ARTHRITIS & RHEUMATISM
Vol. 43, No. 3, March 2000, pp 506–514

CLINICAL IMPROVEMENT AS REFLECTED IN MEASURES OF FUNCTION AND HEALTH-RELATED QUALITY OF LIFE FOLLOWING TREATMENT WITH LEFLUNOMIDE COMPARED WITH METHOTREXATE IN PATIENTS WITH RHEUMATOID ARTHRITIS

Sensitivity and Relative Efficiency to Detect a Treatment Effect in a Twelve-Month, Placebo-Controlled Trial

PETER TUGWELL, GEORGE WELLS, VIBEKE STRAND, ANDREAS MAETZEL, CLAIRE BOMBARDIER, BRUCE CRAWFORD, CATHERINE DORRIER, and ANN THOMPSON, on behalf of the LEFLUNOMIDE RHEUMATOID ARTHRITIS INVESTIGATORS GROUP

Objective. **To examine correlations between clinical improvement as defined by the American College of Rheumatology (ACR) responder analysis and clinical improvement as determined by 4 function and/or health-related quality of life measures, and to estimate the sensitivity and relative efficiency of these measures compared with changes in the tender joint count in patients with rheumatoid arthritis (RA).**

Methods. **A 52-week, multicenter, double-blind controlled trial was conducted to compare treatment with leflunomide (n = 182), methotrexate (n = 180), or placebo (n = 118) in patients with active RA. ACR response rates and improvement in scores on the Health Assessment Questionnaire (HAQ), Problem Elicitation Technique (PET), and Medical Outcomes Survey Short Form 36 (SF-36) were compared in 438 of the patients.**

Results. **In comparing leflunomide with placebo, the patient global assessment, HAQ disability index, and SF-36 bodily pain scale were most responsive to treatment group differences. The modified HAQ (M-HAQ), PET Top 5, SF-36 physical component score, physician global assessment, pain intensity scale, and SF-36 physical functioning scale were more responsive to treatment group differences than was the tender joint count. In comparing methotrexate with placebo, the patient and physician global assessments were most responsive. These 2 measures, as well as the pain intensity scale and the C-reactive protein level, were more responsive to treatment group differences than was the tender joint count, while the SF-36 mental health component score was least responsive. A close correlation between changes in the M-HAQ and HAQ scores indicated that the M-HAQ was similarly responsive to change over time. Improvements in the PET, SF-36 physical component score, bodily pain, and physical functioning scales correlated with the ACR responder status.**

Supported by Hoechst Marion Roussel.

Peter Tugwell, MD, George Wells, PhD, Andreas Maetzel, MD, Claire Bombardier, MD: Ottawa Hospital, Ottawa, Ontario, Canada; Vibeke Strand, MD: Stanford University, Palo Alto, CA; Bruce Crawford: Mapi Values USA, LLC; Catherine Dorrier: Quintiles Transnational; Ann Thompson: Hoechst Marion Roussel, Kansas City, Missouri.

Drs. Tugwell and Strand served as independent paid consultants to Hoechst Marion Roussel, the sponsor of the studies described in this report.

Members of the Leflunomide Rheumatoid Arthritis Investigators Group are as follows: Elizabeth Tindall, MD, Portland, OR; Howard Offenberg, MD, Gainesville, FL; Jeffrey Poiley, MD, Orlando, FL; Joel Rutstein, MD, San Antonio, TX; Frederick Dietz, MD, Rockford, IL; Alan Brodsky, MD, Dallas TX; Robert Harris, MD, Whittier, CA; Mitchell Lowenstein, MD, Palm Harbor, FL; Andrew Baldessare, MD, St. Louis, MO; Paul Howard, MD, Paradise Valley, AZ; Marshall Sack, MD, Austin, TX; John Tesser, MD, Phoenix, AZ; Nathan Wei, MD, Frederick, MD; Robin Dore, MD, Anaheim, CA; Robert Ettlinger, MD, Tacoma, WA; Larry Anderson, MD, Portland, ME; Barry Bockow, MD, Seattle, WA; Michael Liebling, MD, Torrance, CA; Paul Romain, MD, Burlington, MA; Scott Baumgartner, MD, Spokane, WA; Joel Silverfield, MD, Tampa, FL; Andrew Chubick, MD, Dallas, TX; Charles Franklin, MD, Willow Grove, PA; Selwyn Cohen, MD, Trumbull, CT; Gordon Senter, MD, Salisbury, NC; Richard Furie, MD, Manhasset, NY; Sanford Hartman, MD, Decatur, GA; Robert Levy, MD, Olympia, WA; David Yocum, MD, Tucson, AZ; Don Cheatum, MD, Dallas, TX; Sicy Lee, MD, New York, NY; Matthew Heller, MD, Peabody, MA.

Address reprint requests to Peter Tugwell, MD, Chairman, Department of Medicine, Ottawa Hospital, General Campus, 501 Smyth Road, Room LM-12, Ottawa, Ontario K1H8L6, Canada.

Submitted for publication July 12, 1999; accepted in revised form November 11, 1999.

Conclusion. **Both disease-specific and generic measures of function and health-related quality of life detect improvements in RA patients. Using both types of measures for evaluating therapies will identify discernible changes that are important to patients, and will facilitate comparisons across different disease states.**

In the treatment of diseases, such as rheumatoid arthritis (RA), that can interfere with the day-to-day activities of patients, it is useful to assess the effect of treatment on those aspects of the disease that are most important to the patient. These disease parameters may differ from the clinical manifestations of the disease that are most commonly measured by the clinician. Although several well-validated instruments have been designed to measure changes in function and health-related quality of life that can result from effective treatment of the disease, there is limited information on the sensitivity and relative efficiency of these instruments to detect treatment effects. Such changes identify what, to the patient, may be considered true treatment effects, and so should be used in the management of the patient's treatment regimen.

Several disease-specific instruments have been developed to assess functional status in RA. The Health Assessment Questionnaire (HAQ) (1), modified Health Assessment Questionnaire (M-HAQ) (2), and Problem Elicitation Technique (PET) (3) have been validated in randomized, controlled clinical trials (4–7). Responses elicited with these instruments can reflect impairment in performance of daily and other essential activities. The HAQ and M-HAQ are a standard set of questions related to disabilities that are most frequently present in general populations of patients with RA. The PET individualizes this by focusing on those disabilities that are present in each patient due to the arthritis, and asks patients to indicate which disabilities are most important to them and which they would most like to see improved. All 3 instruments have been shown to reflect change in disease status and to correlate with the American College of Rheumatology (ACR) response criteria (8).

In one of the first studies to demonstrate that these instruments can detect small, clinically important improvements, Bombardier et al compared several measures of outcome used in a 6-month, randomized, placebo-controlled trial of auranofin in 294 patients with active RA (9). The HAQ was as responsive as the traditional measures of tender and swollen joint counts, 10-cm visual analog scale (VAS) for pain, and patient global assessment of arthritis. Buchbinder et al estimated the relative efficiency of several clinical and functional outcome measures to detect a treatment effect in a 6-month, randomized, placebo-controlled trial of cyclosporine in 144 patients with active RA (7). The relative efficiency of 4 measures of pain, including the 10-cm VAS pain scale, and 3 measures of function/health-related quality of life (PET, HAQ, and Arthritis Impact Measurement Scales), was compared with that of the tender joint count. Patient and physician global assessments were most responsive to change, but the findings were not statistically different from those demonstrated by the tender joint count. The swollen joint count, 10-cm VAS pain scale, PET, and HAQ were of intermediate responsiveness, with relative efficiencies ranging from 0.33 to 0.58.

Generic health-related quality of life scales allow comparisons of benefit of therapeutic interventions across different disease states. To date, these instruments have been used more frequently in disciplines other than rheumatology, which has been slow to adopt them. In this era of cost constraints, resource allocation may be influenced significantly by the demonstration of benefit, and its relative magnitude, provided by generic measures of health-related quality of life. It is important, therefore, to know if treatment interventions that we believe to be efficacious in patients with rheumatic diseases are confirmed to demonstrate meaningful improvement with the use of generic measures.

The Medical Outcomes Study 36-item short form (SF-36) is a generic instrument for assessing health-related quality of life that has been validated in normal and diseased populations (10). Beaton et al evaluated the SF-36, as well as the Nottingham Health Profile, the Health Status Section of the Ontario Health Survey, the Duke Health Profile, and the Sickness Impact Profile, in 127 workers with musculoskeletal problems (11). In individuals who experienced a change in health, the SF-36 was the most responsive measure. The standardized response means (SRM), defined as the ratio of the mean observed change to the standard deviation of the difference scores, ranged from 0.81 to 1.13. Ruta and colleagues (12) assessed the responsiveness of the SF-36 to change in 240 British RA patients observed over 3 months, comparing it with the ACR response criteria, which included the M-HAQ. Over time, the largest SRMs were evident in the SF-36 bodily pain scale (>0.8) and physical component score (0.61). The SF-36 physical functioning and vitality scales reflected moderate sensitivity, with SRMs between 0.2 and 0.5. These values were in contrast to SRMs of 0.41 and 0.85 for the VAS pain scale and M-HAQ, respectively. The SF-36 was similarly assessed by Wells et al in a multicenter, controlled trial comparing generic quality of life instruments in 40 patients beginning methotrexate therapy who were ex-

amined at baseline and at 3 and 6 months (13). Although the SF-36 physical component score was not as sensitive to change as the Nottingham Health Profile and Rheumatoid Arthritis Quality of Life measure following 6 months of treatment, it showed similar positive (57.1%) and negative (83.3%) agreements with the ACR ≥20% response criteria.

This report presents comparisons of 4 function/health-related quality of life measures—the HAQ, M-HAQ, PET, and SF-36—with respect to their sensitivity and relative efficiency to detect treatment differences in a 12-month, randomized, placebo-controlled clinical trial. The clinical results from this study have been published previously (14,15).

PATIENTS AND METHODS

Patients. A 12-month, multicenter, randomized, double-blind, placebo-controlled study to assess the safety and efficacy of leflunomide treatment compared with placebo and methotrexate was conducted in patients with RA. Patients were randomly assigned in a 3:2:3 distribution to 1 of 3 treatment groups: leflunomide 20 mg daily, placebo, or methotrexate 7.5–15.0 mg weekly. A total of 482 patients were enrolled, and 480 (182 receiving leflunomide, 118 placebo, and 180 methotrexate) were evaluable for clinical response using a modified intent-to-treat analysis, which was defined as all patients who received at least 1 dose of study drug with at least 1 followup visit.

This report presents a secondary analysis of 438 patients (166 receiving leflunomide, 102 placebo, and 170 methotrexate), all of whom completed baseline and 1 or more followup HAQ and SF-36 questionnaires. This population included 438 patients instead of 480 because 4 subjects did not complete a baseline questionnaire (a validated Spanish translation of the SF-36 was lacking at the time of this study's initiation), 20 subjects exited early without completing followup questionnaires, and 18 questionnaires were excluded due to inconsistent responses (9 at baseline and 9 at followup), as calculated by the "response consistency index" developed by the Health Institute (16). The demographics and baseline characteristics of those patients who did not complete the questionnaires were similar to those of the entire protocol population.

Measures of function and health-related quality of life. Information on the following measures was collected at baseline and at each monthly visit as components of the ACR response criteria, which was the primary outcome measure for this study: tender and swollen joint counts (28 joints), patient and physician global assessments of disease activity (on a 0–100 mm VAS), pain intensity scale (0–100 mm VAS), M-HAQ (described below), Westergren erythrocyte sedimentation rate, and C-reactive protein (CRP) level. Function and health-related quality of life were assessed across the 3 treatment groups using the mean change from baseline at 24 and 52 weeks, or at study withdrawal, in each of the following measures: HAQ, PET, and SF-36.

The HAQ, a disease-specific instrument, asks 20 questions to assess 8 functional categories. Responses to each question are scored by patients on a scale of 0 (without difficulty) to 3 (unable to do), with regard to whether or not aids are required to perform these activities. The worst scores in each category are then summed and divided by the number of categories, to give a disability index. Mean changes in each functional category of the HAQ are also reported, as well as an unweighted sum of the means in each category, divided by the number of categories (17). The M-HAQ assessment is a single page of 8 questions about functional activities performed on a daily basis; these are derived from the HAQ, and responses are each scored by patients on a scale of 0 (without difficulty) to 3 (unable to do). The M-HAQ score is derived by summing the total score and dividing by the number of categories (7).

The PET asks patients to identify those functional activities (in this protocol, as enumerated by the HAQ) that are most affected by their RA and that they would most like to see improved by treatment. The PET described in this report is a modification of that used in previous studies where patients were allowed to select any activity and were not limited to items only present in the HAQ. Patients then rank the difficulty, severity, and/or frequency of performing these activities on a 7-point scale, as well as their level of importance. The weighted Top 5 score of the PET is determined by summing the scores for the 5 most important problems, calculated as difficulty multiplied by importance.

The SF-36 is a generic instrument with scores (0–100) based on responses to individual questions, summarized into 8 scales, which include function domains and aspects of well-being: physical functioning, role-physical, bodily pain, general health, vitality, social functioning, role-emotional, and mental health. These 8 scales are also combined into summary phys-

Table 1. Demographics and disease characteristics of the study patients at baseline*

	Leflunomide (n = 182)	Placebo (n = 118)	Methotrexate (n = 180)
Sex, % female	73	70	75
Age, mean ± SD years	54.1 ± 12.0	54.6 ± 10.7	53.3 ± 11.8
Disease duration, mean ± SD years	7.0 ± 8.6	6.9 ± 8.0	6.5 ± 8.1
% with disease ≤2 years	39	33	40
% rheumatoid factor positive	65	60	59
Mean ± SD number DMARDs failed	0.8 ± 1.0	0.9 ± 0.9	0.9 ± 1.0
% with no prior DMARD treatment	45	40	44

* $P > 0.05$ for all baseline comparisons. DMARDs = disease-modifying antirheumatic drugs.

Table 2. Mean changes in function and health-related quality of life measures at end point versus baseline in the intent-to-treat cohort*

	Leflunomide	Placebo	Methotrexate
HAQ disability index			
No.	166	101	169
Baseline score	1.30	1.31	1.30
Mean change	−0.45†‡	0.03	−0.26§
Mean % change versus placebo	37	–	22
M-HAQ			
No.	182	118	180
Baseline score	0.78	0.89	0.79
Mean change	−0.29†‡	0.07	−0.15§
Mean % change versus placebo	45	–	27
PET Top 5			
No.	166	101	170
Baseline score	21.2	22.4	20.4
Mean change	−6.9†‡	−0.66	−3.41§
Mean % change versus placebo	29	–	13
SF-36 physical component			
No.	157	101	162
Baseline score	30.0	28.9	29.7
Mean change	7.6†‡	1.0	4.6§
Mean % change versus placebo	22	–	12
SF-36 mental component			
No.	157	101	162
Baseline score	46.8	48.3	48.5
Mean change	1.5	0.8	0.9
Mean % change versus placebo	1	–	0

* HAQ = Health Assessment Questionnaire; M-HAQ = modified HAQ; PET = Problem Elicitation Technique; SF-36 = Short Form 36 (of the Medical Outcomes Study).
† $P < 0.0001$ versus placebo.
‡ $P < 0.01$ versus methotrexate.
§ $P < 0.05$ versus placebo.

ical and mental component scores, which are again scored from 0 to 100, with higher scores reflecting better quality of life.

Statistical analysis. All analyses were performed on the modified intent-to-treat patient population. Because data from the health-related quality of life instruments were not normally distributed at baseline, the Wilcoxon rank sum test was applied to continuous variables and the Mantel-Haenszel chi-square test to categorical variables, to compare baseline characteristics.

For each outcome measure, several calculations were performed to determine the measure's ability to detect a treatment effect, first for leflunomide treatment compared with placebo, and then for methotrexate compared with placebo. The mean change from the beginning to the end of the study (using the intent-to-treat population, last value carried forward) was calculated for both the treatment and placebo groups. The observed treatment effect was calculated as the percentage difference between the mean change in the treatment group and the mean change in the placebo group.

The standardized effect size (SES) was calculated as the ratio of the treatment effect to the pooled standard deviation of the standard deviations of the mean change scores. An approximate Z test was used to compare the SES of the instrument to the SES for tender joint count, using the following equation:

$$\frac{(SES_{instrument} - SES_{joint\ count})}{\sqrt{2(1/n_t + 1/n_c)(1 - |r|)}}$$

where n_t and n_c is the number of patients in the treatment group and placebo group, respectively, and r is the observed, pooled within-group correlation of the 2 outcome measures.

The relative efficiency was calculated in relation to the tender joint count for each instrument, by taking the square of the SES for the instrument to the SES of the tender joint count. A relative efficiency >1 would imply that the instrument

Table 3. Positive and negative agreement of the generic quality of life measures with the ACR responder criteria*

	20% ACR responder		50% ACR responder		70% ACR responder	
	Yes	No	Yes	No	Yes	No
Leflumomide						
No. of patients	93	85	61	119	36	142
PET (Top 5)	68.8	56.5	55.7	69.2	52.8	80.3
SF-36 mental component	24.7	64.7	8.2	77.8	2.8	82.4
SF-36 physical component	59.1	61.2	44.3	67.5	30.6	74.6
SF-36 pain	73.1	51.8	70.5	53.8	55.6	55.6
SF-36 physical functioning	57.0	52.9	54.1	57.3	44.4	57.7
Methotrexate						
No. of patients	98	82	41	139	17	163
PET (Top 5)	65.9	64.3	51.2	79.1	47.1	85.9
SF-36 mental component	22.0	72.4	9.8	82.0	11.8	84.0
SF-36 physical component	51.2	68.4	56.1	77.7	41.2	80.4
SF-36 pain	68.3	52.0	61.0	61.2	76.5	65.6
SF-36 physical functioning	57.3	61.2	61.0	68.3	58.8	74.2

* Except where otherwise indicated, values are the percentage of American College of Rheumatology (ACR)–defined responder patients with more than 20% improvement in the quality of life measure (positive agreement) or the percentage of nonresponder patients with less than 20% improvement in the quality of life measure (negative agreement). See Table 2 for other definitions.

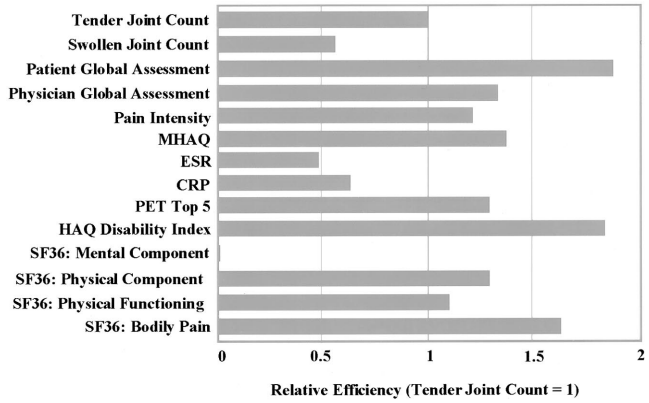

Figure 1. Relative efficiency (compared with tender joint count) of the outcome measures analyzed for leflunomide versus placebo in the intent-to-treat cohort. MHAQ = modified Health Assessment Questionnaire; ESR = erythrocyte sedimentation rate; CRP = C-reactive protein; PET = Problem Elicitation Technique; SF-36 = Medical Outcomes Study Short Form 36.

is more efficient than the tender joint count in detecting a treatment effect.

Comparisons of ≥20%, ≥50%, and ≥70% improvement in the PET weighted Top 5 scores, SF-36 physical and mental component scores, and physical functioning and bodily pain scales in responders versus nonresponders (according to the ACR ≥20%, ≥50%, and ≥70% improvement criteria) were examined. In addition, the HAQ, PET, and SF-36 were examined at baseline for ceiling/floor effects. Ceiling effects refer to the percentage of people scoring the best possible score despite having active disease. Thus, although a small degree of impairment may be present, it would be impossible for that score to improve. Floor effects refer to the percentage of people scoring the worst possible score in a category measured by the questionnaire, even though there may still be room for further deterioration. To allow for the greatest amount of responsiveness in an instrument over time, it is ideal for ceiling/floor effects to be negligible.

RESULTS

There were no significant differences between treatment groups in demographic or baseline disease characteristics at study entry (Table 1).

Active treatment versus placebo. As shown in Table 2, statistically significant improvements in function and health-related quality of life were reported with leflunomide and methotrexate treatment when compared with placebo, as measured by the M-HAQ, disability index of the HAQ, weighted Top 5 score of the PET, and physical component score of the SF-36. There was no significant improvement, however, in the mental health component score of the SF-36. The relative percentage improvement compared with placebo in each of the parameters is also presented in Table 2.

Changes in individual patients. Substantial concordance between patients who were ACR responders and those who demonstrated ≥20%, ≥50%, and ≥70% improvement in these disease-specific and generic measures of function and health-related quality of life was evident. The mental health component score of the SF-36, however, had low positive agreement. Positive and negative agreements are presented in Table 3 for both active treatments. Positive and negative agreements were expressed as the percentage of patients who were classified as either ACR responders or nonresponders according to whether they achieved >20% or <20% improvement, respectively, in each quality of life measure.

There were no potential ceiling effects for either

Table 4. Relative efficiency of various outcome measures to detect a treatment effect for leflunomide versus placebo in the intent-to-treat cohort*

Measure	Observed treatment effect	SES	Relative efficiency	Z statistic	*P*
Tender joint count	−4.69	−0.59	1.00		
Swollen joint count	−2.78	−0.44	0.56	1.40	0.162
Patient global assessment	−2.18	−0.81	1.88	1.97	0.049
Physician global assessment	−1.84	−0.68	1.33	0.92	0.356
Pain intensity	−17.51	−0.65	1.21	0.51	0.809
M-HAQ	−0.36	−0.69	1.37	0.80	0.422
ESR	−8.82	−0.41	0.48	1.18	0.237
CRP	−1.09	−0.47	0.63	0.79	0.428
PET Top 5	−6.26	−0.67	1.29	0.58	0.561
HAQ disability index	−0.42	−0.80	1.84	1.60	0.110
SF-36 mental component	−0.65	−0.06	0.01	3.25	0.001
SF-36 physical component	−8.52	−0.67	1.29	0.58	0.565
SF-36 physical functioning	−14.34	−0.62	1.10	0.21	0.831
SF-36 bodily pain	−15.76	−0.73	1.63	1.19	0.236

* SES = standardized effect size; ESR = erythrocyte sedimentation rate; CRP = C-reactive protein (see Table 2 for other definitions).

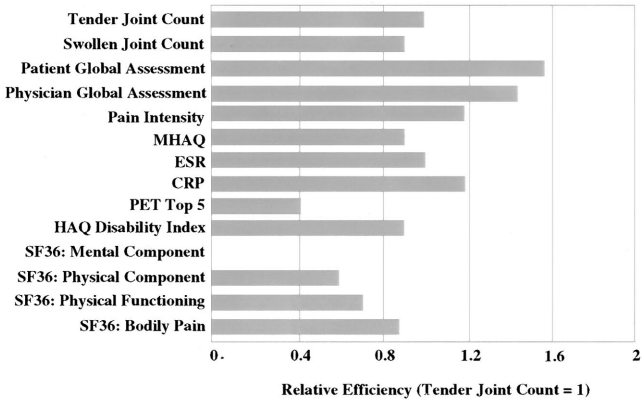

Figure 2. Relative efficiency (compared with tender joint count) of the outcome measures analyzed for methotrexate versus placebo in the intent-to-treat cohort. See Figure 1 for definitions.

the mental or physical component of the SF-36. Among the SF-36 subscales, there was a potential ceiling effect for role-emotional, in which 41% of subjects scored the highest possible value. A potential floor effect was found for the role-physical scale of the SF-36, with 59% of patients scoring the lowest possible value at baseline. Two percent of patients had sufficiently high baseline HAQ scores to preclude improvement of 20% or higher (a potential ceiling effect), and 5% had baseline PET Top 5 values that precluded 20% improvement.

Comparison of outcome measures in terms of ability to detect a treatment effect. The relative efficiencies of the various outcome measures to detect a treatment effect relative to tender joint count (for which the relative efficiency was 1) for leflunomide compared with placebo are shown graphically in Figure 1, and tabulations are presented in Table 4. The relative efficiency ranged from 0.01 (SF-36 mental component score) to 1.88 (patient global assessment of disease activity). In decreasing order, the HAQ disability index (relative efficiency 1.84), SF-36 bodily pain scale (relative efficiency 1.63), M-HAQ (relative efficiency 1.37), physician global assessment (relative efficiency 1.33), PET Top 5 and SF-36 physical component score (both relative efficiency 1.29), pain intensity scale (relative efficiency 1.21), and SF-36 physical functioning scale (relative efficiency 1.10) were more sensitive to treatment group differences than was the tender joint count.

When comparing methotrexate with placebo, patient and physician global assessments were most sensitive to treatment effect. This is presented graphically in Figure 2, and tabulations are presented in Table 5. These 2 measures, as well as the pain intensity scale and CRP level, were more sensitive to treatment effects than was the tender joint count, whereas disease-specific and generic measures of function were not. Nonetheless, the relative efficiencies ranged from 0.42 (PET Top 5) to 0.71–0.88 (SF-36 physical functioning scale, physical component score, and bodily pain scale). The relative efficiencies for both the HAQ and M-HAQ were the same, 0.91, and approached the relative efficiency of the tender joint count.

Table 6 shows the PET Top 5 activities, from among 20 specified in the HAQ, that were selected by patients across all treatment groups as being most important to them. All functional activities were selected by at least 30 patients, and no 1 item was selected by more than 204 of the 437 patients assessed with this measure. The most frequently selected activities were "doing chores," "standing from chair," and "dressing

Table 5. Relative efficiency of various outcome measures to detect a treatment effect for methotrexate versus placebo in the intent-to-treat cohort*

Measure	Observed treatment effect	SES	Relative efficiency	Z statistic	*P*
Tender joint count	−3.58	−0.45	1.00		
Swollen joint count	−2.46	−0.43	0.91	0.18	0.859
Patient global assessment	−1.67	−0.56	1.55	0.95	0.341
Physician global assessment	−1.40	−0.54	1.44	0.68	0.378
Pain intensity	−12.86	−0.49	1.19	0.35	0.729
M-HAQ	−0.22	−0.43	0.91	0.17	0.884
ESR	−9.04	−0.45	1.00	0.00	1.000
CRP	−0.97	−0.49	1.19	0.26	0.791
PET Top 5	−2.75	−0.29	0.42	1.21	0.227
HAQ disability index	−0.23	−0.43	0.91	0.15	0.879
SF-36 mental component	−0.05	−0.01	0.00	2.66	0.008
SF-36 physical component	−3.53	−0.35	0.80	0.76	0.450
SF-36 physical functioning	−8.87	−0.38	0.71	0.52	0.605
SF-36 bodily pain	−8.31	−0.37	0.88	0.68	0.609

* See Tables 2 and 4 for definitions.

Table 6. Frequency of PET Top 5 categories*

	Frequency	%
Do chores	204	42.5
Stand from chair	203	42.3
Dressing self	195	40.6
Get in/out of bed	163	34.0
Get down 5-lb bag	160	33.3
Open milk carton	148	30.8
Take a tub bath	147	30.6
Open jars previously opened	145	30.2
Shampoo hair	118	24.6
Climb up 5 steps	112	23.3
Walk outdoors on flat ground	110	22.9
Get in/out of car	106	22.1
Run errands and shop	92	19.2
Turn faucets on/off	85	17.7
Open car doors	83	17.3
Cut meat	83	17.3
Bend to pick up clothing	82	17.1
Lift glass to mouth	64	13.3
Wash and dry body	62	72.9
Get on/off toilet	38	7.9

* PET = Problem Elicitation Technique.

self," and the least frequently cited was "getting on/off toilet."

DISCUSSION

In this double-blind, placebo-controlled, multicenter trial comparing leflunomide with methotrexate and placebo in 482 patients with active RA, detailed analyses of disease-specific and generic measures demonstrated that leflunomide and methotrexate treatment significantly improve patients' function and health-related quality of life. The study was designed to include disease-specific and generic measures in an attempt to elucidate the similarities and differences of each in assessing function and health-related quality of life in patients with RA.

The M-HAQ, as a component of the ACR responder definition, was required; however, since it has been criticized as not being as sensitive to change as the full HAQ, the full HAQ was also administered. The PET was included because it best represents improvements that are important to the patient. A generic measure, the SF-36 (most notably, the physical functioning domain), was included to allow comparisons across other disease groups. In this way, the benefits realized through treatment of RA can be compared with those associated with treatment of other diseases, providing a general context for assessing the value of therapy. While the patient/physician global assessment is generally the most responsive of the measures, it is important to disaggregate this measure by examining comprehensive measures that reflect the key domains within physical, emotional, and social functioning.

Significant decrements in functional ability and health-related quality of life were apparent at baseline in this population of methotrexate-naive patients with active RA, similar to findings reported by Ruta et al and by Wells et al (12,13). Leflunomide and methotrexate administration resulted in substantive improvements in the HAQ disability index, M-HAQ, PET Top 5 score, SF-36 physical component score, and several scales of the SF-36 including physical functioning, bodily pain, general health profile, vitality, and social role. These changes are clearly important in view of the little-to-no improvement or deterioration observed in the placebo group.

The disease-specific instruments performed well. Close correlations between changes in the M-HAQ and HAQ indicated that the M-HAQ was similarly sensitive to change when administered monthly over time in this clinical trial. The PET questionnaire provided assurance that the changes seen in the HAQ disability index reflected improvement that was judged to be important to the individual patient. Although the PET instrument used in this trial was limited to only items present in the HAQ, this format reduced the time involved in completing the questionnaire and still showed substantial differences from one patient to another as to which disabilities are most important to them.

The generic health-related quality of life SF-36 questionnaire reflected treatment-induced benefits in pain and physical function, which were statistically significantly better than those seen with placebo and statistically different in a few scales between leflunomide and methotrexate. The physical component score incorporates physical functioning, role-physical, pain, and general health as well as all other domains queried in the SF-36. Thus, it measures much more than just functional decrements experienced by RA patients; it measures how these decrements affect a patient in their day-to-day activities. These correlated with similar changes in physical function (by M-HAQ, HAQ, and PET) and the ACR responder status. Improvement compared with placebo occurred in the 4 domains of the SF-36 (pain, vitality, social function, and physical function) that were shown by Ruta et al (12) as most responsive to change in their cohort of patients. These domains also correlated, on an individual patient basis, with ACR responses of $\geq$20%, $\geq$50%, and $\geq$70%.

Concurrent improvement in measures of health-related quality of life best correlated with the ACR response status at the $\geq$20% level. As previously dis-

cussed by Felson and colleagues, response rates of ≥50% and ≥70% discriminate less well from the effects of placebo (18). Although cutoff points other than the 20% used for the ACR 20% measure are currently being analyzed across different data sets to determine the most appropriate response levels, this report presents the performance at a 20% threshold for the various measures, since this is the level currently recommended and currently recognized by regulatory authorities.

While all scales of the SF-36 appear to demonstrate clinically meaningful and statistically significant differences when compared against placebo, in this trial they appeared to perform better in the leflunomide versus placebo comparison than in the methotrexate versus placebo comparison. Although, as previously reported (14,15), there were statistically significantly greater improvements in the M-HAQ, HAQ disability index, PET Top 5 score, and 2 of 8 scales of the SF-36 in the leflunomide group compared with the methotrexate treatment group, caution should be exercised in interpreting apparent differences between the performance profiles shown in Figures 1 and 2. These profiles are ratios rather than absolute numbers, and no consensus has yet been established as to what quantitative change reflects a "minimum clinically important difference," as opposed to being statistically significant. To see if this is a consistent treatment-related effect, this analysis should be repeated on other data sets of leflunomide- and methotrexate-treated patients. Despite some variation across treatment groups (which may or may not be clinically meaningful), an important conclusion here is that all of the health-related quality of life scales and the physical component score, with the exception of the mental health component score, performed sufficiently well that clinically important differences can be detected with statistical significance of less than 0.001.

One reason proposed as to why instruments may lack responsiveness has been the presence of ceiling effects, which is of particular concern in a generic instrument such as the SF-36 where questions are broad and may not capture specific impairment and/or disability due to RA. These data show that even when ~40% of patients have early disease (disease duration of ≤2 years, and/or are DMARD naive), sufficient levels of impairment/disability are reflected in baseline scores so that significant improvement can be demonstrated in patients receiving active treatment. This finding has other implications, in that it will allow comparisons of benefit across the disease states and also can be used in economic evaluations to compare the cost effectiveness of different treatments. This will facilitate stronger advocacy for appropriate resource allocations for disease treatment.

The present study is the second to include an assessment of the SF-36 measure, and the first to identify a treatment effect favoring active treatment compared with placebo. Although it is a generic measure, the SF-36 performed as well as the disease-specific measures, the HAQ and M-HAQ. This correlation requires further confirmation. Furthermore, very few RA studies to date have included a generic measure, so it is important to continue to include both disease-specific and generic measures to allow comparisons with existing studies and pooling for meta-analyses and comparisons across different disease populations.

In conclusion, since these data demonstrate that both disease-specific and generic measures of function and health-related quality of life can detect improvements in RA patients, both should be used in clinical studies evaluating therapies. This will assist in identifying changes that are discernible as well as important to patients' day-to-day life, and will facilitate comparisons across different disease states.

REFERENCES

1. Fries JF, Spitz PW, Young DY. The dimensions of health outcomes: the Health Assessment Questionnaire, disability and pain scales. J Rheumatol 1982;9:789–93.
2. Pincus T, Summey JA, Soraci SA Jr, Wallston KA, Hummon NP. Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment Questionnaire. Arthritis Rheum 1983;26:1346–53.
3. Bell MJ, Bombardier C, Tugwell P. Measurement of functional status, quality of life, and utility in rheumatoid arthritis. Arthritis Rheum 1990;33:591–601.
4. Tugwell P, Bombardier C, Gent M. Low dose cyclosporine versus placebo in patients with rheumatoid arthritis. Lancet 1990;335: 10051–6.
5. The HERA Study Group. A randomized trial of hydroxychloroquine in early rheumatoid arthritis: the HERA Study. Am J Med 1995;98:156–68.
6. Boers M, Verhoeven AC, Markusse HM, van de Laar MA, Westhovens R, van Denderen JC, et al. Randomised comparison of combined step-down prednisolone, methotrexate and sulphasalazine with sulphasalazine alone in early rheumatoid arthritis. Lancet 1997;350:309–18.
7. Buchbinder R, Bombardier C, Yeung M, Tugwell P. Which outcome measures should be used in rheumatoid arthritis clinical trials? Clinical and quality-of-life measures' responsiveness to treatment in a randomized controlled trial. Arthritis Rheum 1995;38:1568–80.
8. Felson DT, Anderson JJ, Boers M, Bombardier C, Furst D, Goldsmith C, et al. The American College of Rheumatology preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum 1995;38:727–35.
9. Bombardier C, Rabat J, and the Auranofin Cooperating Group. A comparison of health related quality of life measures for rheumatoid arthritis research. Control Clin Trials 1991;12:243S–56S.
10. Ware JE, Sherbourne CD. The MOS SF-36 item short-form health

survey (SF-36): conceptual framework and item selection. Med Care 1992;30:473–83.
11. Beaton DE, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. J Clin Epidemiol 1997;50:79–93.
12. Ruta DA, Hurst JP, Kind P, Hunter M, Stubbings A. Measuring health status in British patients with rheumatoid arthritis. Br J Rheumatol 1998;37:425–36.
13. Wells G, Boers M, Shea B, Tugwell P, Westhovens R, Suarez-Almazor M, et al. Sensitivity to change of generic quality of life instruments in patients with rheumatoid arthritis. J Rheumatol. 1999;26:217–21.
14. Strand V, Cohen S, Schiff M, Weaver A, Fleischmann R, Cannon GW, et al, on behalf of the Leflunomide RA Investigators Group. Treatment of active rheumatoid arthritis with leflunomide compared to placebo and methotrexate. Arch Intern Med. 1999;159:2542–50.
15. Strand V, Tugwell P, Bombardier C, Maetzel A, Crawford B, Dorrier C, et al, on behalf of the Leflunomide Rheumatoid Arthritis Investigators Group. Function and health-related quality of life: results from a randomized controlled trial of leflunomide versus methotrexate or placebo in patients with active rheumatoid arthritis. Arthritis Rheum 1999;42:1870–8.
16. Ware JE, Snow KK, Kosinski M, Gandek B. SF-36 health status survey manual. Boston: The Health Institute, New England Medical Center, 1993.
17. Tomlin GS, Holm MB, Rogers JC, Kwoh CK. Comparison of standard and alternative Health Assessment Questionnaire scoring procedures for documenting functional outcomes in patients with rheumatoid arthritis. J Rheumatol 1996;23:1524–30.
18. Felson DT, Anderson JJ, Lange MLM, Wells G, LaValley MP. Should improvement in rheumatoid arthritis clinical trials be defined as fifty percent or seventy percent improvement in core set measures, rather than twenty percent? Arthritis Rheum 1998;41: 1564–70.