The Clinical Assessment of Patients with Psoriatic Arthritis: Results of a Reliability Study of the Spondyloarthritis Research Consortium of Canada

DAFNA D. GLADMAN, RICHARD J. COOK, CATHY SCHENTAG, MARIE FELETAR, ROBERT I. INMAN, CAROL HITCHON, JACOB KARSH, ALICE V. KLINKHOFF, WALTER P. MAKSYMOWYCH, DIANNE P. MOSHER, BINDU NAIR, and MILLICENT A. STONE

ABSTRACT. ***Objectives.*** To evaluate whether rheumatologists experienced in psoriatic arthritis (PsA) assess peripheral and axial involvement in the same way and to consider core clinical measurements that should be included in clinical trials in PsA.

Methods. Ten patients with PsA, representing a broad range of joint inflammation, joint damage, and spinal involvement, were selected for the study. Each patient was examined by each of 10 rheumatologists, members of the Spondyloarthritis Research Consortium of Canada, according to a Latin Square design. Assessments included scoring actively inflamed joints and damaged joints, dactylitis, enthesitis, and spinal measurements. Variance components analyses were conducted for continuous measurements based on models with observer, patient, and order effects. Estimates of intraclass correlation coefficients and associated 95% confidence intervals were obtained.

Results. There was substantial reliability in the assessment of the number of actively inflamed joints and excellent agreement in the number of damaged joints. Only moderate agreement was found for the number of digits with dactylitis. There was excellent agreement among observers in the intermalleolar distance measurements, but there was not as good agreement in the other measurements of spinal mobility. There was good agreement among the observers in detecting plantar fasciitis, however, the other entheses did not fare as well.

Conclusion. In this first multicenter study of the assessment of clinical evaluation of patients with PsA we found that the assessment of peripheral joint disease is reliable although training should be performed prior to initiation of drug trials or comparative studies in this disease. The assessment of back measurements in PsA and other spondyloarthritis requires further study. (J Rheumatol 2004;31:1126–31)

Key Indexing Terms:
PSORIATIC ARTHRITIS SPONDYLITIS RELIABILITY CLINICAL MEASURES

From The Psoriatic Arthritis Program, University of Toronto, Toronto; University of Waterloo, Waterloo; University of Ottawa, Ottawa, Ontario; University of Manitoba, Winnipeg, Manitoba; University of Alberta, Edmonton, Alberta; Dalhousie University, Halifax, Nova Scotia; University of Saskatchewan, Saskatoon, Saskatchewan.

D.D. Gladman, MD, FRCPC, Professor of Medicine, University of Toronto, Director, Psoriatic Arthritis Program; R.J. Cook, PhD, Professor, Department of Statistics and Actuarial Science, University of Waterloo; C. Schentag, MSc, Research Associate; M. Feletar, MBBS(Hons), FRACP, Clinical Research Fellow, Psoriatic Arthritis Program, Centre for Prognosis Studies in the Rheumatic Diseases, Toronto Western Hospital; R.I. Inman, MD, FRCPC, Professor of Medicine, University of Toronto; C. Hitchon, MD, FRCPC, Assistant Professor of Medicine, University of Manitoba; J. Karsh, MDCM, FRCPC, Professor of Medicine, University of Ottawa; A. Klinkhoff, MD, FRCPC, Assistant Professor of Medicine, University of British Columbia; W. Maksymowych, MD, FRCPC, Associate Professor of Medicine, University of Alberta; D. Mosher, MD, FRCPC, Associate Professor of Medicine, Dalhousie University; B. Nair, MD, FRCPC, Assistant Professor of Medicine, University of Saskatchewan. M.A. Stone, MB, MRCP(UK), Assistant Professor of Medicine, University of Toronto.

Address reprint requests to Dr. D. Gladman, Toronto Western Hospital, ECW 5-034B, 399 Bathurst St., Toronto, Ontario, M5T 2S8.
E-mail: dafna.gladman@utoronto.ca

Submitted April 11, 2003; revision accepted December 30, 2003.

Psoriatic arthritis (PsA) is an inflammatory arthritis associated with psoriasis. It has been recognized as a unique entity due to the efforts of Wright and Moll[1]. It is distinguished from rheumatoid arthritis (RA) in that it is usually seronegative for rheumatoid factor; it affects males and females equally; it tends to be asymmetric, especially at onset; it affects distal interphalangeal joints; it affects the spine; it has extraarticular features common to the seronegative spondyloarthropathies; and it is associated with HLA-B*27[2]. Thus PsA has been classified among the spondyloarthropathies. PsA had been considered a mild form of arthritis. The fact that patients with PsA are less tender than patients with RA may have contributed to this concept[3]. However, over the past several years it has been recognized that many patients with PsA develop a severe destructive form of arthritis and become disabled. Indeed the disability noted among patients with PsA is similar to that of patients with RA[4]. Nonetheless, until recently there have been few randomized controlled trials of drug therapy for this condition.

There are no specific measures to assess the physical findings in PsA. The peripheral joints have been assessed either by the American College of Rheumatology (ACR) joint count[5], or by a modification of the Ritchie index[6], both measures developed for RA. While the ACR joint count was validated in PsA, it was proven reliable only within one clinic[5] and has not been tested by other investigators. The assessment of the spondyloarthropathy has included measures used in ankylosing spondylitis (AS)[7,8]. The Assessment in Ankylosing Spondylitis (ASAS) Study Group has defined the core outcome measures for AS[9]. However, the reliability of the back assessment measures has not yet been demonstrated in PsA. An important outcome measure in PsA is the development of damaged joints. The reproducibility of the assessment of peripheral joints and spine by physicians from different clinics has not been studied.

Our objectives were (1) to evaluate whether rheumatologists experienced in PsA assess peripheral and axial involvement in the same way and (2) to consider core clinical measurements that should be included in clinical trials in PsA.

MATERIALS AND METHODS

Patients selection. The sample size was determined to ensure that the width of the 95% confidence interval (CI) for the intraclass correlation coefficient (ICC) would be 0.20 for an ICC estimate of 0.90. From the formula[10] we determined that 10 patients would have to be examined by each of 10 observers. Therefore 10 patients with PsA were recruited for the study from the Psoriatic Arthritis Clinic, Centre for Prognosis Studies in the Rheumatic Diseases at the Toronto Western Hospital. These patients were selected to represent a broad range of joint inflammation, joint damage, and spinal involvement.

Observers. Ten rheumatologists representing centers from across Canada with established interest and experience in PsA participated in the study. These individuals were recruited from among a larger group of Canadian rheumatologists who recently formed the Spondyloarthritis Research Consortium Canada (SPARCC).

Design. All 10 patients were assessed by the same 10 observers according to a Latin Square design[11], which facilitates an analysis of components of variation due to patient, observer, and error, while possibly controlling for order of assessment.

Clinical assessments. Clinical assessments included evaluation of peripheral joint disease, spinal involvement, and presence of enthesitis.

Peripheral joint disease assessment. The number of actively inflamed joints was determined by the number of joints with stress pain, joint line tenderness, and/or swelling. The swollen joints were identified specifically. The number of damaged joints was determined by the number of joints with fixed deformities, flail joints, fused joints, or joints that had undergone surgery. The number of digits with dactylitis, defined as diffuse swelling of a whole digit, was also determined for each patient. Each digit with dactylitis was further defined as being acute or chronic. Grip strength was measured using a sphygmomanometer inflated to 100 mmHg then down to 20 mm/Hg. The patient was asked to squeeze the cuff maximally, and the right and left hand grip strengths were each measured and recorded in mmHg. Examiners were provided with new sphygmomanometers.

Spinal assessment. The following measurements were obtained: chest expansion (cm), measured as the difference between full expiration and full inspiration using a measuring tape at the level of the nipple; finger to floor distance (cm), reflecting the distance from tip of third finger to floor when patient bends forward without bending the knees; lateral flexion of the back (cm), determined by the distance from the tip of the third finger to the floor, as well as the distance between the tip of the third finger and the fibula, without bending the knees or bending forward; and Schober's test based on the change in a 10 cm segment measured in cm placed in the lumbosacral junction between upright position and full flexion. Additional measurements included the Modified Schober's test: a line is drawn at the level of the dimples of Venus to serve as an anchor with a mark 5 cm below the line and 10 cm above the line with the patient upright. The change between the bottom line and top line at full flexion were then measured in cm. The Smythe test begins with the same line drawn between the dimples of Venus, from which 3 consecutive 10 cm segments are marked with the patient in full flexion, the difference in each 10 cm segment from full flexion to full extension is then marked (normally the distance is 2, 3, 4 reflecting the lower thoracic, upper lumbar, lower lumbar regions). The tragus to wall distance, between the tragus of the ear to the wall with patient standing upright position with the heels against the wall, was measured in cm. Occiput to wall distance, between the occiput and wall with the patient standing upright with the heels against the wall, was measured in cm. Cervical rotation (right and left) was graded as 0 normal (> 70°), 1 mild (20–70°), 2 moderate reduction (< 20°) in movement. Cervical bending or lateral flexion (right and left) was graded as 0 normal (> 40°), 1 mild (20–40°), 2 moderate (< 20°) reduction of movement. Cervical flexion was graded as normal or limited. Cervical extension was graded as normal or limited. Sacroiliac pain was tested by the Gaenslen maneuver (patient drops one leg to the side of the examination table, pressure applied on thigh and the opposite iliac bone), FABER test (flexion abduction external rotation of the hip, pressure on abducted thigh and opposite iliac bone); compression over the pelvis with the patient lying on the side. If there was pain in any maneuver, the patient was considered to have sacroiliac stress pain. Intermalleolar distance was measured with the patient lying down, knees straight, spreading feet apart as much as possible. Examiners were provided with tape measures.

Enthesitis. The following entheses were examined: rotator cuff insertion at the shoulder, tibial tuberosity at the knee, Achilles tendon, and plantar fascia insertions in the calcaneus. These are included among the Maastricht AS enthesis score (MASES), which includes 13 anatomic sites[12].

Prior to beginning the evaluations the examiners discussed the definitions for each test and came to a consensus about the general approach to peripheral joint and back examinations, but did not undertake a formal training session. Notes were recorded during these discussions. Each examiner assessed each patient and completed a data collection sheet. The information was then entered into an MS Excel computer database.

Statistical analysis. Variance components analyses were conducted for continuous measurements based on analysis of variance (ANOVA) models with random observer, random patient, and both including and excluding fixed order effects. Estimates of intraclass correlation coefficients (ICC) and associated 95% confidence intervals (CI) were obtained[13]. For categorical measurements, 2 types of analyses were conducted. First, kappa statistics and their associated 95% CI were computed for each categorical measurement[14]. A recently promoted odds ratio (OR) based measure of reliability, called phi, was also computed[15,16]. Phi is computed as $(\sqrt{OR - 1}) \div (\sqrt{OR + 1})$, and like kappa, phi takes on values over the interval [–1, 1], with larger values representing good agreement. Unlike kappa, however, phi can be used in settings where there is a need to adjust for covariates such as the order of the assessment[15,16]. To estimate phi with more than 2 raters, generalized estimating equations were used with the association between assessments parameterized in terms of OR[17]. These regression models included an intercept only, or a single covariate reflecting the order of the assessment, depending on whether order was controlled for or not. We report unadjusted and adjusted estimates of phi and associated 95% CI.

ANOVA was also carried out to assess whether there were differences in the nature of assessments between observers. For convenience we use the term reliability coefficient to refer to the intraclass correlation coeffi-

cient, kappa, and phi, and make it clear which statistic is of interest in each context. Sackett, *et al*[18] suggest that values of kappa in the intervals (0.80, 1.00) represent excellent agreement beyond chance, (0.60, 0.80) substantial agreement beyond chance, (0.40, 0.60) moderate agreement beyond chance, (0.20, 0.40) fair agreement, and (0.0, 0.20) poor agreement beyond chance. For the purpose of interpreting our results we adopt this same classification for the ICC, kappa, and phi but consider 0.4 as minimum level of reliability for each of these measures.

Analyses were computed for each measure, and the results presented in groups according to whether they related to peripheral involvement, spinal involvement, or enthesitis. All statistical analyses were performed using SAS.

RESULTS

Patient characteristics. The 10 patients included 5 men and 5 women with an average age of 52.5 (± 11.7) years and average disease duration of 15.8 (± 10.5) years. Five patients had polyarthritis alone, and 5 had polyarthritis with spondylitis (at least grade 2 sacroiliitis with or without syndesmophytes). Three of the patients had arthritis mutilans. One patient did not have any actively inflamed joints and one did not have any damaged joints. The average numbers of actively inflamed and damaged joints per patient were 22 and 10, respectively. Six patients had dactylitis and 8 had enthesitis (Table 1).

Assessment of peripheral joint disease. The reliability for assessing peripheral joint disease was moderate to substantial. Table 2 displays the estimated reliability coefficients for measures related to peripheral disease activity. There were no meaningful differences between the unadjusted and adjusted analyses and so we discuss only the results from the unadjusted analyses. There was substantial reliability in the assessment of the number of actively inflamed joints and excellent agreement in the number of damaged joints. Only moderate agreement was found for the number of digits with dactylitis. Inter-observer reliability on the clinical assessment of grip strength was found to be excellent for right hands and substantial for left hands. Although the a priori analysis was based on the total number of actively inflamed joints, we also looked at the agreement with regards to the number of swollen joints. Here there was poor agreement with ICC of 0.10 (–0.01, 0.38).

Table 1. Characteristics of the PsA patients participating in the study.

No. of patients	10
Sex, M/F	5/5
Age at study, yrs*	52.5 (11.5)
Disease duration, yrs*	15.8 (10.7)
Arthritis pattern	
Polyarthritis	5
Polyarthritis and spondylitis	5
Arthritis mutilans	3
No. of actively inflamed joints (mean) [median]	22.0 (9) [21.0]
No. of damaged joints (mean) [median]	10.0 (12) [6.0]
Dactylitis	6
Enthesitis	8

* Mean (SD).

Table 2. Reliability analyses for clinical assessments relating to peripheral disease (10 observers, 10 patients), using the intraclass correlation coefficient for each feature.

Feature Assessed	Unadjusted Reliability Coefficient	Unadjusted 95% CI	Adjusted Reliability Coefficient	Adjusted 95% CI
Grip strength, left	0.77	(0.59, 0.92)	0.79	(0.62, 0.93)
Grip strength, right	0.90	(0.80, 0.97)	0.92	(0.83, 0.97)
No. of active joints	0.76	(0.58, 0.92)	0.76	(0.57, 0.92)
No. of swollen joints	0.10	(–0.01, 0.38)	0.10	(–0.01, 0.38)
Dactylitis	0.57	(0.34, 0.82)	0.56	(0.34, 0.82)
No. of damaged joints	0.81	(0.65, 0.94)	0.81	(0.65, 0.94)

Spinal assessment. The findings regarding spinal assessments are reported in Table 3. Among the continuous measures, inter-observer reliability was excellent for the assessments of both right and left lumbar lateral flexion by finger to floor. Substantial agreement was found for the "finger to floor distance" and the intermalleolar distance. Only moderate reliability was found for assessing lumbar lateral flexion by finger to fibula. Other assessments that were less reliable include assessments of chest expansion (moderate), back range of motion (poor), occiput-to-wall distance (moderate), and tragus-to-wall distance (moderate). Fair inter-observer reliability was found for the Schober and poor reliability was found for the modified Schober tests.

Among the categorical measures related to spinal involvement, the findings include fair reliability for cervical rotation (right or left), cervical lateral bending (right or left), and sacroiliac pain. There was substantial to excellent reliability for the assessment of cervical flexion and good reliability for cervical extension. The findings were similar for kappa and phi, and between the unadjusted estimates of phi, and the estimates adjusted for the order of the assessment.

Table 4 reports the findings regarding the analysis of assessments related to enthesitis, which showed overall fair to moderate reliability. The reliability in the assessment of rotator cuff enthesitis was fair, but it was moderate for the assessments of tibial tuberosity enthesitis (left and right, respectively). Achilles enthesitis was moderate, and plantar enthesitis was moderate to substantial.

There was no order effect in this study with the exception of the grip strength, which improved with repeated measurements. There was, however, a significant observer effect for some of the measurements, suggesting that the observers are making their assessments in consistently different ways. Table 5 depicts the percent contribution of observer, patient, and order of examination to the variability detected. The variability due to patients generally exceeded the observer and order effects. Observer effects were greatest for chest expansion, measures of back range of movement, occiput to wall distance, measures of lateral flexion by finger to fibula distance (right and left), and tragus to wall measurements.

Table 3A. Reliability analyses for clinical assessments relating to spinal involvement (10 observers, 10 patients). Continuous variables.

Feature Assessed	Statistic	No. of Assessments*	Unadjusted Reliability Coefficient	Unadjusted 95% CI	Adjusted Reliability Coefficient	Adjusted 95% CI
Chest	ICC	100	0.41	(0.20, 0.73)	0.40	(0.19, 0.71)
Finger-floor distance	ICC	100	0.78	(0.60, 0.92)	0.77	(0.59, 0.92)
Back upper	ICC	100	0.10	(–0.01, 0.38)	0.10	(–0.01, 0.39)
Back middle	ICC	100	0.21	(0.06, 0.53)	0.20	(0.05, 0.52)
Back lower	ICC	100	0.09	(–0.01, 0.37)	0.09	(–0.01, 0.37)
Occiput–wall distance	ICC	100	0.59	(0.37, 0.84)	0.57	(0.34, 0.82)
Intermaleolar distance	ICC	70	0.78	(0.59, 0.93)	0.79	(0.60, 0.93)
R lateral flexion (FFl)	ICC	80	0.80	(0.62, 0.93)	0.80	(0.63, 0.93)
L lateral flexion (FFl)	ICC	80	0.84	(0.69, 0.95)	0.83	(0.68, 0.95)
R lateral flexion (FFib)	ICC	90	0.54	(0.33, 0.81)	0.53	(0.30, 0.80)
L lateral flexion (FFib)	ICC	90	0.58	(0.35, 0.83)	0.57	(0.34, 0.83)
Schober	ICC	100	0.28	(0.10, 0.61)	0.27	(0.10, 0.60)
Modified Schober	ICC	100	0.10	(–0.01, 0.39)	0.10	(–0.01, 0.38)
Tragus to wall	ICC	100	0.53	(0.31, 0.80)	0.51	(0.29, 0.79)

* Three observers did not report intramaleolar distance, 2 did not report lateral flexion finger to floor, and one did not report finger to fibula distance.

Table 3B. Reliablitiy analyses for clinical assessments relating to spinal involvement (10 observers, 10 patients). Categorical variables.

Feature Assessed	Statistic	No. of Assessments*	Unadjusted Reliability Coefficient	Unadjusted 95% CI	Adjusted Reliability Coefficient	Adjusted 95% CI
Cervical rotation R	Kappa	90	0.38	(0.26, 0.51)		
	Phi		0.38	(0.08, 0.61)	0.40	(0.09, 0.63)
Cervical rotation L	Kappa	90	0.29	(0.12, 0.45)		
	Phi		0.35	(0.06, 0.59)	0.36	(0.06, 0.61)
Cervical bending R	Kappa	100	0.28	(0.20, 0.35)		
	Phi		0.25	(0.05, 0.44)	0.25	(0.05, 0.44)
Cervical bending L	Kappa	90	0.23	(0.14, 0.32)		
	Phi		0.21	(0.03, 0.37)	0.22	(0.03, 0.38)
Sacroiliac pain	Kappa	90	0.20	(–0.21, 0.60)		
	Phi		0.29	(0.06, 0.49)	0.31	(0.06, 0.53)
Cervical flexion	Kappa	100	0.78	(0.36, 1.2)		
	Phi		0.84	(0.21, 0.98)	0.89	(0.37, 0.98)
Cervical extension	Kappa	100	0.63	(0.36, 0.90)		
	Phi		0.67	(0.37, 0.85)	0.68	(0.37, 0.85)

* One observer neglected to report on cervical rotation, cervical bending left, and sacroiliac pain.

DISCUSSION

The assessment of patients with PsA has been difficult because of lack of widely accepted classification or diagnostic criteria for the disease. Clinicians have used the Moll and Wright classification[19] as a framework for both classification and diagnosis. Until recently PsA attracted relatively little attention from pharmaceutical companies, perhaps because the disease was considered mild and infrequent. Recently, however, it has become clear that the disease may be severe in a significant proportion of the patients and may be more prevalent than initially thought[20]. With the advent of new and emerging therapies for PsA, it is important that the assessment measures of these patients be reliable and reproducible.

Our study represents the first attempt to evaluate whether the clinical assessments of patients with PsA are reliable. We evaluated the ability of clinicians from different centers to examine the same patients with PsA in a reproducible way. We found excellent agreement in the assessment of peripheral joints, including both joint inflammation and damage. However, while there is excellent agreement, there is still a significant variability due to observer effect. The results suggest that training and standardization of observers are important in order to reduce variability due to systematic effects in both clinical trials and observational cohorts. This variation due to observer effect is particularly important in clinical trials, since it results in the need for larger sample sizes to detect drug effects. Despite the excellent agreement on the total joint count there was not good agreement in the swollen joint count (Table 2). The reason

Table 4. Reliability Analyses for clinical assessment relating to enthesitis (10 observers, 10 patients).

Feature Assessed	Statistic	No. of Assessments*	Unadjusted Reliability Coefficient	Unadjusted 95% CI	Adjusted Reliability Coefficient	Adjusted 95% CI
Rotator cuff right	Kappa	90	0.30	(0.20, 0.40)		
	Phi		0.30	(0.04, 0.52)	0.31	(0.04, 0.54)
Rotator cuff left	Kappa	80	0.43	(0.24, 0.62)		
	Phi		0.45	(0.11, 0.70)	0.46	(0.11, 0.70)
Tibial tuberosity right	Kappa	90	0.64	(0.38, 0.89)		
	Phi		0.67	(0.37, 0.85)	0.73	(0.44, 0.89)
Tibial tuberosity left	Kappa	90	0.54	(0.31, 0.77)		
	Phi		0.58	(0.31, 0.76)	0.60	(0.28, 0.80)
Achilles insertion right	Kappa	90	0.38	(–0.18, 0.94)		
	Phi		0.55	(0.27, 0.75)	0.55	(0.26, 0.75)
Achilles insertion left	Kappa	90	0.47	(0.11, 0.83)		
	Phi		0.56	(0.28, 0.75)	0.56	(0.28, 0.75)
Plantar fascia right	Kappa	90	0.42	(–0.19, 1.02)		
	Phi		0.60	(0.37, 0.76)	0.63	(0.36, 0.80)
Plantar fascia left	Kappa	90	0.61	(0.12, 1.10)		
	Phi		0.72	(0.37, 0.90)	0.72	(0.37, 0.89)

* One observer neglected to report on enthesitis and an additional neglected to report on left rotator cuff.

Table 5. Proportions of variation (% Var) attributable to patients, observers, and order.

Variable	N Assessors*	N Patients	Patient % Var	Patient p	Observer % Var	Observer p	Order % Var	Order p
Active joints	10	10	76.8	0.001	6.8	0.001	1.7	0.492
Damaged joints	10	10	81.7	0.001	5.6	0.001	1.9	0.216
Dactylitis	10	10	58.3	0.001	8.1	0.032	4.0	0.390
Chest expansion	10	10	44.5	0.001	25.8	0.001	3.1	0.511
Finger floor distance	10	10	78.2	0.001	2.9	0.247	1.5	0.720
Back upper	10	10	17.0	0.025	12.1	0.122	11.8	0.134
Back middle	10	10	26.6	0.001	9.9	0.225	5.3	0.684
Back lower	10	10	16.5	0.042	11.6	0.317	8.6	0.378
Occiput–wall distance	10	10	60.8	0.001	14.5	0.001	0.4	0.999
Intermaleolar	7	10	79.5	0.001	3.5	0.088	3.7	0.229
R Lateral flexion (FFl)	8	10	80.9	0.001	3.8	0.037	2.6	0.313
L Lateral flexion (FFl)	8	10	84.7	0.001	4.0	0.009	1.2	0.671
R Lateral flexion (FFib)	9	10	56.8	0.001	13.3	0.001	2.4	0.777
L Lateral flexion (FFib)	9	10	60.2	0.001	11.6	0.002	2.6	0.699
Grip strength right	10	10	90.1	0.001	0.5	0.109	2.4	0.007
Grip strength left	10	10	77.8	0.001	2.8	0.179	4.2	0.033
Schober test	10	10	32.4	0.001	10.8	0.109	5.5	0.568
Modified Schober	10	10	17.6	0.034	10.8	0.234	7.0	0.556
Tragus to wall distance	10	10	54.9	0.001	14.6	0.001	2.4	0.709

* Three observers were not able to carry out the intermalleolar measure, 2 observers did not perform finger to floor (FFl), and 1 observer did not measure finger to fibula (FFib).

for this poor agreement may be the difficulty in assessing joint swelling in patients with PsA, where the effusions may be tight, and particularly in the presence of concomitant joint damage.

Thompson, *et al*[21] compared 4 methods of assessing peripheral joints in RA using a technique similar to the one used in the current study. The observer variation in their study varied from 20 to 30%, compared to 6 to 7% in our study. In another study of clinical assessment variability in rheumatoid arthritis, Klinkhoff, *et al*[22] found that before standardization there was a significant variation due to observers (accounting for 13% of the variation), which was reduced to 3% after standardization of the clinical examination. This was maintained at 6 months' followup. Following a training session these investigators were able to reduce the sample size required to detect a difference of 2 joints from 225 to 91.

The current investigation of inter-observer reliability in PsA detected a smaller percent variation due to observer without a training session. This was seen despite the fact

that patients with PsA have less tenderness and may have tight effusions that are at times harder to detect than patients with RA[3]. However, although the total number of inflamed joints was similar among the observers, there was poor agreement on the number of swollen joints. It is likely that with standardization clinicians can improve the inter-observer variability.

Our study is also the first to address the use of spinal assessment instruments in patients with PsA. We found that while there was excellent agreement among observers in the intermalleolar distance measurements, the agreement in the other measurements of spinal mobility was not as good. In particular there was poor agreement on chest expansion, back movements (both the Schober's, modified Schober's and Smythe tests), occiput to wall, sacroiliac maneuvers, and cervical spine lateral bending and rotation. In a study of reproducibility of spinal measures in AS, Bellamy, *et al*[23] found that the assessment of sacroiliac pain and cervical rotation produced more observer variability. However, these investigators found a much better agreement overall in the spinal assessments among patients with AS. The fact that only 5 patients included in the current study had spondylitis, mainly sacroiliitis, may have contributed to the lack of agreement on the spinal assessments. Moreover, patients with PsA do not have as severe spinal disease as patients with AS. Thus, only some of the spinal assessments may be used in clinical trials in patients with PsA.

The assessment of enthesitis in PsA has not been previously addressed. We found good agreement among the observers in detecting plantar fasciitis and tibial tuberosity; however, the other entheses did not show good agreement. The use of the more extensive enthesitis measure that includes 13 sites is not likely to produce more agreement[12].

Whether those measures that do not show excellent interobserver agreement should be excluded from outcome measures in clinical trials is not clear. It is possible that these measures would perform better if there was appropriate standardization and training of rheumatologists in their use.

In summary, in this first multicenter study of the assessment of clinical evaluation of patients with PsA, we found that the assessment of peripheral joint disease is reliable; however, assessment training should be performed prior to initiation of drug trials or comparative studies in this disease. The assessment of back measurement in PsA and other spondyloarthritis requires further study and standardization of assessment technique.

REFERENCES

1. Wright V, Moll JMH, editors. Psoriatic arthritis. Seronegative polyarthritis. Amsterdam: North Holland Publishing; 1976:169-223.
2. Gladman DD, Rahman P. Psoriatic arthritis. In: Ruddy S, Harris ED, Sledge CB, Budd RC, Sergent JS, editors. Kelly's textbook of rheumatology, 6th edition. Philadelphia: W.B. Saunders; 2001:1071-9.
3. Buskila D, Langevitz P, Gladman DD, Urowitz M, Smythe H. Patients with rheumatoid arthritis are more tender than those with psoriatic arthritis. J Rheumatol 1992;19:1115-9.
4. Sokoll KB, Helliwell PS. Comparison of disability and quality of life in rheumatoid and psoriatic arthritis. J Rheumatol 2001;28:1842-6.
5. Gladman DD, Farewell V, Buskila D, et al. Reliability of measurements of active and damaged joints in psoriatic arthritis. J Rheumatol 1990;17:62-4.
6. Daunt AON, Cox MJ, Robertson JC, Cawley MJD. Indices of disease activity in psoriatic arthritis. J R Soc Med 1987;80:556-8.
7. Hanly J, Russell ML, Gladman DD. Psoriatic spondyloarthropathy: A long term prospective study. Ann Rheum Dis 1988;47:386-93.
8. Gladman DD, Brubacher B, Buskila D, Langevitz P, Farewell VT. Psoriatic spondyloarthropathy in men and women: A clinical, radiographic and HLA study. Clin Invest Med 1992;15:371-5.
9. van der Heijde D, van der Linden S, Dougados M, Bellamy N, Russell AS, Edmonds J. Ankylosing spondylitis: plenary discussion and results of voting on selection of domains and some specific instruments. J Rheumatol 1999;26:1003-5.
10. Donner A, Eliasziw M. Sample size requirements for reliability studies. Stat Med 1987;6:441-8.
11. Montgomery DC. Design and analysis of experiments, 3rd edition. New York: John Wiley & Sons; 1991.
12. Heuft-Dorenbosch L, Spoorenberg A, van Tubergen A, et al. Assessment of enthesitis in ankylosing spondylitis. Ann Rheum Dis 2003;62:127-32.
13. Burdick RK, Graybill FA. Confidence intervals on variance components statistics: Textbooks and monographs. New York: Marcel Dekker; 1992;127.
14. Fleiss JL. Statistical methods for rates and proportions, 2nd edition. New York: Wiley series in probability & mathematical statistics; 1981.
15. Cook RJ, Farewell VT. Conditional inference for subject-specific and marginal agreement: two families of agreement measures. Can J Stat 1995;23:333-44.
16. Meade MO, Cook RJ, Guyatt GH, et al. Interobserver variation in interpreting chest radiographs for the diagnosis of acute respiratory distress syndrome. Am J Respir Crit Care Med 2000;161:85-90.
17. Lipsitz SR, Laird NM, Harrington DP. Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. Biometrika 1991;78:153-60.
18. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology: A basic science for clinical medicine, 2nd edition. Toronto: Little Brown; 1991.
19. Moll JMH, Wright V. Psoriatic arthritis. Semin Arthritis Rheum 1973;3:55-78.
20. Gladman DD. Current concepts in psoriatic arthritis. Curr Opin Rheumatol 2002;14:361-6.
21. Thompson PW, Hart LE, Goldsmith CH, Spector TD, Bell MJ, Ramsden MF. Comparison of four articular indices for use in clinical trials in RA: patient, order and observer variation. J Rheumatol 1991;18:661-5.
22. Klinkhoff AV, Bellamy N, Bombardier C, et al. An experiment in reducing interobserver variability of the examination for joint tenderness. J Rheumatol 1988;15:492-4.
23. Bellamy N, Buchanan WW, Esdaile JM, et al. Ankylosing spondylitis antirheumatic drug trials. I. Effects of standardization procedures on observer dependent outcome measures. J Rheumatol 1991;18:1701-8.
24. Gladman DD, Brubacher B, Buskila D, Langevitz P, Farewell VT. Differences in the expression of spondyloarthropathy: a comparison between ankylosing spondylitis and psoriatic arthritis. Genetic and gender effects. Clin Invest Med 1993;16:1-7.