The research team do not have any information about you at this stage.

Thank you for taking the time to read this.

Yours sincerely

[Insert GP name]

Appendix 9.3: Participant information sheet

Participant Information Sheet

Evaluation of the Polymyalgia Rheumatica-Impact Scale (PMR-IS)

What is this study about?

There is currently no standard way of assessing how the condition polymyalgia rheumatica (PMR) and its treatment is affecting someone.

It is important to be able to fully understand how PMR is affecting a person at any point in time to help patients and their doctors make decisions about treatment.

We have developed a questionnaire, **the PMR-Impact Scale**, that assesses all aspects of how PMR is affecting someone. It was developed from information from interviews with people with PMR and has been shortened and improved by studying responses from many people with the condition.

We now want to test it to see if it is reliable and compares well to other similar questionnaires.

Why have I been invited?

We asked your doctor to identify patients who developed PMR in the last two years and invite them to take part. They have sent this information to you but have not given us any information about you.

Do I have to take part?

No, you do not have to take part in the study if you do not want to. You do not have to give a reason for this.

If you return the completed questionnaires, we will take it that you are agreeing for the information you give us to be used for our study.

What do I have to do if I decide to take part?

Please complete the questionnaire booklet. This contains the PMR-Impact Scale but also two other questionnaires. Please answer them all even if there is some overlap of questions. Please

then put the questionnaire booklet in the pre-paid envelope provided and post it back to the study team.

We would like to send you a second questionnaire booklet to complete about 3 weeks after the first. This will contain the PMR-IS and one other questionnaire as well as some questions about whether your PMR has changed. If you are happy to be sent the second booklet, please sign the consent page on the front of the enclosed booklet before you send it back.

How will the information from this study be used?

The answers to the questions will be analysed to test how the new PMR-Impact Scale compares to the other two questionnaires. We will also test whether the scores on the PMR-Impact Scale stay the same if people feel their illness has stayed the same and whether the scores change as expected when people's PMR improves or worsens.

Will there be any direct benefit from taking part in the study?

There is no direct benefit to your medical care to taking part in this study but you will be helping us to develop a questionnaire which may improve care for others in the future.

Will there be any harm from taking part in the study?

It is very unlikely. If you find any of the questions upsetting you do not have to answer them.

How will we use information about you?

We will need to use information from you for this research project. This information will include your name, contact details, gender, date of birth and questionnaire answers.
People who do not need to know who you are will not be able to see your name or contact details. Your data will have a code number instead.
Once we have finished the study we will keep some of the data so we can check the results. We will write our reports in a way that no-one can work out that you took part in the study.

What are your choice about how your information is used?

You can stop being part of the study at any time without giving a reason but we will keep the information about you that we already have.

Where can you find out more about how your information is used?

You can find out more about how we use your information

- At www.hra.nhs.uk/information-about-patients/
- By asking one of the research team using the contact details below.

Who has reviewed the study?

The South Central-Hampshire research ethics committee have reviewed the study.

Who is doing the research?

This study is being carried out by Dr Helen Twohig, Dr Sara Muller, Dr Caroline Mitchell and Professor Christian Mallen who are all GPs and / or researchers linked to Keele University. It is funded by a Wellcome Trust Doctoral Fellowship awarded to Helen Twohig.

Who can I contact for further information?

If you have any questions or would like more information about this study before deciding whether to take part, please contact the research team using the contact details below.

Who can I contact if I have any complaints about the study?

If you wish to voice concerns or complain about the research in any way please contact Keele University using the contact details below, or your local NHS complaints department.

RESEARCH TEAM CONTACT

Dr Helen Twohig,
Primary Care Centre Versus Arthritis,
School of Primary, Community and Social Care
Keele University, Staffordshire
ST5 5BG
h.j.twohig@keele.ac.uk

INDEPENDENT CONTACT AT KEELE UNIVERSITY

Dr Tracy Nevatte
Directorate of Research, Innovation and Engagement
David Weatherall Building,
Keele University, Staffordshire, ST5 5BG
research.governance@keele.ac.uk

Appendix 9.5: Confirmation of ethical approval

South Central - Hampshire B Research Ethics Committee
Level 3 Block B
Whitefriars
Lewins Mead
Bristol
BS1 2NT

Telephone: 0207 104 8171

Please note: This is the favourable opinion of the REC only and does not allow you to start your study at NHS sites in England until you receive HRA Approval

10 October 2019

Dr Helen Twohig
Wellcome Trust Primary Care Doctoral Fellow
Keele University
Primary Care Centre Versus Arthritis
Keele University
Keele
ST5BG

Dear Dr Twohig

Study title:	**Evaluation of the construct validity, test re-test reliability and responsiveness of the PMR-IS**
REC reference:	**19/SC/0525**
Protocol number:	**RG-0301-19**
IRAS project ID:	**269455**

The Proportionate Review Sub-committee of the South Central - Hampshire B Research Ethics Committee reviewed the above application on 10 October 2019.

Ethical opinion

On behalf of the Committee, the sub-committee gave a favourable ethical opinion of the above research on the basis described in the application form, protocol and supporting documentation, subject to the conditions specified below.

Conditions of the favourable opinion

The REC favourable opinion is subject to the following conditions being met prior to the start of the study.

Confirmation of Capacity and Capability (in England, Northern Ireland and Wales) or NHS management permission (in Scotland) should be sought from all NHS organisations involved in the study in accordance with NHS research governance arrangements. Each NHS organisation must confirm through the signing of agreements and/or other documents that it has given permission for the research to proceed (except where explicitly specified otherwise).

Guidance on applying for HRA and HCRW Approval (England and Wales)/ NHS permission for research is available in the Integrated Research Application System.

For non-NHS sites, site management permission should be obtained in accordance with the procedures of the relevant host organisation.

Sponsors are not required to notify the Committee of management permissions from host organisations.

Registration of Clinical Trials

It is a condition of the REC favourable opinion that **all clinical trials are registered** on a publicly accessible database. For this purpose, 'clinical trials' are defined as the first four project categories in IRAS project filter question 2. Registration is a legal requirement for clinical trials of investigational medicinal products (CTIMPs), except for phase I trials in healthy volunteers (these must still register as a condition of the REC favourable opinion).

Registration should take place as early as possible and within six weeks of recruiting the first research participant at the latest. Failure to register is a breach of these approval conditions, unless a deferral has been agreed by or on behalf of the Research Ethics Committee (see here for more information on requesting a deferral: https://www.hra.nhs.uk/planning-and-improving-research/research-planning/research-registratio n-research-project-identifiers/

As set out in the UK Policy Framework, research sponsors are responsible for making information about research publicly available before it starts e.g. by registering the research project on a publicly accessible register. Further guidance on registration is available at:

https://www.hra.nhs.uk/planning-and-improving-research/research-planning/transparency-respo nsibilities/

You should notify the REC of the registration details. We routinely audit applications for compliance with these conditions.

Publication of Your Research Summary

We will publish your research summary for the above study on the research summaries section of our website, together with your contact details, no earlier than three months from the date of this favourable opinion letter. Should you wish to provide a substitute contact point, make a request to defer, or require further information, please visit:
https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-sum maries/

It is the responsibility of the sponsor to ensure that all the conditions are complied with before the start of the study or its initiation at a particular site (as applicable).

After ethical review: Reporting requirements

The attached document "After ethical review – guidance for researchers" gives detailed guidance on reporting requirements for studies with a favourable opinion, including:

- Notifying substantial amendments
- Adding new sites and investigators
- Notification of serious breaches of the protocol
- Progress and safety reports
- Notifying the end of the study, including early termination of the study
- Final report

The latest guidance on these topics can be found at https://www.hra.nhs.uk/approvals-amendments/managing-your-approval/.

Ethical review of research sites

The favourable opinion applies to all NHS sites taking part in the study, subject to management permission being obtained from the NHS/HSC R&D office prior to the start of the study (see
"Conditions of the favourable opinion").

Approved documents

The documents reviewed and approved were:

Document	*Version*	*Date*
Covering letter on headed paper [Covering letter]	1	20 September 2019
Evidence of Sponsor insurance or indemnity (non NHS Sponsors only) [Sponsor insurance confirmation]	1	31 July 2019
GP/consultant information sheets or letters [Practice invitation letter]	1	20 July 2019
IRAS Application Form [IRAS_Form_23092019]		23 September 2019
IRAS Application Form XML file [IRAS_Form_23092019]		23 September 2019
IRAS Checklist XML [Checklist_23092019]		23 September 2019
Letter from sponsor [Sponsor letter]	1	19 September 2019
Letters of invitation to participant [Participant invite]	2	30 August 2019
Letters of invitation to participant [Covering letter to patients for 2nd questionnaire booklet]	1	01 July 2019
Non-validated questionnaire [Questionnaire booklet 1]	2	30 August 2019
Non-validated questionnaire [Questionnaire booklet 2]	2	30 August 2019
Participant information sheet (PIS) [PIL]	2	30 August 2019
Research protocol or project proposal [Protocol]	2	30 August 2019
Summary CV for Chief Investigator (CI) [CV Helen Twohig]	1	29 July 2019
Summary CV for supervisor (student research) [Supervisor CV]	1	16 July 2019
Validated questionnaire [SF-36 to be part of questionnaire booklet 1]	1	29 July 2019

Membership of the Proportionate Review Sub-Committee

The members of the Sub-Committee who took part in the review are listed on the attached sheet.

Statement of compliance

The Committee is constituted in accordance with the Governance Arrangements for Research Ethics Committees and complies fully with the Standard Operating Procedures for Research Ethics Committees in the UK.

User Feedback

The Health Research Authority is continually striving to provide a high quality service to all applicants and sponsors. You are invited to give your view of the service you have received and the application procedure. If you wish to make your views known please use the feedback form available on the HRA website: http://www.hra.nhs.uk/about-the-hra/governance/quality-assurance/

HRA Learning

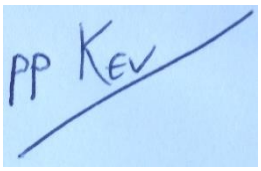
We are pleased to welcome researchers and research staff to our HRA Learning Events and online learning opportunities– see details at:

https://www.hra.nhs.uk/planning-and-improving-research/learning/

With the Committee's best wishes for the success of this project.

19/SC/0525	**Please quote this number on all correspondence**

Yours sincerely



Professor Vincenzo Libri Chair

Email: nrescommittee.southcentral-hampshireb@nhs.net

Enclosures: List of names and professions of members who took part in the review

"After ethical review – guidance for researchers" [SL-AR2]

Copy to: Dr Tracy Nevatte
Dr Helen Twohig, Keele University

South Central - Hampshire B Research Ethics Committee

Attendance at PRS Sub-Committee of the REC meeting on 10 October 2019

Committee Members:

Name	*Profession*	*Present*	*Notes*
Mrs Angela Iveson	Acute Oncology Clinical Nurse Specialist	Yes	
Professor Vincenzo Libri	Medical Consultant in Clinical Pharmacology	Yes	
Miss Rebecca Munro	Student & part time Advice Volunteer	Yes	

Appendix 9.6: The mHAQ

The mHAQ

These questions are about the disability, discomfort and quality of life you have **related to your PMR**.

Please circle one number on each line that best describes your usual activities **over the course of the last week.**

		Without any difficulty	With some difficulty	With much difficulty	Unable to do
a	Dress yourself, including tying shoelaces and doing buttons.....	0	1	2	3
b	Get in and out of bed.....	0	1	2	3
c	Lift a full cup or glass to your mouth.....	0	1	2	3
d	Walk outdoors on flat ground.....	0	1	2	3
e	Wash and dry your entire body.....	0	1	2	3
f	Bend down and pick up clothing from the floor.....	0	1	2	3
g	Turn regular taps on and off.....	0	1	2	3
h	Get in and out of a car.....	0	1	2	3

Appendix 9.7: The RAND SF-36 Questionnaire

RAND HEALTH

RAND > RAND Health > Surveys > RAND Medical Outcomes Study > 36-Item Short Form Survey (SF-36) >

36-Item Short Form Survey Instrument (SF-36)

RAND 36-Item Health Survey 1.0 Questionnaire Items

Choose one option for each questionnaire item.

1. In general, would you say your health is:

- 1 - Excellent
- 2 - Very good
- 3 - Good
- 4 - Fair
- 5 - Poor

2. **Compared to one year ago**, how would you rate your health in general **now**?

- 1 - Much better now than one year ago
- 2 - Somewhat better now than one year ago
- 3 - About the same
- 4 - Somewhat worse now than one year ago
- 5 - Much worse now than one year ago

The following items are about activities you might do during a typical day. Does **your health now limit you** in these activities? If so, how much?

	Yes, limited a lot	Yes, limited a little	No, not limited at all
3. **Vigorous activities**, such as running, lifting heavy objects, participating in strenuous sports	○ 1	○ 2	○ 3
4. **Moderate activities**, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	○ 1	○ 2	○ 3
5. Lifting or carrying groceries	○ 1	○ 2	○ 3
6. Climbing **several** flights of stairs	○ 1	○ 2	○ 3
7. Climbing **one** flight of stairs	○ 1	○ 2	○ 3
8. Bending, kneeling, or stooping	○ 1	○ 2	○ 3
9. Walking **more than a mile**	○ 1	○ 2	○ 3
10. Walking **several blocks**	○ 1	○ 2	○ 3
11. Walking **one block**	○ 1	○ 2	○ 3
12. Bathing or dressing yourself	○ 1	○ 2	○ 3

During the **past 4 weeks**, have you had any of the following problems with your work or other regular daily activities **as a result of your physical health**?

	Yes	No
13. Cut down the **amount of time** you spent on work or other activities	○ 1	○ 2
14. **Accomplished less** than you would like	○ 1	○ 2
15. Were limited in the **kind** of work or other activities	○ 1	○ 2
16. Had **difficulty** performing the work or other activities (for example, it took extra effort)	○ 1	○ 2

During the **past 4 weeks**, have you had any of the following problems with your work or other regular daily activities **as a result of any emotional problems** (such as feeling depressed or anxious)?

	Yes	No
17. Cut down the **amount of time** you spent on work or other activities	○ 1	○ 2
18. **Accomplished less** than you would like	○ 1	○ 2
19. Didn't do work or other activities as **carefully** as usual	○ 1	○ 2

20. During the **past 4 weeks**, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups?

- ○ 1 - Not at all
- ○ 2 - Slightly
- ○ 3 - Moderately
- ○ 4 - Quite a bit
- ○ 5 - Extremely

21. How much **bodily** pain have you had during the **past 4 weeks**?

- ○ 1 - None
- ○ 2 - Very mild
- ○ 3 - Mild
- ○ 4 - Moderate
- ○ 5 - Severe
- ○ 6 - Very severe

---

22. During the **past 4 weeks**, how much did **pain** interfere with your normal work (including both work outside the home and housework)?

- ○ 1 - Not at all
- ○ 2 - A little bit
- ○ 3 - Moderately
- ○ 4 - Quite a bit
- ○ 5 - Extremely

These questions are about how you feel and how things have been with you **during the past 4 weeks**. For each question, please give the one answer that comes closest to the way you have been feeling.

How much of the time during the **past 4 weeks**...

	All of the time	Most of the time	A good bit of the time	Some of the time	A little of the time	None of the time
23. Did you feel full of pep?	1	2	3	4	5	6
24. Have you been a very nervous person?	1	2	3	4	5	6
25. Have you felt so down in the dumps that nothing could cheer you up?	1	2	3	4	5	6
26. Have you felt calm and peaceful?	1	2	3	4	5	6
27. Did you have a lot of energy?	1	2	3	4	5	6
28. Have you felt downhearted and blue?	1	2	3	4	5	6
29. Did you feel worn out?	1	2	3	4	5	6
30. Have you been a happy person?	1	2	3	4	5	6
31. Did you feel tired?	1	2	3	4	5	6

32. During the **past 4 weeks**, how much of the time has **your physical health or emotional problems** interfered with your social activities (like visiting with friends, relatives, etc.)?

- 1 - All of the time
- 2 - Most of the time
- 3 - Some of the time
- 4 - A little of the time
- 5 - None of the time

How TRUE or FALSE is **each** of the following statements for you.

	Definitely true	Mostly true	Don't know	Mostly false	Definitely false
33. I seem to get sick a little easier than other people	1	2	3	4	5
34. I am as healthy as anybody I know	1	2	3	4	5
35. I expect my health to get worse	1	2	3	4	5
36. My health is excellent	1	2	3	4	5

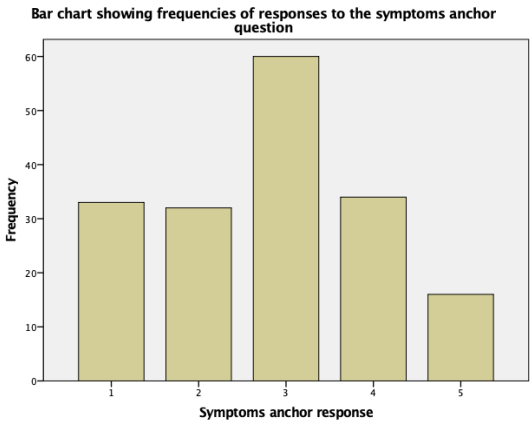
ABOUT

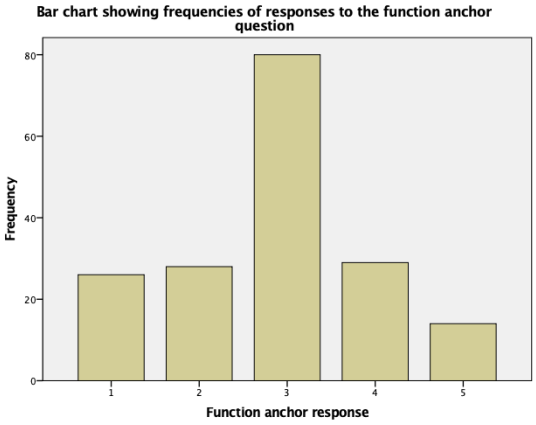
The RAND Corporation is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest.

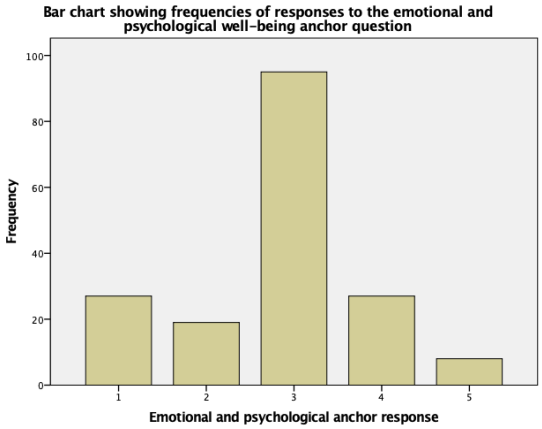
1776 Main Street
Santa Monica, California 90401-3208

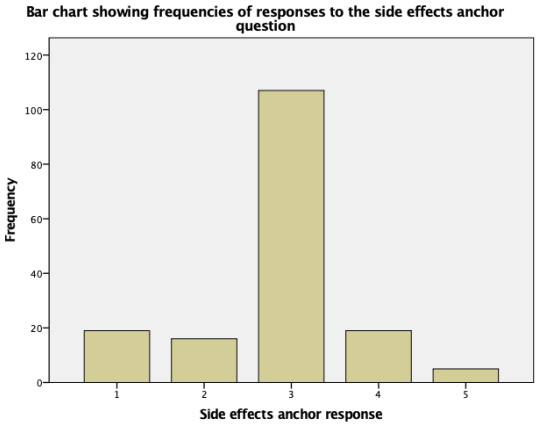
RAND® is a registered trademark. Copyright © 1994-2016 RAND Corporation.

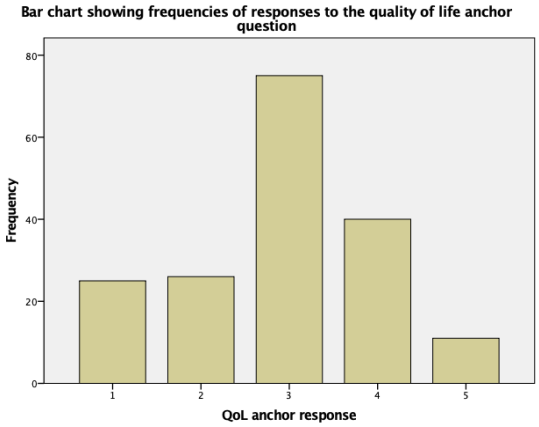
Appendix 9.8: Bar charts of frequencies of responses to the anchor questions







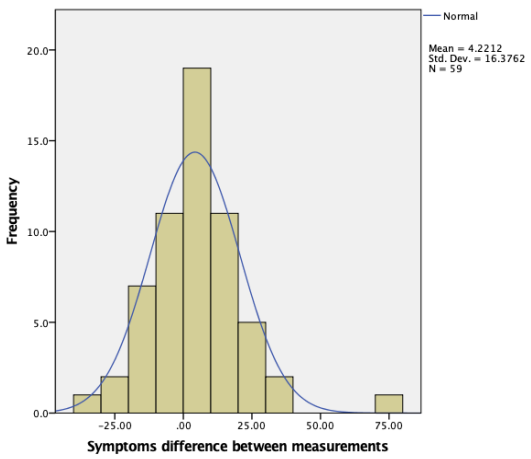


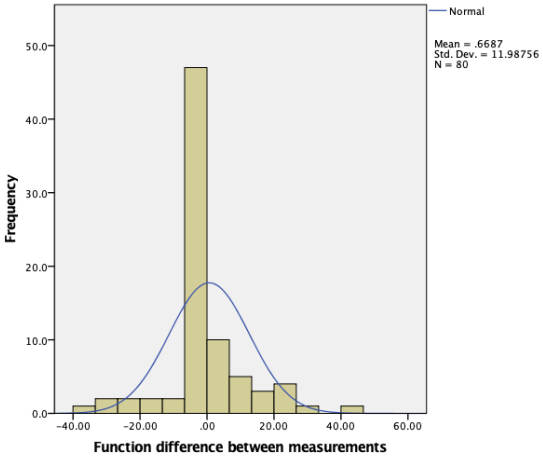


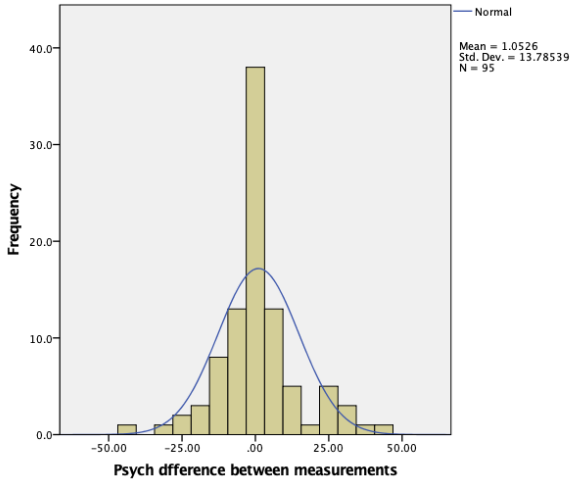
Key:

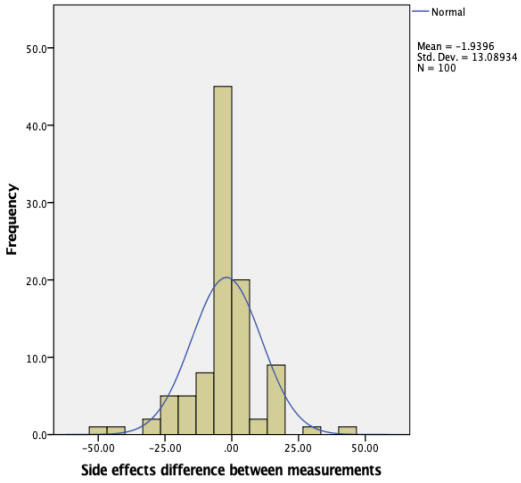
1) Improved a lot
2) Improved a little
3) Stayed the same
4) Worsened a little
5) Worsened a lot

Appendix 9.9: Testing for normality of the differences between the measurements for each scale









Appendix 9.10: Calculating the standard error of the measurement

As described in section 9.3.2, there are several methods for calculating the SEM.

I opted to present the $SEM_{agreement}$ in the results because I wanted to take into account all of the error variance in the sample.

The $SEM_{agreement}$ was calculated from the error variances obtained through a VARCOMP analysis in SPSS. The process for this and the outputs are given below.

For completeness, the $SEM_{consistency}$ is also given in this appendix, calculated by each of the two methods described in section 9.3.2. The results are very similar by all three methods.

VARCOMP analysis

A new database was created in SPSS with time as a stacked variable i.e. a variable called 'index' was created for which values were either 1 or 2 for first and second questionnaire scores for each domain.

Process:

- Select cases where anchor for that domain = 3.
- Analyse > gen linear models > varcomp
- Dependent variable = symptoms score
- Random factors = index and study ID
- Model – custom. Build terms> type is main effects. Select index and study ID.
- Options – method is restricted maximum likelihood.

Output gives estimated variances for:

- Var(StudyID) = variance in the population

- Var(Index1) = systematic error variance i.e. variance due to systematic differences between the two time points
- Var(error) = residual error variance

These variances were then used to calculate $SEM_{agreement}$ using the following formula:

$SEM_{agreement} = \sqrt{}$(systematic error var + residual error var)

Output: symptoms:

Variance Estimates

Component	Estimate
Var(StudyID)	683.741
Var(Index1)	6.576
Var(Error)	133.924

Dependent Variable: Symptoms
Method: Restricted Maximum Likelihood Estimation

Output function:

Variance Estimates

Component	Estimate
Var(StudyID)	399.331
Var(Index1)	.000[a]
Var(Error)	71.176

Dependent Variable: Function
Method: Restricted Maximum Likelihood Estimation
a. This estimate is set to zero because it is redundant.

Output psychological and emotional well being:

Variance Estimates

Component	Estimate
Var(StudyID)	412.497
Var(Index1)	.000[a]
Var(Error)	94.572

Dependent Variable: Psych
Method: Restricted Maximum Likelihood Estimation
a. This estimate is set to zero because it is redundant.

Output side effects:

Variance Estimates

Component	Estimate
Var(StudyID)	427.679
Var(Index1)	.929
Var(Error)	85.781

Dependent Variable: SE
Method: Restricted Maximum Likelihood Estimation

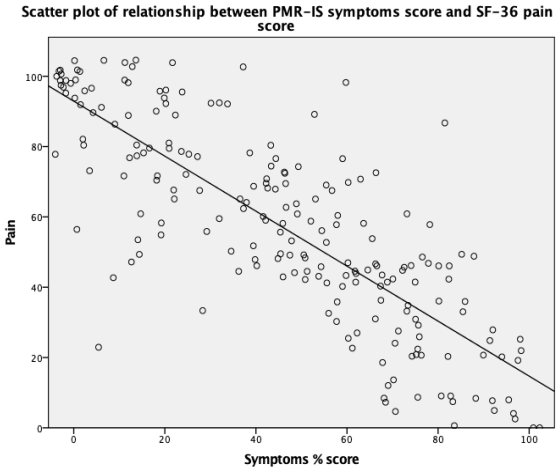
SEM for each domain calculated by the different methods

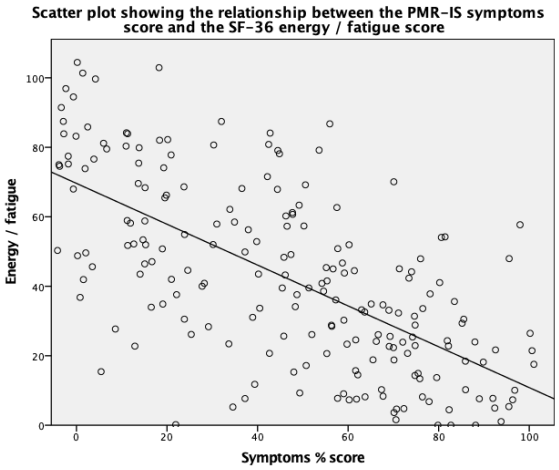
Domain	**n**	**$SEM_{consistency}$ (root mean square average method)**	**$SEM_{consistency}$ ($SD_{difference}$ / √2)**	**$SEM_{agreement}$ (from variance components)**
Symptoms	59	11.86	11.58	11.85
Function	80	8.44	8.48	8.44
Emotional and psychological well-being	95	9.72	9.75	9.72
Steroid side effects	100	9.31	9.26	9.31

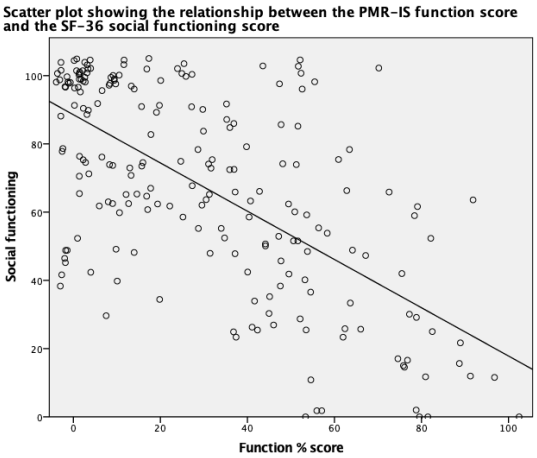
This shows that in my data, $SEM_{agreement}$ is exactly the same as $SEM_{consistency,}$ calculated via the root mean square average method, for all domains except symptoms, where it was only 0.01 different i.e. the systematic error variance is 0 or negligible for all domains.

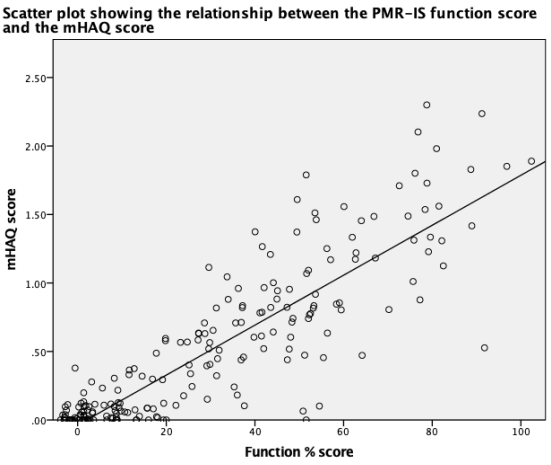
Appendix 9.11: Scatter plots for correlation

Scatter plots to check for a monotonic relationship between the paired variables used in hypothesis testing for construct validity

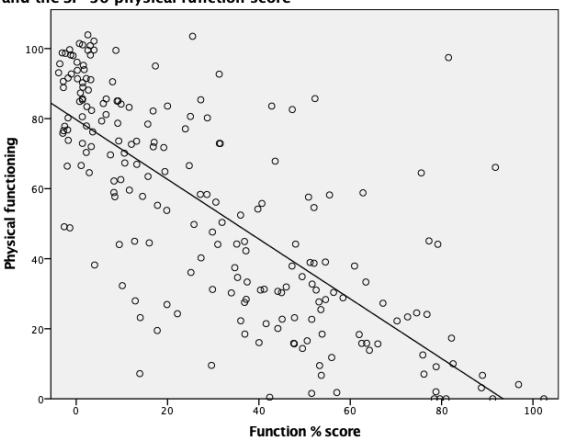




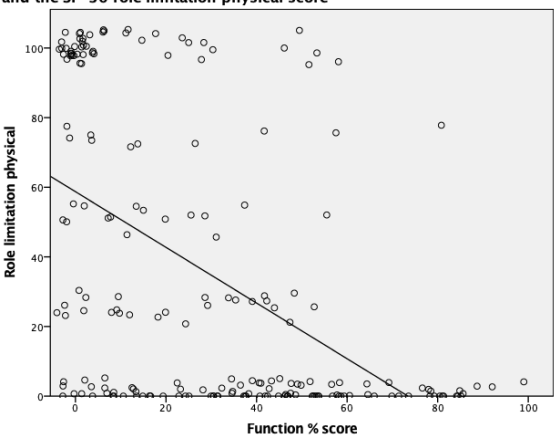




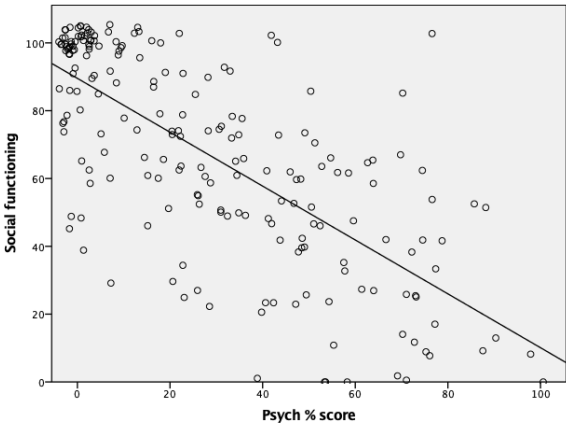
Scatter plot showing the relationship between the PMR-IS function score and the SF-36 physical function score



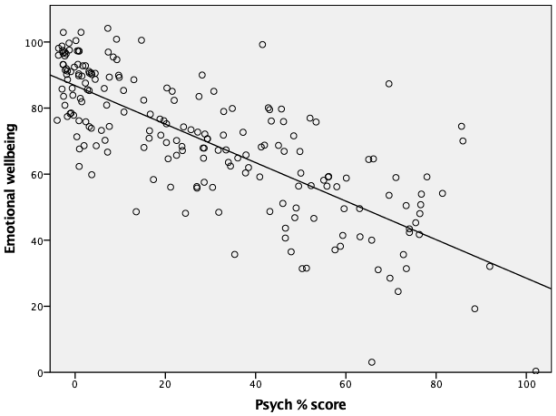
Scatter plot showing the relationship between the PMR-IS function score and the SF-36 role limitation physical score



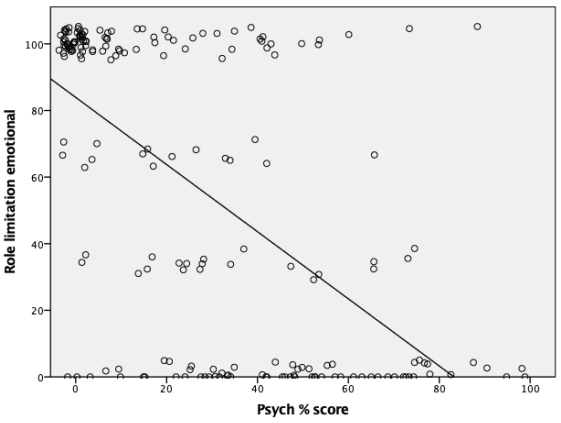
Scatter plot showing the relationship between the PMR-IS emotional and psychological well-being score and the SF-36 social functioning score



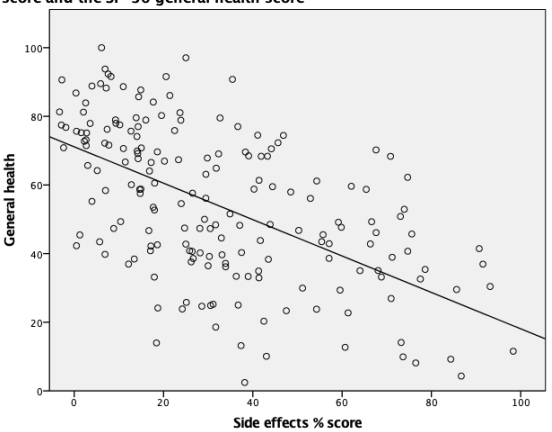
Scatter plot showing the relationship between the PMR-IS emotional and psychological well-being score and the SF-36 emotional well-being score



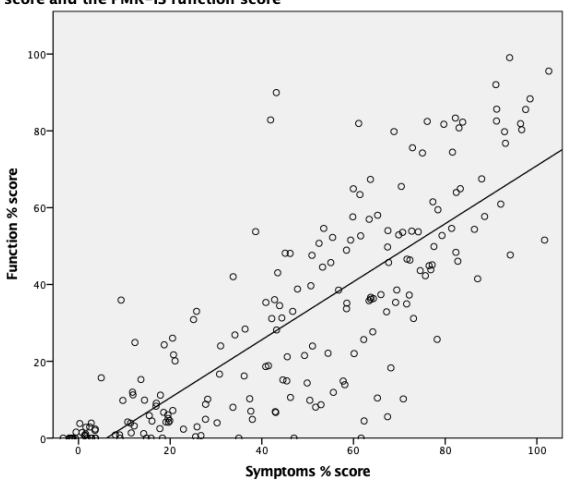
Scatter plot showing the relationship between the PMR-IS emotional and psychological well-being score and the SF-36 role limitation emotional score



Scatter plot showing the relationship between the PMR-IS side effects score and the SF-36 general health score



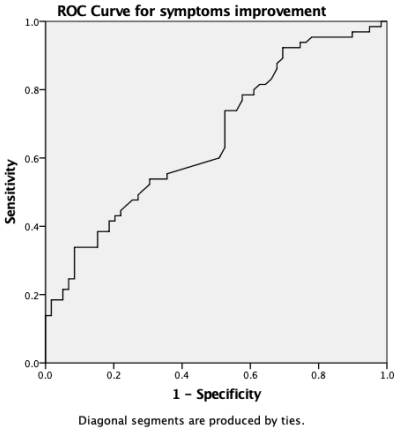
Scatter plot showing the relationship between the PMR-IS symptoms score and the PMR-IS function score

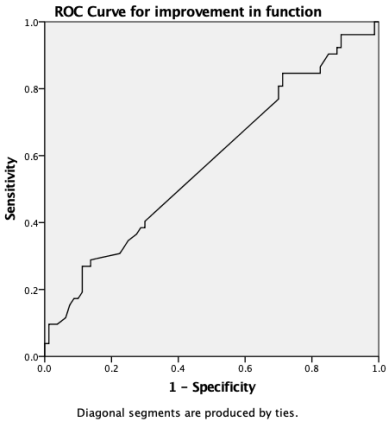


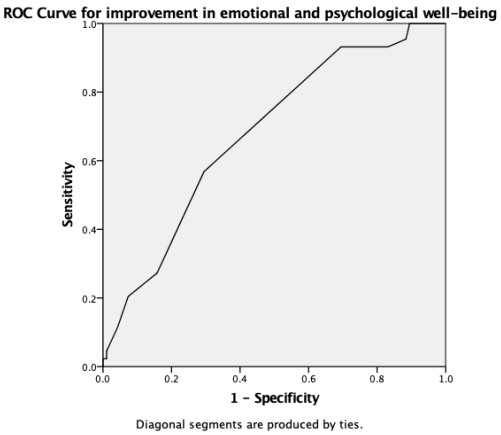
Appendix 9.12: Anchor based ROC method to calculate the MIC improvement for each domain

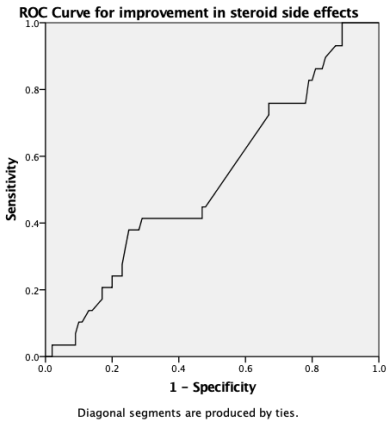
Participants were grouped into those that had remained stable and those that had improved for each domain (those that had worsened were excluded from this analysis).

An ROC curve was plotted in SPSS using the 'change in score' as the test variable and 'improved / stable' as the state variable, with sensitivity (true positives) on the y axis and 1-specicity (false positives) on the x axis.









The coordinate points were imported into Excel and a column for sensitivity minus specificity calculated.

The change score at the point where the sensitivity minus specificity was lowest was taken as the MIC.

Example of table of coordinate points for the symptoms domain

Coordinates of the Curve

Test Result Variable(s): Difference so that improvement is positive

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity	Specificity	Sens-Spec
-73.5	1	1	0	1
-61.5625	1	0.983	0.017	0.983
-42.7083	0.985	0.983	0.017	0.968
-33.3333	0.985	0.966	0.034	0.951
-31.25	0.985	0.949	0.051	0.934
-26.875	0.969	0.949	0.051	0.918
-22.8125	0.969	0.932	0.068	0.901
-21.875	0.969	0.898	0.102	0.867
-20.9375	0.954	0.898	0.102	0.852
-20.3125	0.954	0.881	0.119	0.835
-19.6875	0.954	0.864	0.136	0.818
-19.375	0.954	0.831	0.169	0.785
-18.4375	0.954	0.814	0.186	0.768
-17.1875	0.954	0.797	0.203	0.751
-16.5625	0.954	0.78	0.22	0.734
-15.625	0.938	0.763	0.237	0.701
-14.5833	0.938	0.746	0.254	0.684
-13.3333	0.923	0.746	0.254	0.669
-12.1875	0.923	0.729	0.271	0.652
-11.5625	0.923	0.712	0.288	0.635
-11.25	0.923	0.695	0.305	0.618
-10.9375	0.908	0.695	0.305	0.603
-10.3125	0.892	0.695	0.305	0.587
-9.6875	0.877	0.678	0.322	0.555
-9.0625	0.862	0.678	0.322	0.54
-8.125	0.831	0.661	0.339	0.492
-7.1875	0.815	0.644	0.356	0.459
-6.875	0.815	0.627	0.373	0.442
-6.875	0.8	0.61	0.39	0.41

-6.5625	0.785	0.61	0.39	0.395
-5.9375	0.785	0.593	0.407	0.378
-5.625	0.785	0.576	0.424	0.361
-5	0.769	0.576	0.424	0.345
-4.0625	0.738	0.559	0.441	0.297
-3.4375	0.738	0.525	0.475	0.263
-2.8125	0.723	0.525	0.475	0.248
-2.1875	0.708	0.525	0.475	0.233
-1.875	0.677	0.525	0.475	0.202
-1.5625	0.662	0.525	0.475	0.187
-1.25	0.646	0.525	0.475	0.171
-0.9375	0.631	0.525	0.475	0.156
-0.3125	0.6	0.508	0.492	0.108
0.625	0.554	0.356	0.644	-0.09
1.25	0.538	0.356	0.644	-0.106
1.25	0.538	0.339	0.661	-0.123
1.5625	0.538	0.305	0.695	-0.157
3.125	0.523	0.305	0.695	-0.172
4.6875	0.492	0.271	0.729	-0.237
5	0.477	0.271	0.729	-0.252
5.625	0.477	0.254	0.746	-0.269
6.25	0.446	0.22	0.78	-0.334
6.5625	0.431	0.22	0.78	-0.349
7.1875	0.431	0.203	0.797	-0.366
7.8125	0.415	0.203	0.797	-0.382
8.4375	0.415	0.186	0.814	-0.399
9.0625	0.4	0.186	0.814	-0.414
9.5417	0.385	0.186	0.814	-0.429
10.1667	0.385	0.169	0.831	-0.446
10.625	0.385	0.153	0.847	-0.462
10.625	0.338	0.153	0.847	-0.509
10.65	0.338	0.136	0.864	-0.526
11.1411	0.338	0.119	0.881	-0.543
11.7411	0.338	0.102	0.898	-0.56
12.5446	0.338	0.085	0.915	-0.577
13.9286	0.323	0.085	0.915	-0.592
14.8214	0.308	0.085	0.915	-0.607
15	0.292	0.085	0.915	-0.623
15.3125	0.277	0.085	0.915	-0.638
15.9375	0.262	0.085	0.915	-0.653
16.5625	0.246	0.085	0.915	-0.669
17.1875	0.246	0.068	0.932	-0.686
17.8125	0.231	0.068	0.932	-0.701

18.4375	0.215	0.068	0.932	-0.717
19.0625	0.215	0.051	0.949	-0.734
19.6875	0.2	0.051	0.949	-0.749
20.625	0.185	0.051	0.949	-0.764
21.5625	0.185	0.034	0.966	-0.781
23.125	0.185	0.017	0.983	-0.798
25.9375	0.169	0.017	0.983	-0.814
29.0625	0.154	0.017	0.983	-0.829
32.1875	0.138	0.017	0.983	-0.845
34.0625	0.138	0	1	-0.862
35.3125	0.123	0	1	-0.877
36.5625	0.077	0	1	-0.923
37.1875	0.062	0	1	-0.938
39.0625	0.046	0	1	-0.954
43.4375	0.031	0	1	-0.969
68.125	0.015	0	1	-0.985
91	0	0	1	-1