Test-Retest Reliability of Patient Global Assessment and Physician Global Assessment in Rheumatoid Arthritis

GINA ROHEKAR and JANET POPE

ABSTRACT. ***Objective***. As a guide to treatment of rheumatoid arthritis (RA), physicians use measurement tools to quantify disease activity. The Patient Global Assessment (PGA) asks a patient to rate on a scale how they feel overall. The Physician Global Assessment (MDGA) is a similar item completed by the assessing physician. Both these measures are frequently incorporated into other indices. We studied reliability characteristics for global assessments and compared test-retest reliability of both the PGA and the MDGA, as well as other commonly used measures in RA.

Methods. We studied 122 patients with RA age 17 years or older. Patients who received steroid injection or change in steroid dose at the visit were excluded. Patients completed the HAQ, PGA, visual analog scale for pain (VAS Pain), VAS Fatigue, and VAS Sleep. After seeing their physician, they received another questionnaire to complete within 2 days at the same time of day as clinic visit. Physicians completed the MDGA at the time of the patient's appointment and at the end of their clinic day. Test-retest results were assessed using intraclass correlations (ICC). "Substantial" reliability is between 0.61–0.80 and "almost perfect" > 0.80.

Results. Four rheumatologists and 146 patients participated, with 122 questionnaires returned (response rate 83.6%). Test-retest reliability was 0.702 for PGA, 0.961 for MDGA, and 0.897 for HAQ; VAS results were 0.742 for Pain, 0.741 for Fatigue, and 0.800 for Sleep. The correlation between PGA and MDGA was –0.172.

Conclusion. PGA, MDGA, HAQ, and VAS Pain, VAS Fatigue, and VAS Sleep all showed good to excellent test-retest reliability in RA. MDGA was more reliable than PGA. The correlation between PGA and MDGA was poor. (J Rheumatol First Release Sept 15 2009; doi:10.3899/ jrheum.090084)

Key Indexing Terms:
RHEUMATOID ARTHRITIS DISEASE ACTIVITY REPRODUCIBILITY OF RESULTS
PHYSICIAN GLOBAL ASSESSMENT PATIENT GLOBAL ASSESSMENT

Rheumatoid arthritis (RA) is a chronic, inflammatory, autoimmune disease that mainly presents with pain and swelling in the synovial joints[1] and that can lead to joint destruction. Many patients experience pain and fatigue[2].

Monitoring disease activity and damage in RA may be difficult. Because there is no "gold standard" quantitative measure[3], a number of different measures have been used both clinically and in research. Three such measures used most commonly include American College of Rheumatology (ACR) Core Data Set[4], Disease Activity Score (DAS)[5], and Health Assessment Questionnaire (HAQ)[6]. For the ACR Core Data Set, both Patient Global Assessment (PGA) and Physician Global Assessment (MDGA) are measured. The DAS takes the PGA into account.

The test-retest reliability of the PGA and MDGA in RA patients during and after a physician visit has not yet been reported. Recently, Pincus, *et al* published a study in which the primary outcome was to examine the test-retest reliability of various PGA scales[7]. Since the PGA and MDGA are used in trials testing efficacy of new medications, their reliability is important. We studied reliability characteristics for global assessments and compared test-retest reliability of both the PGA and the MDGA in RA patients. Our secondary objectives were to demonstrate reliability of other commonly used measures in RA including the HAQ and visual analog scale for pain (VAS Pain), VAS Fatigue, and VAS Sleep, and to examine the correlation between PGA and MDGA.

MATERIALS AND METHODS

Our study was approved by the University of Western Ontario Ethics Board and conducted in patients with a rheumatologist-confirmed diagnosis of RA. Participating physicians were 4 rheumatologists from St. Joseph's Health Care (London, Canada), an academic hospital. Patients were required to be age 17 years or older and have RA diagnosed by their rheumatologist. Patients who received a steroid injection (intraarticular or

From the Department of Rheumatology, St. Joseph's Health Care, London, Canada.

Supported by an unrestricted grant by the Lawson Health Research Institute's Internal Research Fund.

G. Rohekar, MD, FRCPC; J. Pope, MD, FRCPC.

Address correspondence to Dr. G. Rohekar, St. Joseph's Health Care, Department of Rheumatology, 268 Grosvenor Street, London, ON, N6A 4V2. E-mail: gina.rohekar@sjhc.london.on.ca

Accepted for publication May 21, 2009.

intramuscular) or who had an alteration in oral steroids were excluded from study of the test-retest reliability of the PGA, but not of the MDGA. Exclusion was stipulated so that patients would not experience a treatment effect of the steroid injection between their first and second surveys.

Upon arrival to the clinic, the participating patient completed a one-page, double-sided questionnaire containing the HAQ, PGA, and VAS Pain, VAS Fatigue, and VAS Sleep. The HAQ is a scale of 20 activities of daily living in 8 categories used to assess functional disability, measuring from 0 to 3[6,8]. A higher HAQ score (> 1.0) predicts worse outcomes[6,8]. Each VAS consisted of a 100 mm line with tall vertical lines at beginning (0 mm) and end (100 mm) points, with small vertical lines at 10 mm intervals on the line. The PGA evaluation question was: "Considering all ways in which illness and health conditions may affect you at this time, please make a mark on the line below to show how well you are doing." The 0 mm-end of the line was marked "Very Well" and the 100 mm-end of the line was marked "Very Poorly." Similar VAS were used for pain, sleep, and fatigue. The question for pain was: "How much PAIN have you had because of your illness in the PAST WEEK? Please indicate on the scale below how severe your pain has been." The left side of the scale was marked "No Pain" and the right side was marked "Very Severe Pain." For fatigue, the questionnaire asked: "How much of a problem has UNUSUAL FATIGUE or tiredness been for you OVER THE PAST WEEK? Place a mark on the line below." The left end of the scale was marked "Fatigue is no problem," and the right side of the scale was marked "Fatigue is a major problem." Similarly, the VAS Sleep asked "How much of a problem has SLEEPING been for you OVER THE PAST WEEK? Place a mark on the line below." The left end of the scale was marked "Sleep is no problem" and the right end was marked "Sleep is a major problem." For the MDGA, the 100 mm VAS did not have vertical markers every 10 mm or anchors at the beginning and the end. The question asked: "On this line, where would you rate the patient's arthritis and how it affects him/her today?" The line was marked with the words "None" on the left side and "Maximum" on the right side.

Following assessment by the physician, patients were given a package containing another questionnaire. The second questionnaire was randomly ordered with regards to repeating the various VAS and HAQ. The patient was instructed by the physician that the second questionnaire was to be completed 1 to 2 days after the visit, at about the same time of day as their clinic visit. The time interval of 1 to 2 days was chosen so as to give enough time to reduce the recall of the first questionnaire, but also because it was unlikely that any medical interventions made at the visit would have yet had an effect.

The participating rheumatologists also completed an initial MDGA at the time of the patient's appointment, and then again at the end of their usual clinic day. This shorter time interval was chosen so that the physician would still remember the patient encounter. Physicians were allowed to refer to their notes from the visit, but they were not allowed to access their original MDGA.

The test-retest reliability was assessed using a single-measure, random model intraclass correlation (ICC), the Pearson correlation coefficient, and Spearman's rho. The ICC was used as the primary outcome since it would be the measure least affected by systematic error[9]. Prior to conducting and analyzing the study, we determined how the results of the ICC should be interpreted. We used the following interpretation as described[10]: ICC < 0.00 = poor correlation, ICC between 0.00 and 0.20 = slight correlation, ICC 0.21 to 0.40 = fair correlation, ICC 0.41 to 0.60 = moderate correlation, ICC 0.61 to 0.80 = substantial correlation, and ICC > 0.80 = almost perfect correlation.

The sample size was calculated according to the method described by Walter, *et al*[11]. For the calculation, we assumed a minimum intraclass correlation coefficient (ICC, p_0) of 0.7 and a desired reliability (p_1) of 0.8 with $\alpha = 0.05$ and ß = 0.20. The values chosen for p_0 and p_1 are based on test-retest reliability scores that fall into the "substantial" (p = 0.61–0.80) to "almost perfect" (p > 0.80) range[10]. Thus, at least 118 patients were required for an adequate sample size.

RESULTS

Questionnaires were given to 146 patients, and 122 patients responded (response rate 83.6%). Physicians completed 166 MDGA (20 of their patients who had received steroid injections or changes in steroid dosing did not qualify for the PGA study).

Table 1 shows the demographics and baseline characteristics of our patients, divided into those who returned their questionnaires and those who did not, to identify any differences between the 2 groups: Respondents were noted to be slightly older as compared to non-responders, and had slightly longer disease duration.

Table 2 summarizes results of our test-retest analysis for primary measures of the PGA and MDGA, for secondary measures HAQ, VAS Pain, VAS Fatigue, and VAS Sleep, and the correlation between initial PGA and initial MDGA. The mean value of the PGA was 28.77 [standard deviation (SD) 24.84]; ICC for test-retest reliability was 0.702. In contrast, the mean value for the MDGA was 22.10 (SD 23.46); ICC was 0.961. When looking at correlations between the PGA and MDGA, the ICC was –0.172.

Results of ICC test-retest reliability of the HAQ, VAS Pain, VAS Fatigue, and VAS Sleep are also presented in Table 2.

Table 1. Demographics and baseline characteristics of patients who did and did not respond to the second questionnaire. Value are mean (SD) unless otherwise indicated.

Characteristics	Responders	Non-responders
Age, yrs	59.91 (11.83)	54.29 (12.14)
Female, %	79.63	80.65
Duration of disease, yrs	9.22 (10.29)	6.12 (7.39)
Initial PGA	28.82 (24.79)	27.25 (29.95)
Initial HAQ	0.656 (0.62)	0.813 (0.63)
Initial MDGA	18.41 (21.52)	31.00 (25.53)

PGA: Patient Global Assessment, HAQ: Health Assessment Questionnaire, MDGA: Physician Global Assessment.

Table 2. Test-retest reliability of items.

Test-Retest Item	Mean (SD)	ICC (95% CI)
PGA VAS	28.77 (24.84)	0.702 (0.56, 0.785)
MDGA VAS	22.10 (23.46)	0.961 (0.947, 0.971)
HAQ	0.661 (0.61)	0.897 (0.855, 0.927)
VAS Pain	33.53 (25.66)	0.742 (0.646, 0.813)
VAS Fatigue	36.46 (27.96)	0.741 (0.646, 0.813)
VAS Sleep	31.58 (29.96)	0.800 (0.723, 0.857)
Initial PGA vs initial MDGA	—	–0.172 (–0.717, 0.200)

PGA: Patient Global Assessment, VAS: Visual Analog Scale. MDGA: Physician Global Assessment, HAQ: Health Assessment Questionnaire, ICC: intraclass correlation, CI: confidence interval.

DISCUSSION

When using quantitative measures to make clinical decisions about a patient's care, it is important to verify that the measures being used are appropriate, reliable, and valid. Given that the PGA and MDGA are often incorporated into other frequently used measurements, such as the DAS or the ACR Core Data set, it is important to demonstrate their validity, particularly when looking at indices, since the validity of an index depends on the validity of the measures that are included[12].

The test-retest reliability of the commonly used measures we studied were very good. Using our predetermined criteria, the PGA, VAS Pain, VAS Fatigue, and VAS Sleep showed "substantial" correlation in terms of test-retest reliability. The MDGA and HAQ were even more highly correlated, both showing "almost perfect" test-retest reliability. The highest test-retest reliability was found for the MDGA; however, the reassessment was done after a shorter time interval versus the other measures. Our results confirm the high consistency of physicians when rating disease activity. While patients also had substantial correlation in their test-retest of the PGA, results were less reliable than their physicians'; however, they did have a 1 to 2 day interval between their assessments. It is possible that the difference in reliability seen when comparing physicians to patients is due to the difference in the way the global assessment was measured in each situation. Since the patients scored their repeat PGA 2 days after the first (vs hours later for the MDGA), it

Table 3. Summary of literature review of test-retest reliability of Patient Global Assessment in rheumatoid arthritis (RA).

Author, Reference	Population and Study Design	Results
Hanly[15]	61 patients with RA; teaching clinic or office practice; initial questionnaire within 24 h or MD assessment; followup assessment on 27 patients after mean 3 mo; used RADAR questionnaire, replacing first question with 10 cm VAS of PGA	Looked mainly at correlation between MD and patient assessments; no information about test-retest reliability of PGA vs PGA or MDGA vs MDGA
Hernandez-Cruz[16]	22 patients with RA; time between assessments 90–120 min	Weighted kappa and ICC calculated; kappa = 0.58; ICC = 0.48
Pincus[17]	Test-retest of PGA examined in a subset of patients; total of 688 patients, of which 162 had RA; 112 patients (all-comers) filled out 2 PGA (one at start of visit, one at end of visit): therefore about 27 patients with RA filled out both;	Spearman's correlation coefficient (rho) used; reported as "kappa scores for all items ranging from 0.65 to 0.81 (all $p < 0.001$)..." (data not shown)
Lassere[18]	24 patients (not RA-specific); questionnaires administered on day 1 and day 8	Looked at ICC and SDD; ICC = 0.94; SDD 95% limits of agreement = –18 to 16
Athale[19]	Paper version of PGA vs computer version: no test-retest of same format; convenience sample of 63 RA patients (complete data for 43 patients)	ICC = 0.911
Pincus[7]	Main goal of study to compare traditional linear PGA VAS to a 21-circle version of the PGA, but test-retest reliability of standard VAS also studied; patients had any rheumatologic diagnosis (not specific to RA); 264 patients studied for test-retest of traditional 10 cm VAS; patients completed 2 assessments at the visit; time separation ≤ 1 h	Spearman rank-order correlation and ICC used to estimate test-retest reliability; Spearman correlation = 0.92; ICC = 0.93; both significant at $p < 0.0001$

RADAR: Rapid Assessment of Disease Activity in Rheumatology; PGA: Patient Global Assessment; MDGA: Physician Global Assessment; ICC: intraclass correlations; SDD: smallest detectable difference.

is possible that the lower correlations we are seeing for PGA are related to an actual change in disease course. Since RA activity tends to fluctuate, this is an important consideration.

As well, it should be noted that the format used for the PGA and the MDGA differed slightly, and therefore potentially led to differences in correlation for the PGA versus the MDGA. While the PGA consisted of a scale with vertical markers every 10 mm, the MDGA did not. Although it is unlikely that this had a large impact on the results, it may have led to differences in how physicians and patients scored their respective global assessments. For example, patients may have had a tendency to cluster their responses around the vertical markers.

There was a poor correlation between the PGA and the MDGA. It is clear from this finding that patients and physicians are not similar in their assessment of RA disease activity. The reasons for this discordance are likely multifactorial. It is possible that patients are using the degree of RA damage, not only RA activity, to form their measurements. As well, they may be incorporating non-RA factors, such as life stresses and other causes of pain, into their assessment of disease activity. On the other hand, physicians may be using joint counts and laboratory measurements as their measure when rating overall disease activity. For example, patients may be basing their assessments on subjective experiences of phenomena such as fatigue and pain, whereas the rheumatologists may be basing their rating on more "objective" observations such as swollen joints. It is clear, however, that the PGA and the MDGA are likely measuring different things when it comes to "disease activity." Indeed, the phrasing of the MDGA as "How would you rate the patient's arthritis and how it affects him/her today?" does not inherently ask the physician to evaluate disease activity. It is likely that individual physicians interpret this question differently. We could not find any reported studies to identify which constructs are measured by physicians in the MDGA. Last, it is important to note that the test-retest reliability of the MDGA, as conducted in this study, did not involve actually seeing the patient a second time. Therefore, it may be seen as testing recall rather than testing reliability of scoring and may be difficult to compare to the PGA.

Some authors have promoted the idea that patient-reported outcomes alone may be sufficient to monitor RA disease activity[13,14]. In light of our findings, we cannot recommend patient-reported outcomes alone.

Table 3 summarizes the other known studies in this area. Pincus, *et al* recently compared various formats of the VAS PGA and VAS pain[7], looking at reliability of different forms of the VAS (for example, 21 circles instead of a straight line) versus the traditional VAS[7]. As part of their study, the test-retest reliability of the usual straight-line VAS was examined and the ICC was found to be 0.93 for PGA and 0.94 for pain[7]. Comparing these results to the results of our study, we see that the ICC for PGA "usual VAS" was higher for Pincus' patients, registering in the "almost perfect" range versus "substantial" in our group. A similar difference in results was found for the pain ICC between our studies. Despite the slightly lower ICC in our patients for pain and PGA on the traditional VAS, the results are concordant in showing good reliability.

Limitations of our study include the lack of generalizability to other populations, such as those with other rheumatic diseases or early RA.

In summary, our study shows that the PGA, MDGA, HAQ, and VAS Pain, VAS Fatigue, and VAS Sleep have good test-retest reliability. There is a large discrepancy between a patient's assessment of disease activity and a physician's assessment of disease activity in RA.

REFERENCES

1. Klippel JH, editor. Primer on the rheumatic diseases. 12th ed. Atlanta, Georgia: The Arthritis Foundation; 2001.
2. Koopman WJ, Boulware DW, Heudebert GR, editors. Clinical primer of rheumatology. Philadelphia, PA: Lippincott Williams & Wilkins; 2003.
3. Pincus T, Sokka T. Quantitative measures for assessing rheumatoid arthritis in clinical trials and clinical care. Best Pract Res Clin Rheumatol 2003;17:753-81.
4. Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum 1993;36:729-40.
5. van der Heijde DM, van 't Hof M, van Riel PL, van de Putte LB. Validity of single variables and indices to measure disease activity in rheumatoid arthritis. J Rheumatol 1993;20:538-41.
6. Bruce B, Fries JF. The Health Assessment Questionnaire (HAQ). Clin Exp Rheumatol 2005;23 Suppl 39:S14-8.
7. Pincus T, Bergman M, Sokka T, Roth J, Swearingen C, Yazici Y. Visual analog scales in formats other than a 10 centimeter horizontal line to assess pain and other clinical data. J Rheumatol 2008;35:1550-8.
8. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.
9. Yen M, Lo LH. Examining test-retest reliability: An intra-class correlation approach. Nurs Res 2002;51:59-62.
10. Koepsell TD, Weiss NS. Epidemiologic methods: Studying the occurrence of illness. New York, New York: Oxford University Press, Inc.; 2003.
11. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med 1998;15;17:101-10.
12. Fransen J, van Riel PL. The disease activity score and the EULAR response criteria. Clin Exp Rheumatol 2005;23 Suppl 39:S93-9.
13. Pincus T, Strand V, Koch G, Amara I, Crawford B, Wolfe F, et al. An index of the three core data set patient questionnaire measures distinguishes efficacy of active treatment from that of placebo as effectively as the American College of Rheumatology 20% response criteria (ACR20) or the disease activity score (DAS) in a rheumatoid arthritis clinical trial. Arthritis Rheum 2003;48:625-30.
14. Pincus T. The American College of Rheumatology (ACR) core data set and derivative "patient only" indices to assess rheumatoid arthritis. Clin Exp Rheumatol 2005;23 Suppl 39:S109-13.
15. Hanly JG, Mosher D, Sutton E, Weerasinghe S, Theriault D. Self-assessment of disease activity by patients with rheumatoid arthritis. J Rheumatol 1996;23:1531-8.

16. Hernandez-Cruz B, Cardiel MH. Intra-observer reliability of commonly used outcome measures in rheumatoid arthritis. Clin Exp Rheumatol 1998;16:459-62.
17. Pincus T, Swearingen C, Wolfe F. Toward a Multidimensional Health Assessment Questionnaire (MDHAQ): Assessment of advanced activities of daily living and psychological status in the patient-friendly health assessment questionnaire format. Arthritis Rheum 1999;42:2220-30.
18. Lassere MN, van der Heijde D, Johnson KR, Boers M, Edmonds J. Reliability of measures of disease activity and disease damage in rheumatoid arthritis: Implications for smallest detectable difference, minimal clinically important difference, and analysis of treatment effects in randomized controlled trials. J Rheumatol 2001;28:892-903.
19. Athale N, Sturley A, Skoczen S, Kavanaugh A, Lenert L. A web-compatible instrument for measuring self-reported disease activity in arthritis. J Rheumatol 2004;31:223-8.