The Journal of Rheumatology

# Test-retest Reliability for HAQ-DI and SF-36 PF for the Measurement of Physical Function in Psoriatic Arthritis

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

**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 36-item Short Form Health Survey 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 Outcome Measures in Rheumatology (OMERACT) Filter 2.1 methodology. We calculated intraclass 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 1 for SF-36 PF in PsA. The quality of the datasets was good. The ICCs for HAQ-DI were good and excellent in study 1 (0.90, 95% CI 0.79–0.95) and study 2 (0.94, 95% CI 0.89–0.97). The ICC for SF-36 PF was excellent (0.96, 95% CI 0.92–0.98). 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.

*Key Indexing Terms:* patient-reported outcome, physical function, psoriatic arthritis, test-retest reliability

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 environmental, psychological, or methodological processes. Test-retest reliability is one of the 7 measurement properties to be evaluated under the Outcome Measures in Rheumatology (OMERACT) Filter 2.1.[1]

Physical function (PF) is one of the core domains to be measured in every randomized controlled trial and longitudinal study in psoriatic arthritis (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 PF.[3] The Health Assessment Questionnaire–Disability Index (HAQ-DI) and the 36-item Short Form Health Survey physical function domain (SF-36 PF) have been evaluated by OMERACT Filter 2.1 and received provisional endorsement from the OMERACT and GRAPPA communities. 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 PF, 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.

## 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, whereas the other had data only for HAQ-DI.

The first dataset was derived from a multicenter study in the United Kingdom that aimed to test modifications of various composite measures in 140 patients with PsA classified by the Classification Criteria for Psoriatic Arthritis (CASPAR). Thirty-one 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 were collected for HAQ-DI and SF-36 PF. All questionnaires were administered in paper and pencil format. Stability of the disease between this short 1-week timepoints was assumed, given PsA is a chronic disease. Prior to participation, all patients signed informed consent forms. 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 center in Singapore aimed to evaluate the validity of the Singapore version of the PsA Quality of Life Index.[5] The HAQ-DI was used as a comparator instrument. Out of the 98 recruited patients with PsA who fulfilled the CASPAR, 38 patients who did not require medication change had test-retest reliability data for HAQ-DI. Data were collected 2 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 the condition between this short 2-week timepoints 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

Institutional Review Board E (Ref: 2012/696/E). Prior to participation, all patients signed informed consent.

The quality of each dataset was evaluated by at least 2 independent working group members using the OMERACT Good Method Checklist,[1] and disputes were reconciled. The OMERACT Good Method Checklist assessed 5 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 questions were 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 rank correlations ($\rho$) between scores in test and retest timepoints were calculated for the HAQ-DI and SF36 PF. Bland-Altman plots were generated. Additionally, the minimal detectable change (MDC) was calculated as follows[6]: standard error of measurement $\times 1.96 \times \sqrt{2}$. 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 (SD) age and duration of illness of these 31 patients were 54 (11.0) years and 5.7 (4.7) years, respectively. The quality of this dataset was determined by 2 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 of –0.02 (SD 0.30, $P = 0.77$, 95% 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 0.79–0.95), and Spearman $\rho$ 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, available with the online version of this article). 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 of 3.10 (SD 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 0.92–0.98), and Spearman $\rho$ 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, available with the online version of this article). The quality of this dataset was evaluated by 2 working group members (YYL and WT) and rated as GREEN

*Table 1.* Quality assessment of databases for test-retest reliability using the OMERACT good method checklist.

| | Study 1, Tillett 2021[11] Rating (Remarks) | Study 2, Leung 2016[5] |
|---|---|---|
| **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 timepoints) | Yes, good methods (Paper and pencil format, patients given specific instructions to administer in the same setting and environment at both timepoints) |
| Was the correct statistic used? • Continuous data: ICC • Dichotomous/ordinal/nominal scores: κ used | Yes, good methods (ICC, Spearman rank correlation, and Bland-Altman plot) | Yes, good methods (ICC, Spearman 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** | | |
| 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 is appropriate 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 timepoints) | NA |
| Was the correct statistic used? • Continuous data: ICC • Dichotomous/ordinal/nominal scores: κ used | Yes, good methods (ICC, Spearman 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: 36-item Short Form Health Survey physical functioning domain.

(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] yrs) had data for test-retest reliability for HAQ-DI. The quality of this dataset was assessed by 2 working group members (YYL and PH) 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 timepoint, respectively, with a mean (SD) difference of 0 (0.19) between timepoints ($P > 0.99$), which is lower than the MDC (0.36). The ICC for HAQ-DI was excellent (0.94, 95% CI 0.89–0.97), and Spearman ρ was 0.83 ($P < 0.001$). Bland-Altman plot data showed minimal dispersion around the line of no difference between baseline and 2-week scores (Supplementary Figure 1B, available with the online version of this article). 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 1 dataset of good quality and adequate performance, test-retest reliability was rated as AMBER (Table 3).

## DISCUSSION
Our article reports the evidence synthesized to support test-retest reliability for HAQ-DI and SF-36 PF in PsA. Using

*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 | | MDC $SEM = SD_{baseline} \times \sqrt{(1 - ICC)}$; $MDC = 1.96 \times SEM \times \sqrt{2}$ | Judgment Interpretation of Adequacy (+, +/−, −) |
|---|---|---|---|---|---|---|---|---|
| | Characteristics of Sample | Characteristics of Testing Situation | Sample Recruited and Sample Considered Stable for Analysis | Scores at Baseline and Retest | Statistic Used | Results | | |
| **HAQ-DI** | | | | | | | | |
| (Study 1) Tillett, 2021[11] | • 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) yrs • Mean (SD) duration of PsA 5.7 (4.7) yrs | • 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 rho (ρ) | • ICC 0.90 (95% CI 0.79–0.95) • ρ = 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 not expected. Bland-Altman plot provided supportive evidence |
| (Study 2) Leung, 2016[5] | • 98 consecutive patients with PsA fulfilled CASPAR recruited for validation of PsAQoL study • Mean (SD) duration of illness 5.5 (8.4) yrs | • 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) yrs | • Mean (SD) T1: 0.38 (0.55); T2: 0.35 (0.56) • Mean difference = 0 (SD 0.19), P > 0.99, 95% CI −0.373 to 0.373 | • ICC • Spearman rho (ρ) | • ICC 0.94 (95% CI 0.89–0.97) • ρ = 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, 2021[11] | • 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) yrs • Mean (SD) duration of PsA 5.7 (4.7) yrs | • 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 rho (ρ) | • ICC 0.96 (95% CI 0.92–0.98) • ρ = 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: minimal detectable change; OMERACT: Outcome Measures in Rheumatology; PsA: psoriatic arthritis; PsAQoL: Psoriatic Arthritis Quality of Life; SEM: standardized error of mean; SF-36 PF: 36-item Short Form Health Survey physical functioning domain; T1: timepoint 1; T2: timepoint 2.

*Table 3.* Summary of test-retest reliability as a measurement property of instrument discrimination.

| | HAQ-DI | SF-36 PF |
|---|---|---|
| Study, year | | |
| Tillett, 2021[11] | Green (+) | Green (+) |
| Leung, 2016[5] | 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; NA: not applicable; SF-36 PF: 36-item Short Form Health Survey physical functioning domain.

the OMERACT methodology, test-retest reliability was rated as GREEN (good to go) for HAQ-DI and as AMBER (some caution) for SF-36 PF. To achieve GREEN, at least 1 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 effect 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.[8,9,10] 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 a prespecified 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.

## ONLINE SUPPLEMENT
Supplementary material accompanies the online version of this article.

## 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 Rheumatol 2019;46:1028-35.
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. Ann Rheum Dis 2017;76:673-80.
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 Suppl 2020;96:46-9.
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 Rheum 2018;47:654-65.
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 Musculoskelet Dis 2016;17:432.
6. Beaton DE, Boers M, Tugwell P. 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; 2017:496-508.
7. Boers M, Kirwan JR, Tugwell P, Beaton D, Bingham III CO, Conaghan PG, et al. The OMERACT handbook. [Internet. Accessed May 4, 2021.] Available from: https://omeracthandbook.org/handbook
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 2011;63 Suppl 11:S4-13.
9. Koh ET, Leong KP, Tsou IY, Lim VH, Pong LY, Chong SY, et al; Tan Tock Seng Hospital Rheumatoid Arthritis (TTSH RA) Study Group. The reliability, validity and sensitivity to change of the Chinese version of SF-36 in oriental patients with rheumatoid arthritis. Rheumatology 2006;45:1023-8.
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 Manag 2018; 26:295-301.
11. Tillett W, FitzGerald O, Coates LC, Packham J, Jadon D, Massarotti M, et al. Composite measures for clinical trials in psoriatic arthritis: testing pain and fatigue modifications in a UK multicenter study. J Rheumatol Suppl 2021;97:39-44.