Published in final edited form as:

J Rheumatol. 2021 October ; 48(10): 1547–1551. doi:10.3899/jrheum.210175.

Test-retest reliability for HAQ-DI and SF-36 PF for the measurement of physical function in psoriatic arthritis.

Ying Ying Leung[1,2], **William Tillett**[3], **Pil Hojgaard**[4,5], **Ana-Maria Orbai**[6], **Richard Holland**[7], **Ashish J Mathew**[8,9,10], **Niti Goel**[11], **Jeffrey Chau**[12], **Chris Lindsay**[13], **Alexis Ogdie**[14], **Laura C Coates**[15], **Robin Christensen**[16], **Philip Mease**[17], **Vibeke Strand**[18], **Dafna D Gladman**[19]

[1.]Singapore General Hospital, Duke-NUS Medical School, Singapore

[2.]Duke-NUS Medical School, Singapore

[3.]Royal National Hospital for Rheumatic Diseases, University of Bath, Bath, United Kingdom

[4.]Department of Rheumatology, Holbaek Hospital, Danmark

[5.]Section for Biostatistics and Evidence-Based Research, the Parker Institute, Bispebjerg and Frederiksberg Hospital, Copenhagen

[6.]Division of Rheumatology, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

[7.]Concord Repatriation General Hospital, Sydney, Australia

[8.]The Copenhagen Center for Arthritis Research, Centre for Rheumatology and Spine Disorders, Rigshospitalet Glostrup, University of Copenhagen, Denmark

[9.]Centre for Prognosis Studies in Rheumatic Diseases, Division of Rheumatology, Department of Medicine, University of Toronto, Toronto, Ontario Canada

[10.]Department of Clinical Immunology & Rheumatology, Christian Medical College, Vellore, India

[11.]Patient Research Partner, Adjunct Assistant Professor, Duke University School of Medicine, Durham, North Carolina, USA

[12.]Patient Research Partner, Hong Kong

[13.]Patient Research Partner, Employed by Aurinia Pharma US Inc., Prosper, Texas USA

[14.]Medicine and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[15.]National Institute for Health Research Clinician Scientist, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom

[16.]Research Unit of Rheumatology, Department of Clinical Research, University of Southern Denmark, Odense University Hospital, Denmark.

Correspondence to: Ying-Ying Leung, MD; Department of Rheumatology and Immunology, Singapore General Hospital, The Academia, level 4, 20 College Road, Singapore 169856, Contact No.: +65 63265276, Fax no.: +65 62203321, katyccc@hotmail.com.

[17]Rheumatology Research, Swedish Medical Center and University of Washington School of Medicine, Seattle, Washington, USA

[18]Division of Immunology/Rheumatology, Stanford University School of Medicine, Palo Alto, California, USA

[19]Medicine, University of Toronto, Senior Scientist, Krembil Research Institute, Director, Psoriatic Arthritis Program, University Health Network, Toronto Western Hospital, Toronto, Ontario, Canada

Abstract

Objective.—Due to no existing data, we aimed to derive evidence to support test-retest reliability for the Health Assessment Questionnaire-Disability Index (HAQ-DI) and Medical Outcome Survey Short-Form-36 item physical functioning domain (SF-36 PF) in psoriatic arthritis (PsA).

Methods.—We identified datasets that collected relevant data for test-retest reliability for HAQ-DI and SF-36 PF; and evaluated them using OMERACT Filter 2.1 methodology. We calculated intra-class correlation coefficients (ICC) as a measure of test-retest reliability. We then conducted a quality assessment and evaluated the adequacy of test-retest reliability performance.

Results.—Two datasets were identified for HAQ-DI and one for SF-36 PF in PsA. The quality of the datasets was good. The ICCs for HAQ-DI were excellent in both datasets: 0.94 (95% CI: 0.88 to 0.97) and 0.94 (95% CI: 0.89 to 0.97). The ICC of SF-36 PF was good (0.89, 95% CI: 0.76 to 0.95). The performance of test-retest reliability for both instruments was judged to be adequate.

Conclusion.—The new data derived support good and reasonable test-retest reliability for HAQ-DI and SF-36 PF in PsA.

Keywords

test-retest reliability; physical function; patient reported outcome; psoriatic arthritis

Introduction

Reliability is a basic and essential measurement property for an instrument to be an accurate representation of the participant's performance rather than due to contextual factors of the testing session such as e.g., environmental, psychological or methodological processes. Test-retest reliability is one of the seven measurement properties to be evaluated under the Outcome Measure in Rheumatology (OMERACT) Filter 2.1 (1).

Physical function is one of the core domains to be measured in every randomized controlled trial and longitudinal study in PsA (2). A Group for Research and Assessment in Psoriasis and Psoriatic Arthritis (GRAPPA) working group was convened to create a standardized core outcome measurement set for PsA to address key outcomes, including physical function (3). The Health Assessment Questionnaire – Disability Index (HAQ-DI) and the physical function domain of the Medical Outcome Survey Short-Form – 36 items (SF-36 PF) have been evaluated by OMERACT Filter 2.1 and received provisional endorsement from the OMERACT and GRAPPA community. Throughout this process, we conducted a systematic literature search to identify all articles that evaluated measurement properties of all patient-

reported outcome measures (PROMs) for PsA (4). No information was available for test-retest reliability for PROMs of physical function, including HAQ-DI and SF-36 PF. To address this gap, the working group members were contacted to identify dataset(s) that had collected data for test-retest reliability for these PROMs. In this article, we report the process to derive evidence to support test-retest reliability for HAQ-DI and SF-36 PF in PsA.

Materials and Methods

Two datasets were identified that had collected the possible information for test-retest reliability at the level required to fulfill OMERACT Filter 2.1 requirements. One dataset evaluated test-retest reliability for both HAQ-DI and SF-36 PF, while the other only had data for HAQ-DI.

The first dataset was derived from a multi-centre study in the United Kingdom which aimed to test modifications of various composite measures in 140 PsA patients classified by the Classification Criteria for Psoriatic Arthritis (CASPAR). 31 patients with stable disease and not requiring medication change were reassessed in clinic 1 week after the initial assessment at baseline, and test-retest reliability data was collected for HAQ-DI and SF-36 PF. All questionnaires were administered in paper and pencil format. Stability of the PsA between this short 1-week time points was assumed, given PsA is a chronic disease. Prior to participation, all patients signed informed consents. Ethical approval for this study was given by the North East York Research Ethics Committee (Ref: 17/NE/0084). All patients signed written consent in accordance with the declaration of Helsinki.

The second dataset was from a study conducted at a single centre in Singapore aimed to evaluate the validity of the Singapore version of the PsA Quality of Life Index. The HAQ-DI was used as comparator instrument (5). Out of the 98 recruited PsA patients who fulfilled the CASPAR, 38 patients who did not require medication change had test-retest reliability data for HAQ-DI. Data were collected two weeks apart in the same environment with specific instructions given to patients (e.g., if the first administration was at home, the second was administered at home). Both sets of questionnaires were administered in paper and pencil format and mailed back to the study team in stamped return envelopes provided. Stability of condition between this short 2-week time points was assumed, given that PsA is a chronic disease and there was no change of medication in these patients. The study protocol was read and approved by the SingHealth Centralized Review Board E (Ref: 2012/696/E). Prior to participation, all patients signed informed consents.

The quality of each dataset was evaluated by at least two independent working group members using the OMERACT Good Method Checklist (1), and disputes were reconciled. The OMERACT Good Method Checklist assessed five questions for test-retest reliability: 1) were the patients stable in the interim time period?; 2) was the time interval appropriate?; 3) were the test conditions similar for the measurements?; 4) was the correct statistic used?; and 5) otherwise good methods?; all answerable with ‘Yes, good methods’ or ‘No, not achieved’. A rating for quality was given as Green (Yes, likely low risk of bias), Amber (Some cautions, but can be used as evidence), or Red (No, do not use this evidence). The dataset would not be evaluated further if rated as Red.

Within each dataset, the intraclass correlation coefficient (ICC) and Spearman's rank correlations (rho) between scores in test and retest time points were calculated for the HAQ-DI and SF36 PF. Bland-Altman plots were generated. Additionally, the minimal detectable change (MDC) was calculated as Standard Error of Measurement × 1.96 × √2 (6). The MDC indicates the minimal amount of change that can be interpreted as a real change.

The adequacy of each instrument for test-retest reliability was presented in data extraction tables, and the adequacy of measuring test-retest reliability for each instrument in each dataset was evaluated as (+) adequate performance, (+/−) equivocal, and (−) poor or less than adequate performance. Intraclass correlation coefficients (ICCs) >0.90 and >0.75 were considered excellent and good, respectively. Summarizing the number of datasets with acceptable quality, the adequacy of measuring test-retest reliability, and the consistency of the data, an overall rating for test-retest reliability was synthesized as recommended by OMERACT (7). An overall rating of GREEN, AMBER or RED was given indicating good to go, caution, or stop, respectively.

Results

From the first study, 31 patients (77% men) had available data for test-retest reliability. The mean (standard deviation, SD) age and duration of illness of these 31 patients were 54 (11.0) years and 5.7 (4.7) years. The quality of this dataset was determined by two working group members (YYL and WT) independently and rated as Green for both HAQ-DI and SF-36 PF (Table 1).

The mean (SD) HAQ-DI at baseline and 1 week were 0.54 (0.62) and 0.52 (0.69), respectively with mean difference (SD) of −0.02 (0.30), p=0.77, 95% confidence interval (CI) −0.13 to 0.09, which is lower than the MDC (0.54). The ICC of HAQ-DI between baseline and 1 week was good (0.90, 95% CI: 0.79 to 0.95), and Spearman's rho was 0.94 (p<0.01). Bland-Altman plots showed reasonably minimal dispersion around the line of no difference between baseline and 1-week scores (Supplementary Figure 1A). The working group judged the adequacy of measurement as (+), adequate (Table 2).

The mean (SD) SF-36 PF at baseline and 1 week was 63.6 (29.6) and 66.7 (29.2), respectively, with a mean difference (SD) of 3.10 (8.06), p=0.05, (95% CI: −0.04 to 6.17), which is lower than the MDC (8.37). The ICC of SF-36 PF was excellent (0.96, 95% CI: 0.92 to 0.98), and Spearman's rho was 0.95 (p<0.01). Bland-Altman plots showed minimal dispersion around the line of no difference between baseline and 1-week scores (Supplementary Figure 2). The quality of this dataset was evaluated by 2 working group members (YYL and PH) and rated as Green (Table 1). The working group judged the adequacy of measurement as (+), adequate (Table 2).

From the second study, 38 patients (44.7% men, mean [SD] age 53.9 [11.5] years) had data for test-retest reliability for HAQ-DI. The quality of this dataset was assessed by 2 working group members (YYL and WT) and rated as Green (Table 1).

The mean (SD) HAQ-DI was 0.38 (0.55) and 0.35 (0.56) at baseline and the 2-week time point respectively, with a mean (SD) difference of 0 (0.19) between time points (p=1.00),

which is lower than the MDC (0.36). The ICC for HAQ-DI was excellent (0.94, 95% CI: 0.89 to 0.97), and Spearman's rho was 0.83 ($p<0.00$). Bland-Altman plot data showed minimal dispersion around the line of no difference between baseline and 2-week scores (Supplementary Figure 1B). The adequacy of HAQ-DI in measuring test-retest reliability in this dataset was rated as (+), adequate (Table 2).

Evidence synthesis

With 2 datasets of good quality and adequate performance, test-retest reliability for HAQ-DI received an overall rating of GREEN. For SF-36 PF with only one dataset of good quality and adequate performance, test-retest reliability was rated as AMBER (Table 3).

Discussion

This article reports the evidence synthesized to support test-retest reliability for HAQ-DI and SF-36 PF in PsA. Using the OMERACT methodology, test-retest reliability for HAQ-DI was rated as GREEN (Good to go) and that for SF-36 PF was AMBER (Some caution). To achieve GREEN, at least one additional good quality dataset showing adequacy of performance for the SF-36 PF is required.

Valid and reliable outcome measurement is essential to understand the impact of diseases in daily clinical practice and interpretation of trials (6). Test-retest reliability is an important measurement property of instrument discrimination. It requires the scores of an instrument to remain the same when the target concept has not changed during a period of time. Both HAQ-DI and SF-36 are generic instruments, and test-retest reliability has been evaluated extensively for the general population and rheumatologic diseases (8–10). However, data for PsA were lacking (4). Data derived for other diseases may not be directly extrapolated to PsA unless the measurement property of the instrument has been carefully tested based on pre-specified hypothesis for test-retest reliability. The current report therefore bridges the gap in providing test-retest reliability data for these instruments.

In conclusion, we have demonstrated test-retest reliability for HAQ-DI and SF-36 PF in PsA, which was judged to be good and reasonable, respectively.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Sources of support

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

Conflict of interest

YYL is funded by the Clinician Scientist award of the National Medical Research Council, Singapore (NMRC/CSA-INV/0022/2017). AMO is funded by the Jerome L. Greene Foundation Scholar Award, the Staurulakis Family Discovery Award, the Rheumatology Research Foundation, and the National Institutes of Health (NIH) through the Rheumatic Diseases Resource-based Core Center (P30-AR053503 Cores A and D, and P30-AR070254, Cores A and B). PH and RC (The Parker Institute, Bispebjerg and Frederiksberg Hospital) is supported by a core grant from the Oak Foundation (OCAY-18-774-OFIL). LCC is funded by a National Institute for Health Research Clinician

Scientist award, and the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). AO is funded by NIH/NIAMS R01 AR072363. WT is supported by the National Institute for Health Research, Programme Grants for Applied Research [Early detection to improve outcome in patients with undiagnosed PsA ('PROMPT'), RP-PG-1212-20007].

All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the funding agencies.

References

1. Beaton DE, Maxwell LJ, Shea BJ, Wells GA, Boers M, Grosskleg S, et al. Instrument selection using the omeract filter 2.1: The omeract methodology. J Rheumatol2019;46:1028–35. [PubMed: 30709952]
2. Orbai AM, de Wit M, Mease P, Shea JA, Gossec L, Leung YY, et al. International patient and physician consensus on a psoriatic arthritis core outcome set for clinical trials. Annals of the rheumatic diseases2017;76:673–80. [PubMed: 27613807]
3. Leung YY, Tillett W, Orbai AM, Ogdie A, Eder L, Coates LC, et al. The grappa-omeract working group: 4 prioritized domains for completing the core outcome measurement set for psoriatic arthritis 2019 updates. J Rheumatol Suppl2020;96:46–9. [PubMed: 32482768]
4. Hojgaard P, Klokker L, Orbai AM, Holmsted K, Bartels EM, Leung YY, et al. A systematic review of measurement properties of patient reported outcome measures in psoriatic arthritis: A grappa-omeract initiative. Semin Arthritis Rheum2018;47:654–65. [PubMed: 29037523]
5. Leung YY, Thumboo J, Rouse M, McKenna SP. Adaptation of chinese and english versions of the psoriatic arthritis quality of life (psaqol) scale for use in singapore. BMC musculoskeletal disorders2016;17:432. [PubMed: 27756259]
6. Beaton DE, Boers M, Tugwell P. Chapter 33 - assessment of health outcomes. In: Firestein GS, Budd RC, Gabriel SE, McInnes IB, O'Dell JR, editors. Kelley and firestein's textbook of rheumatology (tenth edition): Elsevier; 2017. p. 496–508.
7. Boers M, Kirwan JR, Tugwell P, Beaton D, Bingham CO III, Conaghan PG, et al.The omeract handbook. 2017.
8. Maska L, Anderson J, Michaud K. 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 questionnaire ii (haq-ii), improved health assessment questionnaire (improved haq), and rheumatoid arthritis quality of life (raqol). Arthritis Care Res (Hoboken)2011;63Suppl 11:S4–13. [PubMed: 22588760]
9. Koh ET, Leong KP, Tsou IY, Lim VH, Pong LY, Chong SY, et al. The reliability, validity and sensitivity to change of the chinese version of sf-36 in oriental patients with rheumatoid arthritis. Rheumatology (Oxford)2006;45:1023–8. [PubMed: 16495318]
10. Kaya S, Sain Guven G, Teles M, Korku C, Aydan S, Kar A, et al. Validity and reliability of the turkish version of the readiness for hospital discharge scale/short form. J Nurs Manag2018;26:295–301. [PubMed: 29156508]

Table 1.

Quality Assessment of Databases for Test-Retest Reliability Using the OMERACT Good Method Checklist

OMERACT Good Method Checklist for test-retest reliability	***Study 1*** **Tillet 2019 Rating (remarks)**	***Study 2*** **Leung 2016 Rating (remarks)**
HAQ-DI		
Were the patients stable in the interim time period?	Yes, good methods (No change in condition is expected within 1-week interval for patients with stable PsA without medication change)	Yes, good methods (No change in condition is expected within 2-week interval for patients with PsA who did not require medication change)
Was the time interval appropriate?	Yes, good methods (The 1-week interval is appropriate for patients with stable PsA, no change in condition is expected)	Yes, good methods (The 2-week interval is appropriate for patients with stable PsA, no change in condition is expected)
Were the test conditions similar for the measurements? (e.g., type of administration, environment, instructions)	Yes, good methods (paper and pencil format, same environment, same instructions at both time points)	Yes, good methods (Paper and pencil format, patients given specific instructions to administer in the same setting and environment at both time points)
Was the correct statistic used? • Continuous data: intra-class correlation coefficient (ICC) Dichotomous/ordinal/nominal scores: Kappa used	Yes, good methods (ICC. Spearman's rank correlation, and Bland-Altman plot)	Yes, good methods (ICC. Spearman's rank correlation, and Bland-Altman plot)
Otherwise good methods? (Free of any other important flaws).	Yes, good methods (No severe flaws identified)	Yes, good methods (No severe flaws identified)
Overall quality	**Green (Yes, likely low risk of bias)**	**Green (Yes, likely low risk of bias)**
SF-36 PF		
	Study 1 **Tillet 2019 Rating (remarks)**	***Study 2*** **Leung 2016 Rating (remarks)**
Were the patients stable in the interim time period?	Yes, good methods (No change in condition is expected within 1-week interval for patients with stable PsA without medication change)	NA
Was the time interval appropriate?	Yes, good methods (The 1-week interval for patients with stable PsA, no change in condition is expected)	NA
Were the test conditions similar for the measurements? (e.g., type of administration, environment, instructions)	Yes, good methods (Paper and pencil format, same environment, same instructions at both time points)	NA
Was the correct statistic used? • Continuous data: intra-class correlation coefficient (ICC) Dichotomous/ordinal/nominal scores: Kappa used.	Yes, good methods (ICC, Spearman's rank correlation, and Bland-Altman plot)	NA
Otherwise good methods? (Free of any other important flaws).	Yes, good methods (no flaws identified)	NA
Overall quality	**Green (Yes, likely low risk of bias)**	NA

HAQ-DI: Health Assessment Questionnaire-Disability Index; ICC: intraclass correlation coefficient; NA: no data available; OMERACT: Outcome Measures in Rheumatology; PsA: psoriatic arthritis; SF-36 PF: Medical Outcomes Survey Short-Form 36 item physical functioning domain.

J Rheumatol. Author manuscript; available in PMC 2022 October 01.

Table 2.

Report of Studies of Test-Retest Reliability for HAQ-DI and SF-36 PF in PsA with OMERACT Filter 2.1

Author, year	Study description		Results					Judgement
	Characteristics of sample	Characteristics of testing situation	Sample recruited and sample considered stable for analysis	Scores at baseline and retest	Statistic used	Results	Minimal detectable change (95%CI) $SEM=SD_{baseline} \times \sqrt{(1-ICC)}$ $MDC=1.96 \times SEM \times \sqrt{2}$	Interpretation of adequacy (+, +/−, −)
HAQ-DI								
Study 1 Tillett 2019	140 consecutive patients with PsA fulfilled CASPAR, recruited for validation of composite measures	• 1 week apart • Assumed no change in condition	• 31 patients (77% men) who required no medication change had data for test-retest reliability • Mean (SD) age 54 (11) years • Mean (SD) duration of PsA 5.7 (4.7) years	Mean (SD) T1: 0.54 (0.62) T2: 0.52 (0.69) Mean (SD) difference = −0.02 (0.30), p=0.77 95% CI: −0.13 to 0.09	• ICC • Spearman's rho (r)	• ICC=0.90 (95% CI: 0.79–0.95) • r=0.94 (p<0.01)	SEM=0.62 × √(1–0.90) =0.20 MDC=1.96 × 0.20 × √2 = 0.54	(+) Good ICC and correlation between scores in a situation where changes in scores were not expected. Bland-Altman plot provided supportive evidence.
Study 2 Leung 2016	• 98 consecutive patients with PsA fulfilled CASPAR recruited for validation of PsAQoL study • Mean (SD) duration of illness 5.5 (8.4) years	• 2 weeks apart • Stability between timepoints assumed, given that PsA is a chronic illness and no change in medicine between T1 and T2	• 38 patients (44.7% men) who required no medication change had data for test-retest reliability • Mean (SD) age 53.9 (11.5) years	Mean (SD) T1: 0.38 (0.55) T2: 0.35 (0.56) Mean difference=0 (SD=0.19), p=1.00 95% CI: −0.373 to 0.373	• ICC • Spearman's rho (r)	• ICC=0.94 (95% CI: 0.89–0.97) • r=0.83 (p<0.001)	SEM=0.55 × √(1–0.94) =0.13 MDC=1.96 × 0.13 × √2 = 0.36	(+) Excellent ICC and correlation between scores in a situation where changes in scores were not expected. Bland-Altman plot provided supportive evidence.
SF-36 PF								
Study 1 Tillett 2019	140 consecutive patients with PsA fulfilled CASPAR, recruited for validation of composite measures	• 1 week apart • Assumed no change in condition	• 31 patients who required no medication change had data for test-retest reliability • Mean (SD) age 54 (11) years • Mean (SD) duration of PsA 5.7 (4.7) years	Mean (SD) T1: 63.6 (29.6) T2: 66.7 (29.2) Mean (SD) difference = 3.10 (8.06), p = 0.05 95% CI: −0.04 to 6.17	• ICC • Spearman's rho (r)	ICC=0.96 (95% CI: 0.92 to 0.98) r=0.95 (p<0.01)	SEM=29.6 × √(1–0.96) = 5.92 MDC=1.96 × 5.92 × √2 = 8.37	(+) Excellent ICC and correlation between scores in a situation where changes in scores were not expected. Bland-Altman plot provided supportive evidence.

(+) adequate performance; (+/−) equivocal performance; (−): inadequate performance.

CASPAR: Classification criteria for Psoriatic Arthritis; ICC: intraclass correlation coefficient; HAQ-DI: Health Assessment Questionnaire-Disability Index; MDC: minimally detectable change; OMERACT: Outcome Measures in Rheumatology; PsA: psoriatic arthritis; PsAQoL: Psoriatic Arthritis Quality of Life index; r: Spearman's rho correlation; SEM: standardized error of mean; SD: standard deviation; SF-36 PF: Medical Outcome Survey Short-Form - 36 item physical functioning domain; T1: time point 1; T2: time point 2.

Table 3.

Summary of Test-Retest Reliability

Measurement Property	**Discrimination: Test-retest reliability**	
Instrument:	HAQ-DI	SF-36 PF
Author/year		
Tillett 2019	Green (+)	Green (+)
Leung 2016	Green (+)	NA
Total available studies for each property	**2**	**1**
Total studies available for synthesis	**2**	**1**
Overall rating	**GREEN**	**AMBER**

(+) adequate performance; (+/−) equivocal performance; (−): inadequate performance.

HAQ-DI: Health Assessment Questionnaire-Disability Index; SF-36 PF: Medical Outcomes Survey Short-Form 36 item physical functioning domain.