

Single-joint Outcome Measures: Preliminary Validation of Patient-reported Outcomes and Physical Examination

ALISON E. HEALD, EDWARD J. FUDMAN, PERVIN ANKLESARIA, and PHILIP J. MEASE, on behalf of the 13G01 Study Team

ABSTRACT. Objective. To assess the validity, responsiveness, and reliability of single-joint outcome measures for determining target joint (TJ) response in patients with inflammatory arthritis.

Methods. Patient-reported outcomes (PRO), consisting of responses to single questions about TJ global status on a 100-mm visual analog scale (VAS; TJ global score), function on a 100-mm VAS (TJ function score), and pain on a 5-point Likert scale (TJ pain score) were piloted in 66 inflammatory arthritis subjects in a phase 1/2 clinical study of an intraarticular gene transfer agent and compared to physical examination measures (TJ swelling, TJ tenderness) and validated function questionnaires (Disabilities of the Arm, Shoulder and Hand scale, Rheumatoid Arthritis Outcome Score, and the Health Assessment Questionnaire). Construct validity was assessed by evaluating the correlation between the single-joint outcome measures and validated function questionnaires using Spearman's rank correlation. Responsiveness or sensitivity to change was assessed through calculating effect size and standardized response means (SRM). Reliability of physical examination measures was assessed by determining interobserver agreement.

Results. The single-joint PRO were highly correlated with each other and correlated well with validated functional measures. The TJ global score exhibited modest effect size and modest SRM that correlated well with the patient's assessment of response on a 100-mm VAS. Physical examination measures exhibited high interrater reliability, but correlated less well with validated functional measures and the patient's assessment of response.

Conclusion. Single-joint PRO, particularly the TJ global score, are simple to administer and demonstrate construct validity and responsiveness in patients with inflammatory arthritis. (ClinicalTrials.gov identifier NCT00126724) (First Release March 15 2010; J Rheumatol 2010; 37:1042–8; doi:10.3899/jrheum.090827)

Key Indexing Terms:

OUTCOME ASSESSMENT
RHEUMATOID ARTHRITIS

SELF-ASSESSMENT

VISUAL ANALOG SCALE
PSORIATIC ARTHRITIS

Outcome measures, such as the American College of Rheumatology responder criteria¹ and the European League Against Rheumatism response as measured by the Disease Activity Score^{2,3}, have been developed and validated to assess response to systemic agents in inflammatory arthritis. In contrast, no validated measures for assessing the response of single joints exist^{4,5}, despite a history of local treatment of individual refractory joints with intraarticular steroid

injections⁶, radioactive synovectomy⁷, and surgery⁸. Now that new local therapies for refractory joints are under development, it is important that single-joint outcome measures be developed and validated as well^{4,5}. These single-joint outcome measures need to be developed in parallel with local therapies for refractory joints, so that the novel therapies can be properly evaluated. Systemic outcome measures cannot be used to assess single joints, because they are not sensitive enough to evaluate changes in a single joint after local therapies.

The phase 2 portion of a phase 1/2 study of an intraarticular gene transfer agent⁹ provided an opportunity to compare the utility of patient-reported outcomes (PRO) and physical examination in assessing response of a single joint to local treatment. The Outcome Measures in Rheumatology Clinical Trials (OMERACT) filter^{10,11} was applied to assess truth (face validity, construct validity), discrimination (reliability, responsiveness), and feasibility of these measures.

From Targeted Genetics Corporation; Seattle Rheumatology Associates, Swedish Medical Center, Seattle, Washington; and Austin Rheumatology Research, Austin, Texas, USA.

Supported by Targeted Genetics Corporation.

A.E. Heald, MD; P. Anklesaria, PhD, Targeted Genetics Corporation; E.J. Fudman, MD, Austin Rheumatology Research; P.J. Mease, MD, Seattle Rheumatology Associates, Swedish Medical Center.

Address correspondence to Dr. P.J. Mease, Seattle Rheumatology Associates, 1101 Madison St., Suite 1000, Seattle, WA 98104, USA.

E-mail: pmease@nwlink.com

Accepted for publication December 18, 2009.

MATERIALS AND METHODS

Study setting and subjects. Single-joint outcome measures were piloted and evaluated in the phase 2 portion of a multicenter, randomized, double-blind, placebo-controlled trial of intraarticular administration of a gene transfer agent in subjects with rheumatoid arthritis (RA), psoriatic arthritis (PsA), or ankylosing spondylitis (AS), and persistent moderate or severe inflammation in the knee, ankle, elbow, wrist, or metacarpophalangeal (MCP) joint⁹. In the parent study (ClinicalTrials.gov identifier NCT00126724), subjects were randomized to 1 of 3 doses of the gene transfer agent or placebo. The assessments outlined below were completed by study subjects and study staff without knowledge of treatment assignment. The research was conducted in compliance with the Helsinki Declaration. The trial was approved by the Institutional Review Board at each participating site.

Patient-reported outcomes. Subjects completed validated questionnaires to assess target joint (TJ) function, consisting of the Disabilities of the Arm, Shoulder and Hand (DASH) scale^{12,13} for subjects whose TJ was in the upper extremity, and the Rheumatoid Arthritis Outcome Score (RAOS)^{14,15} for subjects whose TJ was in the lower extremity. Subjects also completed the Health Assessment Questionnaire (HAQ)^{16,17} as an assessment of overall function.

After completing the questionnaires, subjects responded to 3 more questions about the TJ, each on a 100-mm visual analog scale (VAS): (1) "Considering all the ways your target joint affects you, place a vertical line on the scale to show how you have been doing in the past week," (0 = no symptoms and 10 = severe symptoms; TJ global score); (2) "To what extent has your target joint impaired (prevented) you from doing your usual activities during the past week?" (0 = no impairment and 10 = severe impairment; TJ function score); and (3) "Comparing your target joint today with how it felt just prior to your last injection, how satisfied are you with the results of the study drug injection?" (0 = not satisfied at all and 10 = very satisfied; TJ response score).

The DASH and HAQ scores were calculated according to published criteria^{12,17}. For the RAOS, 3 subscales (pain, stiffness, and function) were calculated using the questions that were originally derived from the Western Ontario and McMaster Osteoarthritis Index (WOMAC), and scored in a similar manner¹⁸. The response to each question was on a 5-point scale ranging from 0 = none to 4 = extreme. The pain subscale was calculated by adding the responses to 5 questions about pain, dividing by 20, and multiplying by 100, to obtain a score that ranged from 0 to 100. The stiffness subscale was calculated by adding the responses to 2 questions about stiffness, to obtain a score that ranged from 0 to 8. The function subscale was calculated by adding the responses to the 17 questions about physical function, dividing by 68, and multiplying by 100 to obtain a score that ranged from 0 to 100.

Physical examination. The TJ was evaluated by 2 independent examiners for tenderness on a 4-point scale ranging from 0 = none to 3 = severe (TJ tenderness) and swelling on a scale ranging from 0 = none to 3 = severe (TJ swelling), based on guidelines published in the *Dictionary of Rheumatic Diseases*¹⁹. The 2 examiners performed the joint assessments independently, each without knowledge of the other's assessment.

Outcome measure selection. Prior to analysis of the data, PRO measures considered to be most clinically relevant were prospectively selected by 2 rheumatologists. Both are clinician researchers, and one co-chairs the OMERACT single-joint assessment working group, which is charged with conducting research in this subject area. The outcome measures selected included the TJ global score, the TJ function score, the TJ pain score, and a question about the average severity of pain in the past week included in both the DASH and RAOS, whose response ranged from 0 = none to 4 = extreme. Outcome measures considered to be of interest as surrogate markers of disease included TJ swelling and TJ tenderness.

Analysis. The OMERACT filter was applied to assess truth (face validity, construct validity), discrimination (reliability, responsiveness), and feasibility of the selected outcome measures^{10,11}. Face validity and feasibility were assessed through item selection by the expert panel of rheumatolo-

gists as outlined above. Construct validity was assessed by comparing the results of the candidate single-joint outcome measures (TJ global score, TJ function score, TJ pain score, TJ swelling, TJ tenderness) with validated functional measures (HAQ scores, RAOS pain, stiffness and function subscales, and DASH scores) 12 weeks after injection of a study drug using Spearman's rank correlation coefficient. Twelve weeks after injection was chosen as the time for evaluation, because the dataset at that time had few missing datapoints, and sufficient time had passed since study entry to allow a greater range of responses, because TJ swelling was required to be moderate (grade 2) or severe (grade 3) at study entry.

The sensitivities to change for each measure were compared using both effect sizes and standardized response means (SRM). Effect size was defined as the absolute mean change from baseline to Week 12 for each measure divided by the baseline SD of that measure^{20,21}. SRM was defined as the absolute mean change from baseline to Week 12 divided by the SD of the change from baseline to Week 12²². Effect sizes and SRM were considered large (> 0.8), moderate (0.5 to 0.8), or small (0.2 to 0.5)^{22,23}. In addition, the correlation between the patient-reported response (TJ response score) and the change from baseline for each measure was assessed using Spearman's rank correlation coefficient.

Reliability of the physical examination measures of TJ swelling and TJ tenderness was determined by assessing interobserver agreement. The proportion of observations with complete agreement, defined as when both observers graded the swelling or tenderness to be identical, was determined at each time after first injection of the study drug. In addition, weighted κ coefficients were calculated.

RESULTS

Study population. Single-joint outcome measures were piloted in 66 subjects in the study, including 55 (83%) with RA, 9 (14%) with PsA, and 2 (3%) with AS. The mean age of the 51 women and 15 men was 53.3 years, with a range of 22 to 76 years. The TJ included 20 knees (30%), 14 ankles (21%), 18 wrists (27%), 9 MCP joints (14%), and 5 elbows (8%). The construct validity and responsiveness of the proposed single-joint outcome measures were assessed in 61 subjects with complete data at baseline and Week 12. The HAQ was administered to all 61 subjects. The RAOS was administered to 31 subjects whose TJ included 18 knees and 13 ankles. The DASH scale was administered to 30 subjects whose TJ included 16 wrists, 9 MCP joints, and 5 elbows. The reliability of the physical examination measures was assessed in 63 subjects whose TJ was assessed by 2 independent examiners on the same day at 1 or more times.

Construct validity. The distributions of the proposed single-joint measures at baseline and Week 12 are shown in Figures 1, 2, and 3. The TJ global score and TJ function score were fairly evenly distributed at both baseline and Week 12, without ceiling or floor effects (Figure 1). TJ swelling was either moderate or severe at baseline as per study entry criteria, with the exception of 1 subject whose swelling decreased from moderate to mild between screening and baseline (Figure 3A). TJ swelling became more normally distributed by Week 12. Similarly, TJ tenderness (Figure 3B) and the TJ pain score (Figure 2) were more severe at baseline than Week 12.

The correlation between single-joint outcome measures

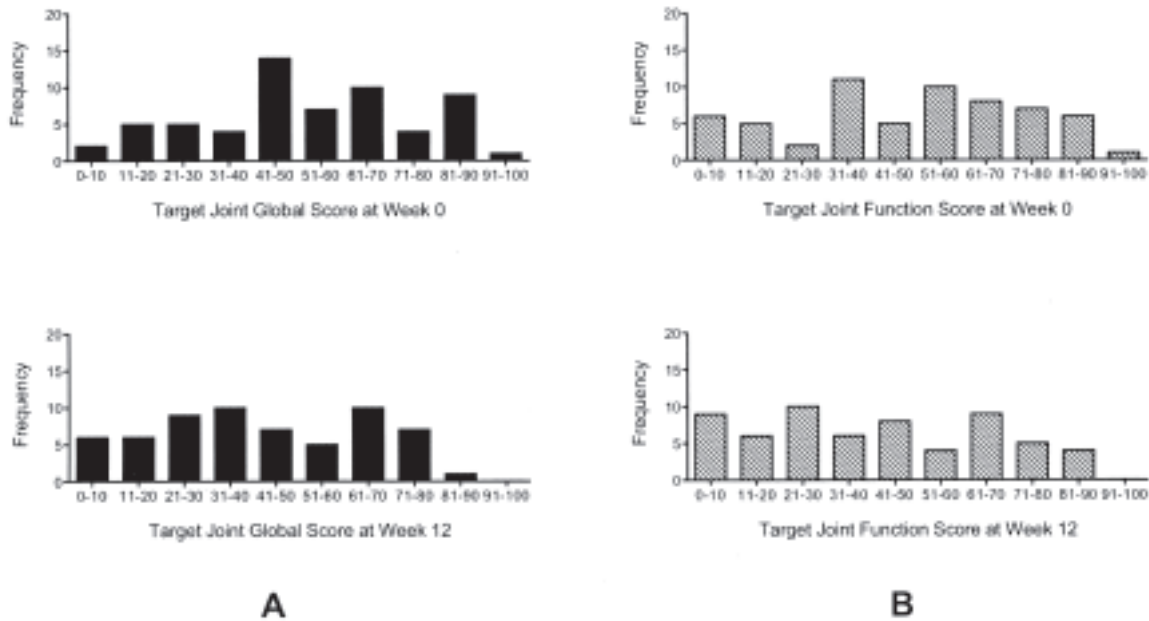


Figure 1. A. Distribution of target joint global scores at Week 0 and Week 12; 0 = no symptoms, 100 = severe symptoms. B. Distribution of target joint function scores at Week 0 and Week 12; 0 = no impairment, 100 = severe impairment.

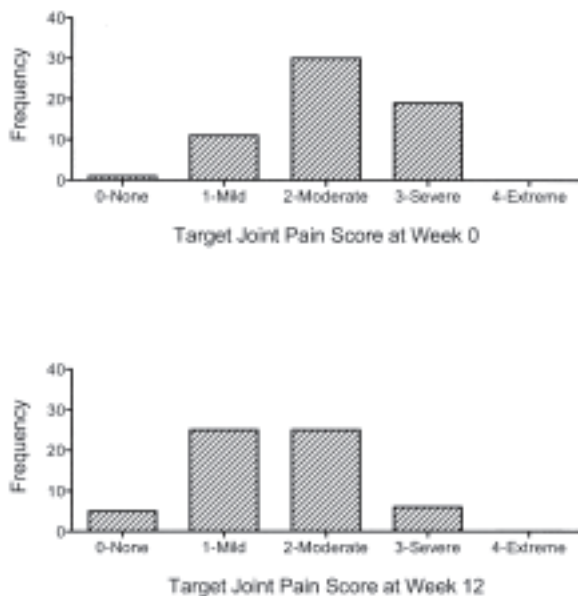


Figure 2. Distribution of target joint pain scores at Week 0 and Week 12.

and validated functional measures is shown in Table 1. The single-joint PRO (TJ global score, TJ function score, TJ pain score) were highly correlated with each other, with Spearman's rank correlation coefficients ranging from 0.76 to 0.90 at Week 12, all with high levels of statistical significance. The single-joint PRO also correlated well with the RAOS and moderately well with the DASH and HAQ. Physical examination measures (TJ swelling, TJ tenderness) correlated well with each other ($r = 0.58, p < 0.0001$), but TJ tenderness was better correlated with single-joint PRO and

validated functional measures than TJ swelling. TJ swelling did not correlate with single-joint PRO or validated functional measures. Of note, the RAOS pain, stiffness, and function subscales were highly correlated with each other, with Spearman's rank correlation coefficients ranging from 0.78 to 0.87 at Week 12, all with $p < 0.0001$ (data not shown).

Responsiveness. Measurements of responsiveness or sensitivity to change are presented in Table 2. TJ swelling had a large effect size and moderate SRM. The RAOS stiffness subscale had a moderate effect size and a large SRM. TJ global score, TJ pain score, and RAOS pain subscale had both a moderate effect size and a moderate SRM. TJ tenderness and the RAOS function subscale had either a moderate effect size or moderate SRM. The HAQ and DASH scale had both a small effect size and a small SRM. Correlation with the TJ response score was strongest for the TJ global score among all joints ($r = -0.37, p = 0.01$), and for the DASH among upper extremity joints ($r = -0.42, p = 0.04$).

Reliability. The reliability of physical examination measures, as assessed by the proportion of observations with complete agreement, is shown in Figure 4. Interobserver reliability was very good, with rates of complete agreement between assessors ranging from 67% to 78% for TJ swelling and 75% to 84% for TJ tenderness at 7 timepoints. Major disagreements (a difference of 2 or more points) were noted for TJ swelling at rates of 0% (3 timepoints) to 5% (1 timepoint) and for TJ tenderness at rates of 0% (1 timepoint) to 3% (3 timepoints). Weighted kappa coefficients ranged from 0.45 (Week 0) to 0.87 (Weeks 18 and 24) for TJ swelling and from 0.61 (Week 0) to 0.75 (Week 24) for TJ tenderness.

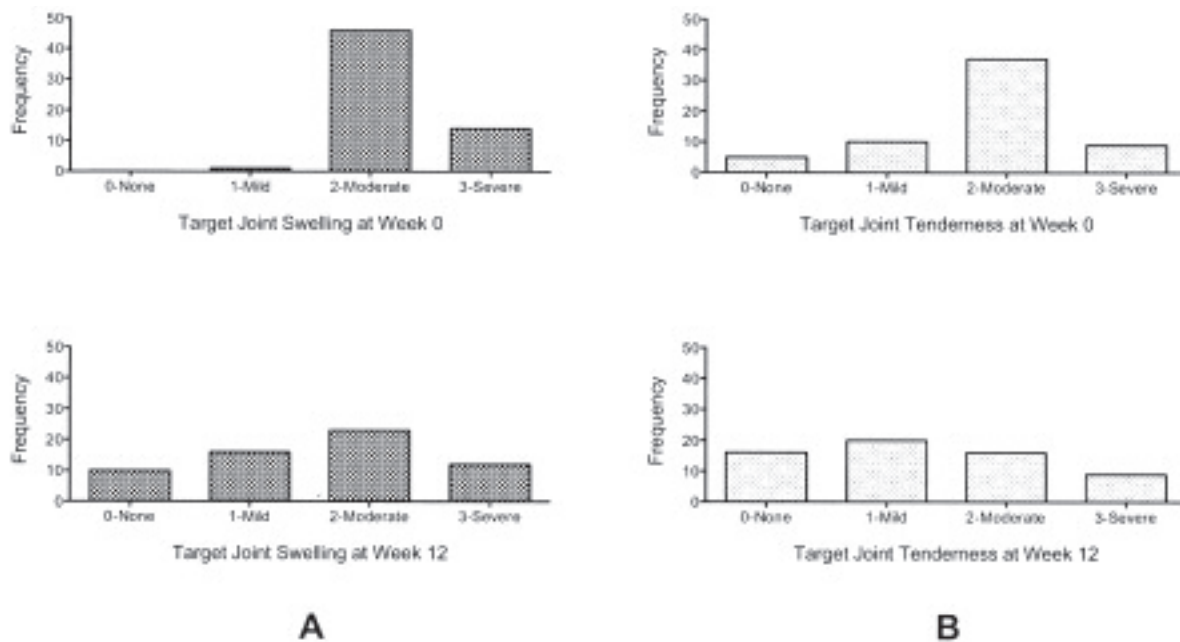


Figure 3. A. Distribution of target joint swelling on physical examination at Week 0 and Week 12. B. Distribution of target joint tenderness on physical examination at Week 0 and Week 12.

Table 1. Construct validity: Spearman's rank correlation for single-joint measures and validated functional measures.

	TJ Global	Single-joint PRO	TJ Pain	Physical Examination	
	Score, r (p)	TJ Function Score, r (p)		TJ Swelling r (p)	TJ Tenderness r (p)
Single-joint PRO					
TJ global score	1	—	—	0.30 (0.02)	0.54 (< 0.0001)
TJ function score	0.90 (< 0.0001)	1	—	0.15 (0.26)	0.42 (0.0007)
TJ pain score	0.81 (< 0.0001)	0.76 (< 0.0001)	1	0.26 (0.05)	0.57 (< 0.0001)
RAOS subscales					
Pain	0.85 (< 0.0001)	0.82 (< 0.0001)	0.83 (< 0.0001)	0.27 (0.15)	0.43 (0.01)
Stiffness	0.70 (< 0.0001)	0.65 (0.0001)	0.76 (< 0.0001)	0.25 (0.18)	0.44 (0.01)
Function	0.79 (< 0.0001)	0.83 (< 0.0001)	0.78 (< 0.0001)	0.18 (0.34)	0.40 (0.03)
DASH	0.53 (0.002)	0.59 (0.0006)	0.61 (0.0004)	-0.01 (0.97)	0.42 (0.02)
HAQ	0.47 (0.0002)	0.59 (< 0.0001)	0.46 (0.0002)	-0.03 (0.83)	0.30 (0.02)

DASH: Disability of Arm, Shoulder, Hand; HAQ: Health Assessment Questionnaire; PRO: patient-reported outcomes; r: Spearman's rank correlation coefficient; RAOS: Rheumatoid Arthritis Outcome Score; TJ: target joint.

DISCUSSION

The need to develop single-joint outcome measures has become more pressing, as more local therapies for refractory joints are under development^{4,5}. Ideally, a single-joint outcome measure would be simple to administer and applicable to all joints. The DASH and RAOS have been developed to assess upper and lower extremities, respectively, but are cumbersome, and neither can be applied to all types of joints.

The single-joint outcome measures (TJ global score, TJ function score, and TJ pain score) evaluated here are simple to administer. Their face validity and feasibility were judged to be good, based on their selection by the expert panel of

rheumatologists. The 3 single-joint outcome measures are highly correlated, suggesting that one may be used in lieu of the others. In comparing the 3 proposed single-joint outcome measures, the TJ global score may prove to be most useful. It demonstrated good construct validity, as seen by the high correlation with the HAQ, RAOS subscales, and DASH score, and good responsiveness, as demonstrated by the moderate effect size and moderate SRM.

Single-joint physical examination measures (TJ swelling, TJ tenderness) were very reliable, with high rates of inter-observer agreement, but they did not correlate as well with validated functional measures as the single-joint PRO. The effect sizes for TJ swelling and TJ tenderness were large and

Table 2. Responsiveness: effect size, standardized response means (SRM), and correlation with target joint response for single-joint measures and validated functional measures.

	Baseline	Change from Baseline to Week 12	Effect Size	SRM	Correlation with TJ Response Score	
					r	p
Single-joint PRO						
TJ global score (0–100)	54 ± 23	–13 ± 20	0.56	0.62	–0.37	0.01
TJ function score (0–100)	49 ± 26	–10 ± 23	0.37	0.41	–0.21	0.17
TJ pain score (0–4)	2.1 ± 0.7	–0.6 ± 0.8	0.77	0.75	–0.28	0.07
Physical examination						
TJ swelling (0–3)	2.2 ± 0.5	–0.6 ± 0.9	1.35	0.69	–0.27	0.08
TJ tenderness (0–3)	1.8 ± 0.8	–0.5 ± 1.1	0.66	0.47	–0.29	0.06
HAQ (0–3)	1.1 ± 0.7	–0.1 ± 0.3	0.10	0.25	–0.22	0.16
RAOS						
Pain subscale (0–100)	48 ± 23	–13 ± 21	0.59	0.65	–0.09	0.71
Stiffness subscale (0–8)	4.3 ± 1.7	–1.2 ± 1.4	0.68	0.83	–0.32	0.16
Function subscale (0–100)	43 ± 26	–10 ± 18	0.40	0.58	–0.16	0.50
DASH (0–100)	37 ± 20	–4 ± 14	0.22	0.31	–0.42	0.04

DASH: Disability of Arm, Shoulder, Hand; HAQ: Health Assessment Questionnaire; PRO: patient-reported outcomes; r: Spearman's rank correlation coefficient; RAOS: Rheumatoid and Arthritis Outcome Score; SRM: standardized response mean; TJ: target joint.

moderate, respectively, with SRM that were moderate and small, respectively, but changes in these physical examination measures did not correlate well with the patient's assessment of response (TJ response score). Physical examination and PRO measure different features of affected joints, and so are unlikely to be highly correlated. Both have their advantages and limitations, and should be used in parallel.

The lack of a “gold standard” for single-joint response hampers the evaluation of prospective single-joint measures in this study. Ideally, the prospective single-joint measures would be compared with a measure of biologic activity. Magnetic resonance imaging (MRI) scans were planned as part of this clinical study protocol, but too few MRI scans were conducted to allow for a meaningful comparison. The rheumatologists considered TJ swelling to be a surrogate for disease activity, but TJ swelling did not correlate well with the patient's symptoms. The candidate single-joint measures were compared with the patient's assessment of response, as recorded on a 100-mm VAS as the TJ response score. However, the TJ response score did not meet criteria of face validity, as the expert panel of rheumatologists felt that patients would have difficulty remembering what their TJ felt like 12 weeks previously.

Based on this preliminary assessment, single-joint PRO, specifically the TJ global response, show promise as valid measures of response in the single joint. However, additional work is required to establish the best single-joint outcome measure. Input from patients with inflammatory arthritis should be obtained to increase the face validity of prospective single-joint outcome measures. Similar to the RAOS

pain, stiffness, and function subscores, the TJ global score, TJ function score, and TJ pain score were highly correlated. Input from patients with inflammatory arthritis may provide valuable insight into distinguishing these measures, or concluding that the concepts are indistinguishable for a single joint from the patient perspective. In addition, there may be room for improvement in the measures themselves. For example, the TJ pain score was based on the response on a 5-point Likert scale to a question embedded within DASH and RAOS, since that was what was available in this study. A pain score based on a 100-mm VAS, similar to the TJ global score and TJ function score, may prove to have better metrics.

Additional work is also required to increase the reliability and responsiveness of the prospective outcomes measures. The test-retest reliability should be assessed by administering the measures at 2 different times in close proximity. If no suitable “gold standard” of response can be identified, it would be helpful to administer these measures after administration of a local treatment that is known to be effective, rather than evaluating it in the context of a clinical study, where it is not known if the intervention is effective. Although a simple patient-reported outcome is desirable for feasibility, a composite measure incorporating both a patient-reported outcome and physician assessment of inflammation may prove to be more useful in determining single-joint response.

ACKNOWLEDGMENT

The 13G01 Study Team: Philip J. Mease, Nathan Wei, Edward J. Fudman, Alan Kivitz, Joy Schechtman, Robert G. Trapp, Kathryn Hobbs, Maria

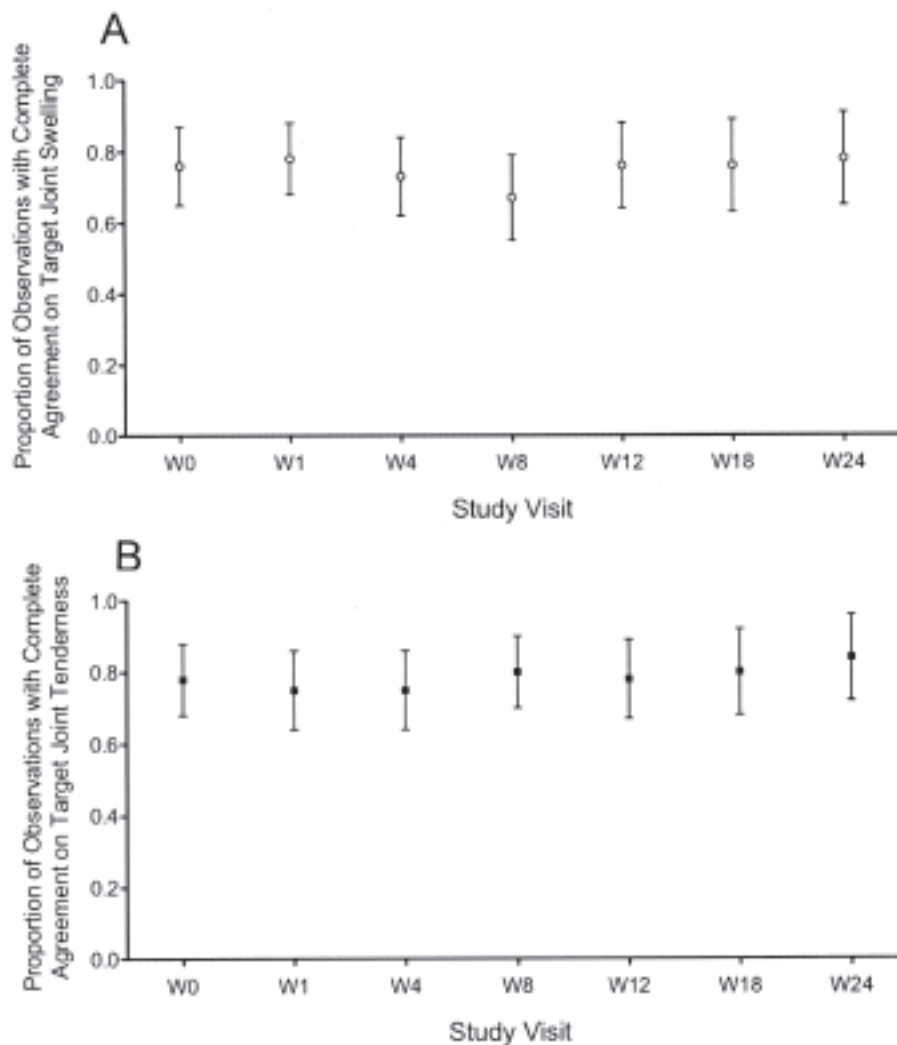


Figure 4. Proportion of observations with complete agreement on target joint swelling (A) and target joint tenderness (B) on physical examination. Swelling was graded on a scale from 0 (none) to 3 (severe). Tenderness was graded on a scale from 0 (none) to 3 (severe). Complete agreement was achieved when both observers graded the swelling or tenderness to be identical. Error bars represent 95% confidence intervals.

Greenwald, Antony Hou, Stephen Bookbinder, Galen Graham, Craig Wiesenhutter, Larry Willis, Eric Ruderman, Joseph Z. Forstot, Michael Maricic, Charles Pritchard, Kathryn Dao, Francis Burch, Darrell Fiske, and Malin Prupas.

REFERENCES

1. Felson DT, Anderson JJ, Boers M, Bombardier C, Furst D, Goldsmith C, et al. American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. *Arthritis Rheum* 1995;38:727-35.
2. van Riel PL, van Gestel AM. Clinical outcome measures in rheumatoid arthritis. *Ann Rheum Dis* 2000;59 Suppl 1:i28-31.
3. van Riel PL, van Gestel AM, van de Putte LB. Development and validation of response criteria in rheumatoid arthritis: steps towards an international consensus on prognostic markers. *Br J Rheumatol* 1996;35 Suppl 2:4-7.
4. Giles JT, Mease P, Boers M, Bresnihan B, Conaghan PG, Heald A, et al. Assessing single joints in arthritis clinical trials. *J Rheumatol* 2007;34:641-7.
5. Keen HI, Bingham CO 3rd, Bradley LA, Conaghan PG, Heald AE, Kaeley GS, et al. Assessing single joints in arthritis clinical trials. *J Rheumatol* 2009;34:2092-6.
6. Schumacher HR, Chen LX. Injectable corticosteroids in treatment of arthritis of the knee. *Am J Med* 2005;118:1208-14.
7. Schneider P, Farahati J, Reiners C. Radiosynovectomy in rheumatology, orthopedics, and hemophilia. *J Nucl Med* 2005;46 Suppl 1:48S-54S.
8. Fiocco U, Cozzi L, Rigon C, Chieco-Bianchi F, Baldovin M, Cassisi GA, et al. Arthroscopic synovectomy in rheumatoid and psoriatic knee joint synovitis: long-term outcome. *Br J Rheumatol* 1996;35:463-70.
9. Mease PJ, Wei N, Fudman EJ, Kivitz AJ, Schechtman J, Trapp RG, et al. Safety, tolerability and clinical outcomes after intra-articular injection of a recombinant adeno-associated vector containing a TNF- α antagonist gene: results of a phase 1/2 study. *J Rheumatol* 2010; [In press].

10. Bellamy N. Clinimetric concepts in outcome assessment: the OMERACT filter. *J Rheumatol* 1999;26:948-50.
11. Boers M, Brooks P, Strand CV, Tugwell P. The OMERACT filter for Outcome Measures in Rheumatology. *J Rheumatol* 1998;25:198-9.
12. Navsarikar A, Gladman DD, Husted JA, Cook RJ. Validity assessment of the Disabilities of Arm, Shoulder, and Hand Questionnaire (DASH) for patients with psoriatic arthritis. *J Rheumatol* 1999;26:2191-4.
13. Raven EE, Haverkamp D, Sierevelt IN, van Montfoort DO, Pöhl RG, Blankevoort L, et al. Construct validity and reliability of the disability of arm, shoulder and hand questionnaire for upper extremity complaints in rheumatoid arthritis. *J Rheumatol* 2008;35:2334-8.
14. Bremander AB, Petersson IF, Roos EM. Validation of the Rheumatoid and Arthritis Outcome Score (RAOS) for the lower extremity. *Health Qual Life Outcomes* 2003;1:55.
15. Karatepe AG, Günaydin R, Türkmen G, Kaya T, Ozbek G. The validity and reliability of the Turkish version of the Rheumatoid and Arthritis Outcome Score (RAOS) in patients with rheumatoid arthritis. *Clin Rheumatol* 2009;28:47-51.
16. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: Dimensions and practical applications. *Health Qual Life Outcomes* 2003;1:20.
17. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: a review of its history, issues, progress, and documentation. *J Rheumatol* 2003;30:167-78.
18. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *J Rheumatol* 1988;15:1833-40.
19. American Rheumatism Association. Appendix A — Joint Examination. In: *Dictionary of the rheumatic diseases, volume 1: signs and symptoms*. 3rd ed. New York: Contact Associates International, Ltd.; 1988:75-6.
20. Hawley DJ, Wolfe F. Sensitivity to change of the Health Assessment Questionnaire (HAQ) and other clinical and health status measures in rheumatoid arthritis: results of short-term clinical trials and observational studies versus long-term observational studies. *Arthritis Care Res* 1992;5:130-6.
21. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care* 1989;27(3 Suppl):S178-89.
22. Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. *Arthritis Rheum* 1985;28:542-7.
23. Cohen J. *Statistical power analysis for the behavior sciences*. 2nd ed. New York: Academic Press; 1977.