Seminars in Arthritis and Rheumatism 62 (2023) 152239

Contents lists available at ScienceDirect

Seminars in Arthritis and Rheumatism

journal homepage: www.elsevier.com/locate/semarthrit

Test-retest reliability of pain VAS/NRS, stiffness VAS/NRS, HAQ-DI and mHAQ in polymyalgia rheumatica: An OMERACT study

Jessica L. Leung [a,b,*], Helen Twohig [c], Sara Muller [c], Lara Maxwell [d], Sarah L. Mackie [e,f], Lorna M Neill [g], Claire E. Owen [a,b], for the OMERACT PMR Working Group

[a] Austin Health, 145 Studley Road, Heidelberg, Victoria, Australia
[b] The University of Melbourne, Swanston Street, Parkville, Victoria, Australia
[c] School of Medicine, Keele University, Keele, Staffordshire, ST5 5BG, United Kingdom
[d] Faculty of Medicine, University of Ottawa, 451 Smyth Road, Ottawa, Ontario, Canada
[e] Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds, Woodhouse, Leeds, LS2 9JT, United Kingdom
[f] NIHR Leeds Biomedical Research Unit, Leeds Teaching Hospitals NHS Trust, Beckett Street, Leeds, West Yorkshire, LS9 7TF, United Kingdom
[g] OMERACT Patient Research Partner, PMR-GCA Scotland Unit, Leeds Teaching Hospitals NHS Trust, Leeds, UK

ARTICLE INFO

Keywords:
Polymyalgia rheumatica
Test-retest reliability
Responsiveness
Outcome measures
Visual analogue scale
Numerical rating score
Health Assessment Questionnaire-Disability Index
Modified Health Assessment Questionnaire
OMERACT

ABSTRACT

Objective: To examine the test-retest reliability of four measurement instruments in polymyalgia rheumatica (PMR): pain severity visual analogue scale (VAS) / numerical rating score (NRS), stiffness severity VAS/NRS, the Health Assessment Questionnaire-Disability Index (HAQ-DI) and the modified Health Assessment Questionnaire (mHAQ).

Method: Two prospectively collected datasets were used. All participants had a diagnosis of PMR and only those with stable disease were included in analyses. Measurement instruments were administered twice, with a testing interval of two to six weeks. The intra-class correlation coefficient (ICC) was calculated using a two-way mixed effects model looking for absolute agreement. ICC values of 0.8-0.9 were deemed representative of good test-retest reliability, whilst values >0.9 were representative of excellent test-retest reliability.

Results: From the first dataset, 38 participants were analysed. The ICC between baseline and 2 weeks for pain VAS, stiffness VAS, HAQ-DI and mHAQ were 0.84, 0.82, 0.92 and 0.92 respectively.

From the second dataset, 58 participants were included in the analysis for pain NRS, 59 for stiffness NRS and 78 for mHAQ. The ICC between baseline and follow-up for pain NRS, stiffness NRS and mHAQ were 0.80, 0.83 and 0.87 respectively.

Conclusion: Pain severity VAS/NRS, stiffness severity VAS/NRS, HAQ-DI and mHAQ all demonstrate good to excellent test-retest reliability in a PMR patient population.

Statement of Clinical Significance

Polymyalgia rheumatica (PMR) is a common inflammatory condition occurring in older people, yet remains poorly understood with limited evidenced-based treatment options. Advancement of PMR research is hampered by the absence of validated outcome measures. To date, there has been only one other study to examine the test-retest reliability of any outcome measure in PMR, specifically the pain Visual Analogue Scale (VAS) and the modified Health Assessment Questionnaire (mHAQ). No studies have previously assessed test-retest reliability of stiffness VAS/ Numerical Rating Scale (NRS) or Health Assessment Questionnaire-Disability Index (HAQ-DI) in a PMR population.

Here, we present findings that support the test-retest reliability of pain severity VAS/NRS and mHAQ, and for the first time demonstrate good to excellent test-retest reliability of stiffness severity VAS/NRS and HAQ-DI. This study contributes to work undertaken by the PMR Working Group to establish an OMERACT-endorsed core outcome measurement set.

* Corresponding author at: Rheumatology Department, Austin Health, 145 Studley Road, Heidelberg, Victoria, Australia
E-mail addresses: jessica.leung@austin.org.au (J.L. Leung), h.j.twohig1@keele.ac.uk (H. Twohig), s.muller@keele.ac.uk (S. Muller), lmaxwell@uottawa.ca (L. Maxwell), s.l.mackie@leeds.ac.uk (S.L. Mackie), gdneill@gotadsl.co.uk (L.M. Neill), claire.owen@austin.org.au (C.E. Owen).

https://doi.org/10.1016/j.semarthrit.2023.152239

Available online 5 July 2023

Introduction

Polymyalgia rheumatica (PMR) is a rheumatic disease characterised by chronic inflammation of musculotendinous structures throughout the shoulder and pelvic girdle. Despite a lifetime incidence second only to rheumatoid arthritis [1], there remains a relative paucity of quality PMR research, and further progress is impeded by the lack of validated outcome measures. Selection and validation of outcome measures in PMR is therefore critical for the advancement of research into the condition and ultimately for the improvement of patient care.

Previous work by the Outcome Measures in Rheumatology (OMERACT) PMR working group has led to the endorsement of a core domain set of four disease aspects that should be measured in all PMR clinical trials: pain, stiffness, physical function, and laboratory markers of systemic inflammation [2].

Subsequent work has identified candidate instruments that achieved adequate domain match and feasibility requirements as outlined in OMERACT Filter 2.1 [3–6]. The selected instruments include a visual analogue scale (VAS) or numerical rating score (NRS) for the pain and stiffness domains, and the Health Assessment Questionnaire-Disability Index (HAQ-DI) or modified Health Assessment Questionnaire (mHAQ) for the physical function domain [7,8]. All instruments are completed by the patient.

Measurement properties of these candidate instruments must be examined to ensure validity in a PMR patient population. This includes test-retest reliability, a measure of the consistency of scores when a test is repeated in a stable situation.

A systematic literature review found a lack of high-quality studies specifically measuring the psychometric properties of instruments in a PMR population [9]. Working group members therefore contributed raw datasets to conduct these analyses, and new studies were designed where necessary.

Here, we present work pertaining to the test-retest reliability of four candidate instruments in PMR.

Methods

Datasets

Two datasets contributed to this analysis.

Dataset 1: Melbourne, Australia

Patients diagnosed clinically with PMR by a rheumatologist at least six months prior were recruited prospectively from a tertiary hospital outpatient clinic in Australia. Eligible participants were identified consecutively as they attended for a routine clinic appointment. All participants had stable disease and a treatment change was not planned for at least two weeks after inclusion into the study. A sample size of 50 participants was targeted as recommended by COSMIN [10,11].

Paper questionnaires were completed in a waiting area after their clinic appointment and repeated at home two weeks later, including a pain severity VAS (0-10cm), stiffness severity VAS (0-10cm) and HAQ-DI. The VAS question stems were worded "How would you rate the level of [pain / stiffness] you are currently experiencing from PMR?", with the anchors "No [pain / stiffness]" and "[Pain / Stiffness] as bad as it could be" at either end of the scale. At the follow-up survey, participants were also asked an anchor question to confirm disease stability. Only participants who reported that their condition was "the same as before", "a bit better than before" or "a bit worse than before" were included in this analysis, whilst those who reported that their condition was "much [better / worse] than before" were excluded. The study received ethical approval from the Austin Health Human Research Ethics Committee (reference HREC/57623/Austin-2019) and all participants gave written informed consent for publication prior to recruitment.

Dataset 2: Keele, United Kingdom

This was a nested study within a larger prospective cohort evaluating the psychometric properties of a novel patient-reported outcome measure in PMR, the PMR-Impact Scale [12]. Patients diagnosed with PMR within the prior three years were identified through primary care practices and one secondary care site in the United Kingdom. Participants completed an initial postal questionnaire booklet including pain severity NRS (0-10), stiffness severity NRS (0-10) and mHAQ, and a second questionnaire booklet two to six weeks later, which also included a series of anchor questions to confirm disease stability. The NRS question stems were worded "How bad has the [pain / stiffness] caused by your PMR been during the last week?", with the anchors "No [pain / stiffness]" and "Severe [pain / stiffness]" at either end of the scale. Only participants who reported that their symptoms / function had "stayed the same" on a domain-specific anchor question were included in the test-retest reliability analysis. The study received UK NHS Health Research Authority and Research Ethics Committee approval (REC reference 19/SC/0525) and all participants gave written informed consent for publication prior to recruitment.

Statistical analysis

Within each dataset and for each PRO, the intra-class correlation coefficient (ICC) was calculated using a two-way mixed effects model looking for absolute agreement in scores. ICC values of 0.8-0.9 and >0.9 were considered to represent good and excellent test-retest reliability respectively.

Standard error of the measurement (SEM) was calculated using the formula $SEM = SD_{difference} / \sqrt{2}$. The smallest detectable change (SDC), indicating the minimum change score that would be needed to represent meaningful change, was calculated at both individual and group level ($SDC_{individual} = 1.96 \times \sqrt{2} \times SEM$ and $SDC_{group} = SDC_{individual} / \sqrt{n}$). Bland-Altman plots were graphed.

Table 1
Dataset characteristics

	Dataset 1	*Dataset 2*
Country	Australia	United Kingdom
Sample size	38	58 (for pain NRS) 59 (for stiffness NRS) 78 (for mHAQ)
Age (mean, SD), years	70.6 (8.5)	72.2 (8.1)
Female (%)	61%	57.1%
Disease duration (mean, SD), months	38.4 (48.5)	16.1 (8.9)
Measurement instruments tested	- Pain severity VAS - Stiffness severity VAS - HAQ-DI - mHAQ	- Pain severity NRS - Stiffness severity NRS - mHAQ
Method of confirming disease stability	Participant report that condition was "the same as before", "a bit better than before" or "a bit worse than before"	Participant report that symptom / function had "stayed the same"
Question stem used for VAS / NRS	"How would you rate the level of [pain / stiffness] you are currently experiencing from PMR?"	"How bad has the [pain / stiffness] caused by your PMR been during the last week?"
Lower anchor used for VAS / NRS	"No [pain / stiffness]"	"No [pain / stiffness]"
Upper anchor used for VAS / NRS	"[Pain / Stiffness] as bad as it could be"	"Severe [pain / stiffness]"

Results

Dataset 1: Melbourne, Australia

In the first study, 48 participants were recruited. Of these, 38 participants with confirmed stable disease were included in this analysis (Table 1).

The ICC between baseline and 2 weeks for pain VAS, stiffness VAS, HAQ-DI and mHAQ were 0.84, 0.82, 0.92 and 0.92 respectively (Table 2). SDC_{group} figures were low for all instruments, meaning that small changes in score at the group level can be attributed to true change rather than measurement error. Bland-Altman plots demonstrated an acceptable degree of minor deviation of most datapoints from the line of no difference (Figure 1).

Dataset 2: Keele, UK

In the second study, 210 first booklets and 179 paired booklets were returned. Of these, 58 participants with confirmed stable disease were included in the analysis for pain NRS, 59 for stiffness NRS and 78 for mHAQ (Table 1).

The ICC between baseline and 2 weeks for pain NRS, stiffness NRS and mHAQ were 0.80, 0.83 and 0.87 respectively (Table 2). SDC_{group} figures were low for all instruments. Bland-Altman plots demonstrated an acceptable degree of minor deviation of most datapoints from the line of no difference (Figure 2).

Discussion

Our analysis has found that pain VAS/NRS and stiffness VAS/NRS have good test-retest reliability (ICC >0.80) in a PMR patient population. Test-retest reliability of HAQ-DI and mHAQ were found to be excellent (ICC >0.90) in the first study and good for mHAQ (HAQ-DI not tested) in the second study.

Prior to this work, there has only been a single published study examining test-retest reliability of any instruments in a PMR patient population. Our results are similar to those reported by Matteson et al, who determined an ICC of 0.82 for pain VAS in 14 patients with PMR tested over a one-week period [13]. On the other hand, mHAQ demonstrated stronger test-retest reliability in our study compared to that estimated by Matteson et al, who reported an ICC of 0.72 [13]. This may relate to their inclusion of only newly diagnosed, corticosteroid-naïve patients, who commenced treatment after their initial visit and conceivably may have experienced an improvement in symptoms over the two timepoints. In contrast, participants in our study were not newly diagnosed and our analysis only included participants who rated their disease as stable. Furthermore, all participants in our first dataset continued steady treatment doses throughout the assessment period.

As demonstrated in a recent systematic literature review, no previous study has examined test-retest reliability of stiffness VAS / NRS or HAQ-DI in a PMR population [9]. Our results are therefore novel, providing preliminary confidence to researchers using these instruments of the validity of results in this patient population.

Our study has several strengths. Both datasets were designed a priori to support test-retest analysis and therefore deliberately included stable patients tested under similar conditions over an appropriate time interval. This study therefore fulfills requirements outlined in the COSMIN-OMERACT Good Methods Checklist[6]. Our sample size also makes this the largest test-retest reliability study to have been undertaken in PMR to date, although the number of participants in our first dataset does not meet the minimum of 50 participants recommended by COSMIN [10,11].

We have examined both VAS and NRS, thus providing data on each, but also limiting the amount of data available for each individual instrument. Whilst it has been demonstrated that VAS and NRS are relatively interchangeable in certain situations [14], this has not been tested in a PMR patient population, so data were not combined. Future studies should explore the relative merits of VAS versus NRS in a PMR population.

The study also has some limitations. In the first dataset, the setting in which the instruments were administered differed as the initial questionnaires were completed in clinic whilst the follow-up questionnaires were completed at home. However, the questionnaires were completed in the waiting room with no undue influence by a clinician, therefore we do not expect this difference in setting to have a major impact on results. In addition, the time of day that the questionnaires should be completed was not mandated and may have differed between the two questionnaires. Given the typical diurnal variation of PMR symptoms, it is unclear what effect this may have had on results, although any variability would have been mitigated by the exclusive inclusion of participants who reported disease / symptom stability.

Finally, the VAS / NRS question stem and response anchors used in our study were slightly different in each of our two datasets. There is currently no standardised approach and many studies do not report the wording used. We propose that future work should collaborate with patient research partners to determine the optimal phrasing for these patient-reported outcomes in PMR, and this should remain consistent in future studies.

Conclusion

To conclude, pain VAS/NRS and stiffness VAS/NRS demonstrate good test-retest reliability and HAQ-DI / mHAQ demonstrate excellent test-retest reliability in PMR patients. Crucially, our findings give confidence in the validity of clinical trial results using these PROs in a PMR population, ultimately contributing to the development of PMR therapy.

Table 2
Statistical analysis of test-retest reliability for measurement instruments in PMR

Instrument	Dataset*	Initial score, mean (SD)	Retest score, mean (SD)	$ICC_{agreement}$ (95% CI)	Mean difference (LoA)	SEM	$SDC_{individual}$	SDC_{group}
Domain: Pain								
Pain VAS (0-10cm)	1	2.35 (2.02)	2.58 (1.87)	0.84 (0.69, 0.92)	0.23 (-2.63, 3.09)	1.03	2.86	0.46
Pain NRS (0-10)	2	2.67 (2.84)	3.07 (2.59)	0.80 (0.68, 0.88)	-0.40 (-3.72, 2.92)	1.20	3.33	0.44
Domain: Stiffness								
Stiffness VAS (0-10cm)	1	2.61 (2.46)	2.70 (2.10)	0.82 (0.65, 0.91)	-0.90 (-3.62, 3.44)	1.27	3.53	0.57
Stiffness NRS (0-10)	2	2.83 (2.68)	3.35 (2.83)	0.83 (0.73, 0.91)	-0.51 (-3.51, 2.50)	1.09	3.02	0.39
Domain: Physical Function								
HAQ-DI	1	0.54 (0.50)	0.42 (0.39)	0.92 (0.80, 0.96)	-0.12 (-0.58, 0.34)	0.17	0.46	0.07
mHAQ	1	0.34 (0.40)	0.29 (0.32)	0.92 (0.85, 0.96)	-0.05 (-0.42, 0.32)	0.13	0.37	0.06
mHAQ	2	0.32 (0.45)	0.34 (0.45)	0.87 (0.80, 0.91)	0.02 (-0.43, 0.48)	0.16	0.44	0.05

*Dataset 1 refers to the Melbourne dataset and dataset 2 refers to the Keele dataset.

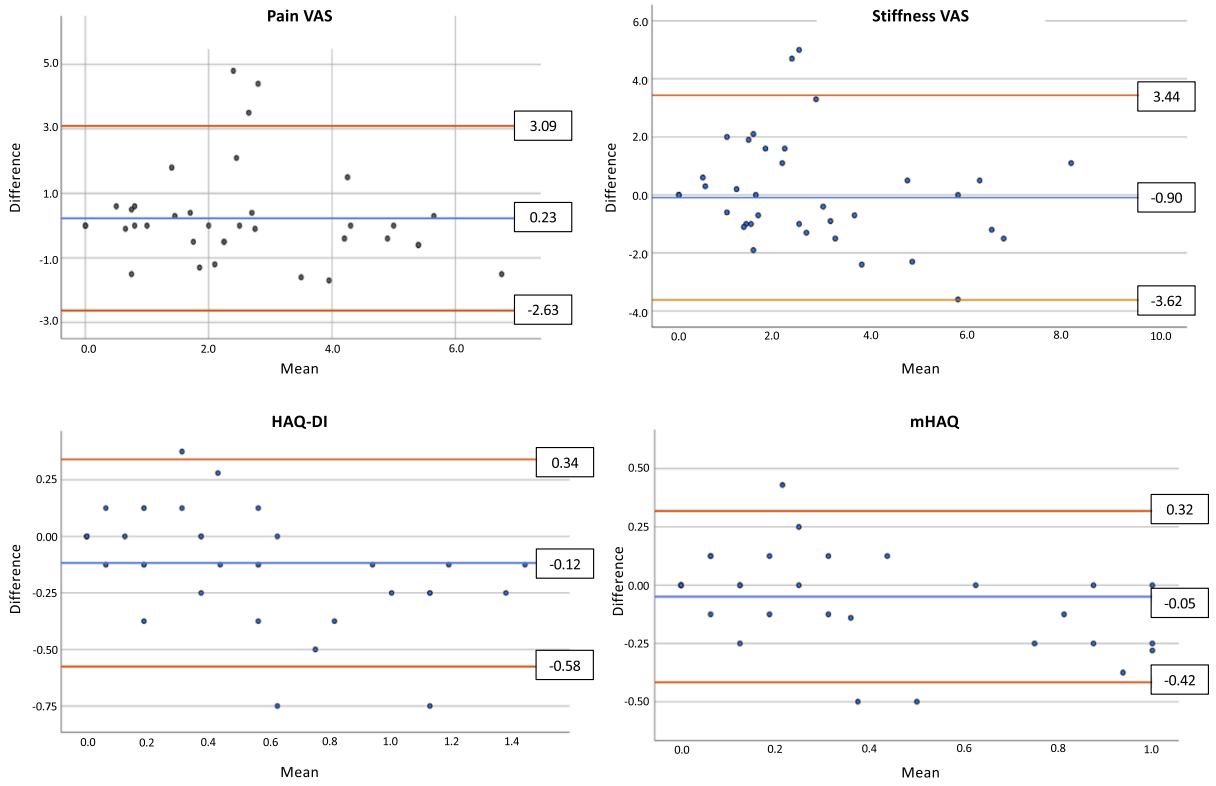

Fig. 1. Bland-Altman Plots (Dataset 1)

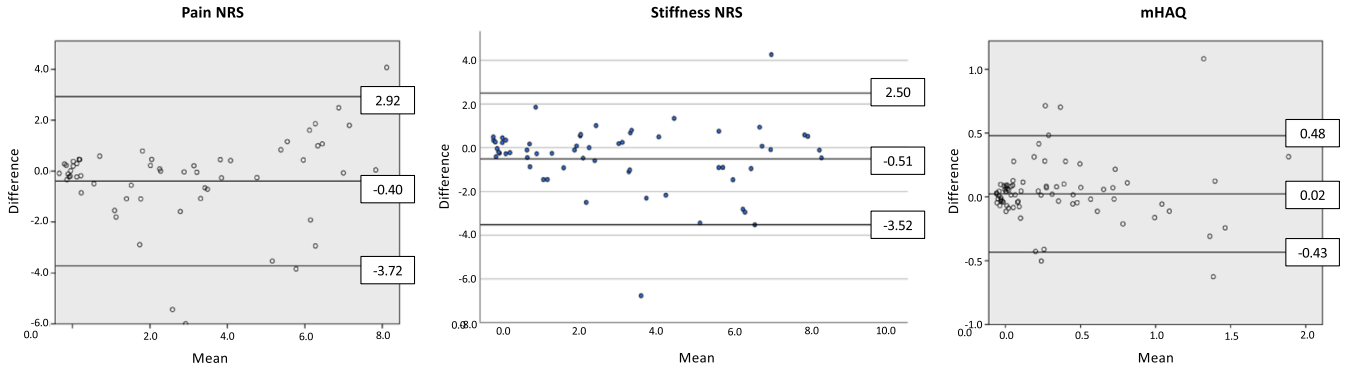

Fig. 2. Bland-Altman Plots (Dataset 2)

Pending evaluation of their other measurement properties, the results of this study support the incorporation of these instruments in a future OMERACT-endorsed core outcome measurement set.

Author Contribution Statement

Jessica L. Leung: Conceptualization, Methodology, Formal Analysis, Investigation, Writing – Original Draft. **Helen Twohig:** Conceptualization, Methodology, Formal Analysis, Investigation, Writing – Review & Editing. **Sara Muller:** Conceptualization, Methodology, Formal Analysis, Writing – Review & Editing. **Lara Maxwell:** Methodology, Writing – Review & Editing. **Sarah L. Mackie:** Conceptualization, Methodology, Writing – Review & Editing. **Lorna M. Neill:** Conceptualization, Writing – Review & Editing. **Claire E. Owen:** Conceptualization, Methodology, Investigation, Writing – Review & Editing.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

H.T. received a salary and research funding from a primary care doctoral fellowship from the Wellcome Trust; data contributed to this manuscript was gathered during this doctoral fellowship. H.T. is currently funded by an NIHR Clinical Lectureship in primary care. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR or the Department of Health and Social Care.

S.M. is funded by the NIHR Applied Research Collaboration West Midlands. This article presents independent research funded by the NIHR and the NIHR Applied Research Collaboration West Midlands. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR or the Department of Health and Social Care.

Declaration of Competing Interest

Jessica L. Leung has received speaker honoraria from Novartis, Eli Lilly and AbbVie, as well as payment for advisory board participation from Fresenius Kabi, Eli Lilly and AbbVie.

Helen Twohig has no declarations of interest.

Sara Muller is a trustee of the PMRGCAuk charity.

Lara Maxwell has no declarations of interest.

Sarah L. Mackie's institution has received grant funding from the National Institute for Health Research, Leeds Hospitals Charity and Vifor Pharmaceuticals, and payment for clinical trial participation from Sanofi. She has received support for virtual attendance at ACR Convergence 2021. She is also the patron of the PMRGCAuk charity (unpaid role) and is supported in part by the NIHR Leeds Bioedical Research Centre.

Lorna M. Neill has received honoraria from AbbVie and is a trustee of PMR-GCA Scotland Charity.

Claire E. Owen has received speaking honoraria from AbbVie and Novartis, as well as payment for advisory board participation from AbbVie.

References

[1] Crowson CS, et al. The lifetime risk of adult-onset rheumatoid arthritis and other inflammatory autoimmune rheumatic diseases. Arthritis Rheum 2011;63(3):633–9.
[2] Mackie SL, et al. The OMERACT Core Domain Set for Outcome Measures for Clinical Trials in Polymyalgia Rheumatica. J Rheumatol 2017;44(10):1515–21.
[3] Yates M, et al. Feasibility and Face Validity of Outcome Measures for Use in Future Studies of Polymyalgia Rheumatica: An OMERACT Study. J Rheumatol 2020;47(9):1379–84.
[4] Owen CE, et al. Toward a Core Outcome Measurement Set for Polymyalgia Rheumatica: Report from the OMERACT 2018 Special Interest Group. J Rheumatol 2019;46(10):1360–4.
[5] Maxwell LJ, et al. The evolution of instrument selection for inclusion in core outcome sets at OMERACT: Filter 2.2. Semin Arthritis Rheum 2021;51(6):1320–30.
[6] Beaton, D., et al., Chapter 5: Instrument selection for Core Outcome Measurement Sets. The OMERACT Handbook, version 2.1 Updated June 2nd 2021. 2021.
[7] Pincus T, et al. Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment Questionnaire. Arthritis Rheum 1983;26(11):1346–53.
[8] Fries JF, et al. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23(2):137–45.
[9] Twohig H, et al. Outcomes Measured in Polymyalgia Rheumatica and Measurement Properties of Instruments Considered for the OMERACT Core Outcome Set: A Systematic Review. J Rheumatol 2021;48(6):883–93.
[10] Terwee CB, et al. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2012;21(4):651–7.
[11] Mokkink LB, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010;19(4):539–49.
[12] Twohig H, et al. Development and psychometric evaluation of the PMR-Impact Scale: a new patient reported outcome measure for polymyalgia rheumatica. Rheumatology (Oxford) 2022.
[13] Matteson EL, et al. Patient-reported outcomes in polymyalgia rheumatica. J Rheumatol 2012;39(4):795–803.
[14] Shafshak TS, Elnemr R. The Visual Analogue Scale Versus Numerical Rating Scale in Measuring Pain Severity and Predicting Disability in Low Back Pain. J Clin Rheumatol 2021;27(7):282–5.