Reprinted from:
The Journal of Rheumatology 1988;15:1833–40

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

NICHOLAS BELLAMY, W. WATSON BUCHANAN, CHARLES H. GOLDSMITH, JANE CAMPBELL, and LARRY W. STITT

Abstract. **Within the context of a double blind randomized controlled parallel trial of 2 nonsteroidal antiinflammatory drugs, we validated WOMAC, a new multidimensional, self-administered health status instrument for patients with osteoarthritis of the hip or knee. The pain, stiffness and physical function subscales fulfil conventional criteria for face, content and construct validity, reliability, responsiveness and relative efficiency. WOMAC is a disease-specific purpose built high performance instrument for evaluative research in osteoarthritis clinical trials.** (*J Rheumatol 1988;***15:1833–1840)**

Key Indexing Terms:
WOMAC OSTEOARTHRITIS VALIDITY
HEALTH STATUS INSTRUMENT RELIABILITY RESPONSIVENESS

We reported on the inadequacy of outcome measurement procedures in osteoarthritis (OA) trials of nonsteroidal antiinflammatory (NSAID) drugs[1]. In an attempt to rationalize measurement in OA, we first probed the symptomatology of hip and knee OA by interviewing 100 patients with OA, and identified 41 items on 5 dimensions (Table 1) which characterize the disorder[2,3]. To validate these items, we have now conducted a double blind, randomized, controlled parallel design trial of isoxicam versus piroxicam in elderly patients with primary OA of the hip or knee. Our goal was to assess the reliability, construct validity and responsiveness of the 41 items previously mentioned. The 2 separate arms of the trial served as independent tests of item responsiveness. To circumvent the problem of simultaneously attempting to assess the performance of a new health status instrument as well as a new antirheumatic compound, we used other reported[4] outcome measures to evaluate the comparative efficacy and tolerability of isoxicam and piroxicam. Thus, using this innovative approach, we defined the clinimetric properties of a health status instrument termed ''WOMAC'' (the Western Ontario and McMaster Universities Osteoarthritis Index), within the context of a traditional clinical trial.

MATERIALS AND METHODS

Fifty-seven patients with symptomatic OA of the hip or knee requiring NSAID therapy were entered in the study. To be eligible patients had to be 55–85 years of age, have definite radiographic evidence of primary OA in the hip or knee, and fulfil defined inclusion and exclusion criteria[4].

Patients were assessed at enrolment (Visit 1), and again one week later without any change in therapy (Visit 2). Thereafter, patients underwent a one-week NSAID-free washout period and were then reassessed (Visit 3). Finally, patients were evaluated following 2 (Visit 4), 4 (Visit 5), and 6 (Visit 6) weeks of active treatment. The initial drug dosage was piroxicam 10 mg OD or isoxicam 100 mg OD, this being increased at Visit 4 to 20 mg OD or 200 mg OD, respectively, in patients failing to respond to the lower dosage[4].

The primary outcome measures employed were the WOMAC OA Index (Test Form, Table 1) and 2 forms of global assessment. WOMAC was self-administered while the global assessments (on each of the 5 dimensions) were made both by trained interviewers (interviewer global assessment) and study patients (patient global assessment). To address issues relating to scaling, patients were given (in random sequence) 2 versions of WOMAC to complete. Both contained identical questions but one required responses on 5-point (none, slight, moderate, severe, extreme) Likert scales[5] while the other required responses on 10 cm horizontal visual analogue scales (VAS) with terminal descriptors[6]. Individual item scores were determined by reading the patient's response to each question. Aggregate scores for each dimension were determined by summing the component item scores for each dimension. The WOMAC final battery was determined by summing the aggregate scores for the pain, stiffness and physical function dimensions. For reasons of feasibility, only the first 3 pain and 7 physical questions were duplicated on both scales (Table 1). Similarly, only the question pertaining to severity of morning stiffness, the 1st, 2nd and 6th social, and the 1st, 2nd, 4th and 5th emotional questions were duplicated on the VAS

From the University of Western Ontario, London, and McMaster University, Hamilton, ON, Canada.

Supported in part by a grant from the Arthritis Society of Canada and in part by a grant from Parke-Davis Canada Inc.

N. Bellamy, MD, MSc, FRCP(Edin, Glas, and C), FACP, Associate Professor of Medicine, Epidemiology and Biostatistics, University of Western Ontario; W.W. Buchanan, MD, FRCP(Edin, Glas, and C), Professor of Medicine; C.H. Goldsmith, PhD, Professor of Clinical Epidemiology and Biostatistics, McMaster University; J. Campbell, BA, Research Assistant, University of Western Ontario; L.W. Stitt, MSc, Statistical Consultant/Analyst and Lecturer, Department of Statistical and Actuarial Sciences, University of Western Ontario.

Address requests for reprints to Dr. N. Bellamy, Ste. 402A, W.T., Victoria Hospital, 800 Commissioners Rd. E., London, ON, Canada, N6A 4G5.

Submitted August 19, 1988 revision accepted September 21, 1988.

scale (Table 1). As with WOMAC, the interviewer and patient global assessment scores on single questions which separately probed the overall status of the patient on each of the 5 dimensions were made on both Likert and VAS scales. Patients completed WOMAC, interviewer and patient global assessments at all 6 visits. To test the construct validity of WOMAC, the following secondary outcome measures were concurrently applied: (1) joint tenderness (modified Doyle Index [hip and knee only])[7], (2) Lequesne Index[8], (3) Bradburn Index of Well Being[9], and (4) social component of the McMaster Health Index Questionnaire (MHIQ)[10]. These measures were selected as being capable of validating the 5 different WOMAC dimensions i.e., Pain (Doyle, Lequesne-Pain), Stiffness (Lequesne-Stiffness), Physical Function (Lequesne-Physical Function), Emotional Function (Bradburn), and Social Function (MHIQ-Social). The Doyle and Lequesne indices were selected since they were developed specifically for patients with OA. The Bradburn and MHIQ indices were selected because of our familiarity with them. Finally, 3 tertiary outcome measures were used: 50′ walking time, total range of movement (ROM), intermalleolar straddle. These commonly used measures of drug efficacy were selected to assess the relative efficiency of the final WOMAC battery against traditional measures, and not as supplementary measures required for validation purposes. We have not, therefore, reported statistical p values for these variables but used the data to calculate the relative efficiency of WOMAC. Individual item and aggregate item data were analyzed for each separate WOMAC dimension using both Student's t test[11] and Wilcoxon's nonparametric test[12] to assess item and dimension responsiveness (Visit 6 vs Visit 3) and the effect of parametric versus nonparametric statistical treatment of the data. Internal consistency (Visit 3) was tested using Cronbach's alpha[13], test-retest reliability (Visit 1 vs Visit 2) using Kendall's tau c statistic[14], and construct validity (Visit 3) determined using Pearson's correlation coefficient[15]. Relative efficiency was calculated (Visit 6 vs Visit 3) using the method employed by Liang, *et al:* [16] e.g., relative efficiency for WOMAC vs Walktime (WT) = $(t_{WOMAC}/t_{WT})^2$. We have not reported response data for Visit 4 (as this represented a titration step) or for Visit 5 (as this was used to assess tolerability after incremental dosing at Visit 4). However, data on these visits can be found in the paper reporting drug efficacy[4].

RESULTS

Fifty-seven patients were enrolled in the study: 28 (14 males, 14 females) received isoxicam and 29 (12 males, 17 females) received piroxicam. The mean age was 66.5 years in each group (varying from 55 to 82). The mean disease duration (i.e., symptomatology) was 8.7 years (varying from 1 to 30) in the piroxicam group and 9.3 years (varying from 2 to 26) in the isoxicam group. The knee was selected as the most severely affected joint in 39 patients (isoxicam 21, piroxicam 18) compared to the hip in 18 patients (isoxicam 7, piroxicam 11). The above differences between the 2 groups were not statistically significant. The means and standard deviations (Visit 3) for primary, secondary, and tertiary outcome measures are illustrated in Tables 2 and 3.

Pain — Responsiveness (Table 1)

Isoxicam. On Likert scaling using Wilcoxon's test, all 5 items significantly improved by Visit 6 ($p \leq 0.019$), while on VAS scaling all items achieved p values of ≤ 0.001. With the interviewer and patient global assessments and aggregate score strategies, p values of ≤ 0.001 were achieved regardless of scale (Likert vs VAS) or type of analysis (Student's t test vs Wilcoxon). When the p values derived by parametric and nonparametric analysis were compared for all 13 analyses performed using individual item = 8, aggregate

Table 1. *Summary of item content of original test form of WOMAC***

Pain†		
1	Walking	(2.58)*
2	Stair climbing	(2.62)*
3	Nocturnal	(2.63)*
4	Rest	(2.57)
5	Weight bearing	(2.51)
Stiffness†		
1	Morning stiffness	(2.52)*
2	Stiffness occurring later in the day	(2.30)
Physical Function†		
1	Descending stairs	(2.60)*
2	Ascending stairs	(2.54)*
3	Rising from sitting	(2.32)*
4	Standing	(2.64)*
5	Bending to floor	(2.51)*
6	Walking on flat	(2.40)*
7	Getting in/out car	(2.26)*
8	Going shopping	(2.40)
9	Putting on socks	(2.38)
10	Rising from bed	(2.37)
11	Taking off socks	(2.37)
12	Lying in bed	(2.36)
13	Getting in/out bath	(2.30)
14	Sitting	(2.54)
15	Getting on/off toilet	(2.67)
16	Heavy domestic duties	(2.43)
17	Light domestic duties	(2.26)
Social Function		
1	Leisure activities	(2.56)*
2	Community events	(2.15)*
3	Church attendance	(2.52)
4	With spouse	(2.65)
5	With family	(2.67)
6	With friends	(2.64)*
7	With others	(2.55)
Emotional Function		
1	Anxiety	(2.64)*
2	Irritability	(2.59)*
3	Frustration	(2.44)
4	Depression	(2.49)*
5	Relaxation	(2.39)*
6	Insomnia	(2.58)
7	Boredom	(2.62)
8	Loneliness	(2.26)
9	Stress	(2.19)
10	Wellbeing	(2.62)

* These items were duplicated on VAS scales

** These item numbers correspond to those in text and Table 4.

† Dimensions retained in final WOMAC instrument.

( ) Numbers in parentheses represent previously published[2] mean importance scores for each item. (Scale: 0=none, 1=slight, 2=moderate, 3=very, 4=extreme importance.)

Table 2. *Primary outcome measures: Visit 3 means (m) and standard deviations (s)*

Variable			Pain		Stiffness		Physical Function		Social Function		Emotional Function	
			Likert*	VAS†	Likert*	VAS†	Likert*	VAS†	Likert*	VAS†	Likert*	VAS†
WOMAC††	C	m	10.3	158.9	4.4	51.7	32.2	342.4	5.9	100.5	10.1	104.8
		s	4.4	69.0	1.8	28.5	13.8	154.1	5.5	75.8	8.5	90.1
	I	m	9.6	153.1	4.3	52.9	31.0	334.4	6.0	102.7	10.8	112.5
		s	4.0	66.3	1.6	29.7	13.5	152.6	5.2	79.5	9.6	107.3
	P	m	10.9	165.4	4.6	50.5	33.4	350.4	5.9	98.3	9.5	97.0
		s	4.7	72.6	1.9	27.7	14.3	157.9	6.0	73.3	7.5	69.3
Patient global assessment	C	m	2.5	55.3	2.4	53.3	1.9	42.1	1.1	27.2	0.7	21.0
		s	1.0	28.2	1.0	29.0	1.0	26.4	1.0	27.1	0.9	22.6
	I	m	2.3	51.7	2.3	50.1	1.8	40.1	1.1	28.5	0.8	22.8
		s	1.1	29.8	1.1	30.9	1.0	25.9	1.0	29.0	1.0	26.4
	P	m	2.6	58.6	2.4	56.1	2.0	43.8	1.1	25.9	0.6	19.4
		s	1.0	26.7	1.0	27.4	0.9	27.2	1.0	25.7	0.7	18.9
Interviewer global assessment	C	m	2.6	—**	2.1	—	2.3	···	1.1		1.2	–
		s	0.9	—	1.0	—	0.8	--	1.1	—	1.0	·
	I	m	2.5	—	2.2	–	2.1	—	1.1	–	1.2	—
		s	0.9	—	1.0	·	0.8		1.1	· ·	1.1	—
	P	m	2.7	—	2.0	···	2.4	—	1.0	–	1.1	—
		s	0.8	—	0.9	–	0.9	--	1.0	·	0.8	—

* Scored on values 0-4, where 0=none, 1=slight, 2=moderate, 3=very, 4=extreme.

† 0–100 mm horizontal VAS scale with terminal descriptors None and Extreme

†† Sum of WOMAC test questionnaire items: Aggregate score (pain), AI (stiffness), AI (physical), AI (social), AI (emotional).

** IgA only scored on Likert scale, *not* on VAS scale. Interviewer global assessment

C - Combined group (isoxicam + piroxicam), I - isoxicam, P - piroxicam.

Table 3. *Secondary and tertiary outcome measures: Visit 3 means (m) and standard deviations (s)*

	Combined		Isoxicam		Piroxicam	
Secondary	m	s	m	s	m	s
Bradburn total score	−3.6	3.3	−3.0	3.3	−4.2	3.2
Modified Doyle total score	2.8	1.6	2.5	1.5	3.0	1.6
Lequesne pain score	4.4	1.1	4.5	1.1	4.3	1.1
Lequesne stiffness score	1.5	0.5	1.5	0.6	1.4	0.5
Lequesne physical score	5.7	2.3	5.6	2.5	5.9	2.0
MHIQ social score	15.8	1.7	15.6	1.9	16.0	1.5
	Combined		Isoxicam		Piroxicam	
Tertiary	m	s	m	s	m	s
Walk time (s)	17.2	7.2	17.1	8.3	17.3	6.0
Intermalleolar straddle	79.2	18.2	80.6	17.8	77.7	18.8
ROM (°)	225.1	26.3	224.8	26.9	225.5	26.1

score = 2, interviewer global assessment = 1, and patient global assessment = 2 strategies, there was exact agreement (to 3 decimal places) in 46% of the cases, while in 54% the parametric value was smaller. The correlation coefficient between scores on Likert and VAS scales was 0.82 for patient global assessment.

Piroxicam. Regardless of the type of statistical analysis used, 80% of the items significantly improved on Likert scaling by Visit 6 ($p \leq 0.005$), but #2 failed to significantly improve. On VAS scaling using Wilcoxon's test, all 3 items achieved p values of ≤ 0.019. With respect to the interviewer and patient global assessments and aggregate score strategies, p values of ≤ 0.003 were achieved regardless of scale. When all 13 comparative analyses (individual item = 8, aggregate score = 2, interviewer global assessment = 1, patient global assessment = 2) were considered, the parametric p values were smaller in 77% of the cases and larger in 15%, while in 8% there was exact agreement. The correlation coefficient between scores on Likert and VAS scales was 0.86 for patient global assessment.

Pain — reliability. From Likert scaled responses to the 5 component items the internal consistency of the pain dimension was 0.86 for isoxicam and 0.89 for piroxicam. The corresponding values for the 3 VAS scaled responses were 0.81 and 0.73, respectively. The test-retest reliability for the combined group (i.e., isoxicam + piroxicam) was 0.68 on the Likert scale and 0.64 on the VAS scale.

Pain — validity. Higher levels of correlation (as expressed by the correlation coefficients and the proportion of items displaying a statistically significant correlation) were noted on both Likert and VAS responses between the test items and the Lequesne pain and physical function components and the Doyle Index, than between these same items and the Lequesne stiffness component, the Bradburn Index and the MHIQ social component (Table 4).

Stiffness — Responsiveness (Table 1)

Isoxicam. On Likert scaling using Wilcoxon's test, both items (morning stiffness, stiffness occurring later in the day) significantly improved by Visit 6 ($p \leq 0.004$), while on VAS scaling morning stiffness achieved a p value of ≤ 0.001 regardless of type of analysis used. With the interviewer global assessment, patient global assessment, and aggregate score strategies, p values of ≤ 0.001 were achieved regardless of scale or type of analysis. When all 7 comparative analyses (individual item = 3, aggregate score = 1, interviewer global assessment = 1, patient global assessment = 2) were considered, the p values showed exact agreement in 43% of the cases, while in 57% of the cases the parametric value was smaller. The correlation coefficient between scores on Likert and VAS scales was 0.91 for patient global assessment.

Piroxicam. On Likert scaling using Wilcoxon's test, both items significantly improved by Visit 6 ($p \leq 0.030$), while on VAS scaling, morning stiffness achieved a p value of 0.002. With respect to the interviewer and patient global assessments and aggregate score strategies, significant improvement was detected on each ($p \leq 0.013$). However, p values were smaller for patient global assessment on VAS scaling than on Likert scaling. When all 7 comparative analyses (individual item = 3, aggregate score = 1, interviewer global assessment = 1, patient global assessment = 2) were considered the parametric p value was smaller in 100% of cases. The correlation coefficient between scores on Likert and VAS scales was 0.87 for patient global assessment.

Stiffness — reliability. From Likert scaled responses to the 2 component items, the internal consistency of the stiffness dimension was 0.90 for isoxicam and 0.91 for piroxicam. Only morning stiffness was probed on the VAS scale, and, consequently, interitem reliability was not determined. Test-retest reliability for the combined group was 0.48 on the Likert scale and 0.61 on the VAS.

Stiffness — validity. The highest levels of correlation noted on both Likert and VAS scaled responses were between the 2 test items and the Doyle Index (morning stiffness $r = 0.45$, late day stiffness $r = 0.46$) and the Lequesne pain, physical function and stiffness components (morning stiffness $r = 0.22$, late day stiffness $r = 0.23$). No significant correlation was noted between the test items and the other scales (Table 4).

Physical function — Responsiveness (Table 1)

Isoxicam. On Likert scaling using Wilcoxon's test, 15 of the 17 physical function items significantly improved by Visit 6 ($p \leq 0.001$ for 4 items, $0.002 \leq p \leq 0.005$ for 8 items, and $0.008 \leq p \leq 0.009$ for 3 items). Items 10 and 14 achieved p values of 0.057 and 0.059, respectively, at Visit 6. On VAS scaling significant improvement occurred on all items, the p value being ≤ 0.001 regardless of type of analysis used. The interviewer and patient global assessments and aggregate score strategies detected significant improvements ($p < 0.003$). When p values derived by parametric and nonparametric tests were compared for all 29 analyses (individual item = 24, aggregate score = 2, interviewer global assessment = 1, patient global assessment = 2) performed, the parametric value was smaller in 62% of the cases, while in 38% of the cases there was exact agreement. The correlation coefficient between scores on Likert and VAS scales was 0.83 for patient global assessment.

Piroxicam. On Likert scaling using Wilcoxon's test, 12 of the 17 physical function items significantly improved by Visit 6 ($0.002 \leq p \leq 0.009$ for 6 items, $0.013 \leq p \leq 0.019$ for 4 items, and $0.024 \leq p \leq 0.027$ for 2 items). On VAS scaling significant improvement occurred on all items, the p value being ≤ 0.001 for 3 items, 0.006 for one item, $0.010 \leq p \leq 0.011$ for 2 items, and 0.044 for one item. With the interviewer and patient global assessments and aggregate score strategies, significant improvement was detected by Visit 6 ($p \leq 0.002$ for aggregate score and interviewer global assessment; $0.004 \leq p \leq 0.010$ for patient global assessment). When p values derived by parametric and nonparametric tests were compared for all 29 analyses performed using individual item = 24, aggregate score = 2, interviewer global assessment = 1, and patient global assessment = 2 strategies, the parametric p value was smaller in 79% of the cases, larger in 7%, while in 14% there was exact agreement. The correlation coefficient between scores on Likert and VAS scales was 0.90 for patient global assessment.

Physical function — reliability. From Likert scaled responses to the 17 component items, the internal consistency of the physical function dimension was 0.95 for isoxicam and 0.95 for piroxicam. The corresponding values for the 7 VAS scaled responses were 0.91 and 0.89, respectively. Test-retest reliability for the combined group was 0.68 on the Likert scale and 0.72 on the VAS scale.

Physical function — validity. Higher levels of correlation were noted on both Likert and VAS scaled responses between

Table 4. *Construct validity analysis: Correlation of WOMAC test items with Lequesne, Modified Doyle, Bradburn, and MHIQ indices*

Domain		Lequesne Pain	Lequesne Stiffness	Lequesne Physical	Doyle Tenderness	Bradburn Emotional	MHIQ Social
Pain							
Likert	1	(0.46/0.57	(0.14/0.35)	(0.30/0.55)	(0.25/0.46)	(−0.06/0.15)	(−0.16/−0.00)
(n=5)	2	1–5	1,4,5	1–5	1,3–5	—	—
	3	100	60	100	80	0	0
VAS	1	(0.39/0.62)	(0.04/0.24)	(0.36/0.50)	(0.36/0.57)	(−0.08/0.04)	(−0.07/0.05)
(n=3)	2	1–3	—	1–3	1–3	—	—
	3	100	0	100	100	0	0
Stiffness							
Likert	1	(0.32/0.45)	(0.22/0.23)	(0.29/0.32)	(0.45/0.46)	(−0.22/−0.09)	(−0.13/−0.08)
(n=2)	2	AMS[4], GEL[5]	—	AMS,GEL	AMS,GEL	—	—
	3	100	0	100	100	0	0
VAS	1	(0.32)	(0.27)	(0.35)	(0.47)	(−0.21)	(−0.11)
(n=1)	2	AMS	AMS	AMS	AMS	—	—
	3	100	100	100	100	0	0
Physical Function							
Likert	1	(0.15/0.51)	(−0.04/0.33)	(0.20/0.54)	(0.14/0.52)	(−0.14/0.24)	(−0.21/0.15)
(n=17)	2	1–4,6,8, 10,12–17	7,15,16	3–17	3,4,6–17	—	—
	3	77	18	88	82	0	0
VAS	1	(0.32/0.50)	(0.01/0.31)	(0.36/0.59)	(0.28/0.54)	(−0.14/0.22)	(−0.31/−0.00)
(n=7)	2	1–7	6	1–7	1–7	–	5
	3	100	14	100	100	0	14
Social Function							
Likert	1	(0.21/0.35)	(0.17/0.34)	(0.24/0.37)	(0.09/0.35)	(−0.03/0.29)	(−0.12/0.11)
(n=7)	2	1,2,5–7	2	1,2,4–7	1–3	—	—
	3	71	14	86	43	0	0
VAS	1	(0.28/0.35)	(0.22/0.37)	(0.42/0.49)	(0.36/0.46)	(−0.14/0.09)	(−0.06/0.05)
(n=3)	2	1–3	1,2	1–3	1–3	—	—
	3	100	67	100	100	0	0
Emotional Function							
Likert	1	(0.15/0.35)	(0.03/0.37)	(0.18/0.46)	(−0.04/0.20)	(0.14/0.45)	(−0.30/−0.04)
(n=10)	2	4–6,10	2,5	1–6,9,10	—	1,7–9	—
	3	40	20	80	0	40	0
VAS	1	(0.30/0.34)	(0.20/0.36)	(0.44/0.54)	(0.14/0.23)	(0.34/0.38)	(−0.25/−0.18)
(n=4)	2	1–4	4	1–4	—	1–4	—
	3	100	25	100	0	100	0

1 Min/max of Pearson correlation coefficients between individual test items and comparison indices.
2 Test item number showing statistically significant correlation with comparison indices ($p \leq 0.05$).
3 Percentage of test items showing statistically significant correlation with comparison indices.
4 AMS – stiffness after first wakening in the morning.
5 GEL – stiffness after sitting, lying, or resting later in the day.
NB – Test item numbers correspond to those identified in Table 1.

the test items and the physical component of the Lequesne Index than with these same items and the Doyle Index, the Lequesne pain and stiffness components, the Bradburn Index, and the MHIQ social component (Table 4).

Social function — Responsiveness (Table 1)

Isoxicam. On Likert scaling using Wilcoxon's test, 4 of the items achieved statistically significant improvement at Visit 6 (p = 0.011, 0.018, 0.033, and 0.043). On VAS scaling p values of 0.001 were achieved by the first 2 items, while a p value of 0.023 was attained by the 3rd item. Although the patient global assessment strategy resulted in a p value of 0.020 on VAS scaling, it did not statistically improve on

Likert scaling (p = 0.110). The aggregate score and interviewer global assessment strategies resulted in p values of ≤ 0.001 and 0.015, respectively. When all 15 comparative analyses (individual item = 10, aggregate score = 2, interviewer global assessment = 1, patient global assessment = 2) were considered, there was absolute agreement of the p value in 13% of the cases, while in 67% the parametric p values were smaller and in 20% the nonparametric p values were smaller. The correlation coefficient between scores on Likert and VAS scales was 0.87 for patient global assessment.

Piroxicam. On Likert scaling none of the items improved significantly by Visit 6 regardless of type of analysis. However, on VAS scaling item one achieved a p value of 0.002, while item 2 achieved p values of 0.010 (t test) and 0.006 (Wilcoxon). Neither the interviewer global assessment nor patient global assessment strategies detected significant improvement regardless of scale or type of analysis. However, the aggregate score strategy resulted in improvement on the VAS ($p \leq 0.008$). When all 15 comparative scale analyses (individual item = 10, aggregate score = 2, interviewer global assessment = 1, patient global assessment = 2) were considered there was exact agreement of the p value in 7% of the cases, while in 73% the parametric p values were smaller and in 20% the nonparametric p values were smaller. The correlation coefficient between scores on Likert and VAS scales was 0.62 for patient global assessment.

Social function — reliability. From Likert scaled responses to the 7 component items, the internal consistency of the social function dimension was 0.89 for isoxicam and 0.93 for piroxicam. Corresponding values for the 3 VAS scaled items were 0.89 and 0.93, respectively. Test-retest reliability for the combined groups was 0.61 on the Likert scale, and 0.59 on the VAS scale.

Social function — validity. Higher levels of correlation were noted between the test items and the Lequesne physical function and pain components (on Likert scaled responses) than with the Doyle Index or Lequesne stiffness component (Table 4). With VAS scaled responses, higher levels of correlation were noted between test items and the Lequesne physical function component, the Doyle Index, and the Lequesne pain component than with these same items and the Lequesne stiffness component. Regardless of scale, no significant correlation was noted between the test items and the Bradburn Index or with the MHIQ social component.

Emotional function — Responsiveness (Table 1)

Isoxicam. On Likert scaling using Wilcoxon's test, half of the items improved significantly by Visit 6 ($p \leq 0.043$). In contrast, all VAS scaled responses achieved p values of ≤ 0.014. Both the aggregate score and interviewer global assessment strategies showed significant improvement on Likert scaling ($p \leq 0.004$). However, the patient global assessment strategy did not detect improvement on either scale ($p \geq 0.090$). When all 19 comparative analyses (individual item = 14, aggregate score = 2, interviewer global assessment = 1, patient global assessment = 2) were considered, there was exact agreement of the p value in 10% of the cases, while in 74% the parametric p values were smaller, and in 16% the nonparametric p values were smaller. The correlation coefficient between scores on Likert and VAS scales was 0.91 for patient global assessment.

Piroxicam. On Likert scaling using Wilcoxon's test, 4 of the 10 items improved significantly by Visit 6 ($p \leq 0.050$). In contrast, the 4 VAS scaled responses all achieved p values ≤ 0.032. Although both the interviewer global assessment and aggregate score strategies demonstrated significant improvement on Likert scaling (p = 0.004 and p = 0.022, respectively), the patient global assessment strategy did not detect improvement (Likert p = 0.779, VAS p = 0.187). When all 19 comparative analyses (individual item = 14, aggregate score = 2, interviewer global assessment = 1, patient global assessment = 2) were considered, there was exact agreement of the p value in 8% of the cases while in 92% the parametric p values were smaller on Likert scaling. On VAS scaling, however, the nonparametric p values were smaller in 100% of cases. The correlation coefficient between scores on Likert and VAS scales was 0.66 for patient global assessment.

Emotional function — reliability. From Likert scaled responses to the 10 component questions, the internal consistency of the emotional function dimension was 0.96 for isoxicam and 0.91 for piroxicam. The corresponding values for the 4 VAS scaled items were 0.98 and 0.88, respectively. The test-retest reliability for the combined group was 0.72 on the Likert scale and 0.66 on the VAS scale.

Emotional function — validity. Higher levels of correlation were noted on both Likert and VAS scaled responses between the test items and the Lequesne physical function component, the Bradburn Index, and the Lequesne pain component than with the Lequesne stiffness component (Table 4). No significant correlation was noted between test items and the MHIQ social component or the Doyle Index.

Relative efficiency. When considering both treatment groups combined, 5 pain, 2 stiffness and 17 physical function items achieved statistical significance ($p \leq 0.005$) by Visit 6. Since only 3 emotional items and none of the social items achieved this level of significance, emotional and social dimensions were not subjected to relative efficiency testing (Table 5). In 83% of analyses the relative efficiency of WOMAC was > 1, i.e., more efficient than the tertiary measures. Relative efficiency values < 1 were largely accounted for by walking time scores for the piroxicam group. In 78% of comparisons the relative efficiency for VAS scaled responses was numerically greater than the corresponding Likert scaled responses.

Table 5. *Relative efficiency* of WOMAC versus tertiary outcome variables*

Tertiary Outcome Variable	Study Group	Pain		Stiffness		Physical Function		WOMAC Final Battery (FB)**	
		Likert	VAS	Likert	VAS	Likert	VAS	Likert	VAS
Walk time	C	1.4	1.5	0.7	1.4	1.2	1.3	1.4	1.5
	I	2.3	2.9	2 0	2.2	2.9	2.4	3.1	2.8
	P	0.8	0.7	0.3	0.9	0.5	0.7	0.6	0.8
Intermalleolar straddle	C	5.2	5.8	2.8	5.5	4.8	5.1	5.4	6.0
	I	2.7	3.4	2.4	2.6	3.4	2.9	3.6	3.3
	P	11.8	11.0	4.5	13.4	7.7	10.2	9.4	12.5
ROM	C	1.4	1.6	0.8	1.5	1.3	1.4	1.4	1.7
	I	1.3	1.6	1.1	1.2	1.6	1.4	1.7	1.6
	P	1.6	1.6	0.7	1.9	1.0	1.6	1.3	1.9

* Relative efficiency = $(t_1/t_2)^2$ e.g., for isoxicam $(t_{\text{pain (VAS)}}/t_{\text{walk time}})^2=2.9$.

** WOMAC (FB) = $AI_{\text{pain}} + AI_{\text{stiffness}} + AI_{\text{physical function}}$.

C−Combined group (isoxicam + piroxicam), I-isoxicam, P-piroxicam

DISCUSSION

In developing a new health status measure we were guided by 4 principles: adequate responsiveness, reliability and validity, and superior relative efficiency over selected traditional measures. We elected to employ a double blind, randomized, controlled parallel design since both groups of patients at Visit 3 would have a high probability of being similar with respect to their pretreatment status and response potential. Furthermore, if the 2 agents are similar in efficacy then the 2 arms of the study may be used for conducting separate tests of index responsiveness in 2 clinically equivalent groups of patients. Indeed, since no significant between-group differences were detected (Tables 2 and 3) and no significant between-drug differences were identified using the reported independent outcome measures[4], and accepting the possibility of a Type II error, nevertheless we regard the design as a legitimate and novel approach to index validation.

Responsiveness. Although isoxicam was voluntarily suspended worldwide by Warner-Lambert International in October 1985, this was not for lack of efficacy but rather for reasons of toxicity apparently related to a manufacturing problem in France[17]. Since drug efficacy was not at issue, we regard this voluntary suspension as irrelevant to the validation of WOMAC. Twenty-seven (5 pain, 2 stiffness, 17 physical, 0 social, 3 emotional) of the original 41 WOMAC items achieved statistical significance with p values ≤ 0.005 by Visit 6 for the combined group. The use of multiple analytic comparisons may result in an increase in Type I errors[18]. Even correcting for this statistical nuance, however, and accepting a high degree of covariance among Index items, the p values attained were extremely good and indicative of a high level of responsiveness for these 27 WOMAC items. Comparative analyses of nonparametric vs parametric treatment of the data suggest that while in many instances there is agreement between the 2 (and therefore that either analysis may be used), nonparametric methods may provide a more conservative estimate of the response, and for conceptual reasons relating to normality of the data, may be regarded as the preferred analytic technique. Since these observations are of limited generalizability, we are continuing to perform both parametric and nonparametric comparisons on instrument data.

Reliability. Reliability coefficients (i.e., Cronbach's alpha) of ≥ 0.80 are generally regarded as acceptable. Index items exceeded 0.85 on both VAS and Likert scales in all but one instance (pain — VAS piroxicam group = 0.73). The values achieved for test-retest reliability were somewhat lower than those for internal consistency. Nevertheless, we regard them as entirely adequate considering that (a) the test-retest interval was one week, and (b) the Kendall's tau c statistic tends to generate slightly lower coefficients of correlation[19]. We believe that the principal explanation for our lower test-retest values lies in the excessive interval (1 week) between the 2 administrations. Indeed, given the high internal consistency and sensitivity of WOMAC, and considering the constantly fluctuating symptomatology of OA, one can predict that test-retest reliability values will only be moderate. These data indicate, therefore, that all 5 WOMAC dimensions on both VAS and Likert scales are of adequate reliability.

Validity. Opinions differ as to which items should be incorporated in outcome measurement, and which numerical weights assigned to the clinical importance of different items[20]. We believe, however, that the item content of WOMAC should be generally acceptable since it is based not only on a review of both the clinimetric and OA literatures[1], but on the opinions of 100 patients with symptomatic OA who provided data on the dimensionality of their symptoms and assigned importance scores for each item subsequently used in constructing WOMAC[2]. For criterion validity testing, coefficients ≥ 0.80 are generally regarded as acceptable. However, no irrefutable gold standard currently exists against which to test criterion validity. We have, therefore, tested construct validity against other indices which probe the 5 Index dimensions of interest. Since these comparators are not gold standards, lower levels of correlation are expected. In general, however, the Index items should show a statistically significant correlation with other indices probing the same dimension (convergent construct validity). Furthermore, Index items should also show higher levels of correlation with other indices probing the same dimension than with indices probing other (particularly unrelated) dimensions (divergent construct validity). These criteria were fulfilled by the pain, stiffness and physical function components of the Index. It should be noted that since physical disability is often secondary to pain, it is not surprising that these 2 dimensions are often associated. We observed that both stiffness items showed a better correlation with the modified Doyle score than with the Lequesne stiffness component. However, we believe this to be due to the fact that the Lequesne Index probes duration of stiffness while WOMAC probes its severity. Given the interrelationship between discomfort and disability, an association between stiffness, pain, tenderness and physical function is predictable. Of note, the VAS scaled stiffness item showed a statistically significant correlation with the Lequesne stiffness component. The social component of WOMAC failed to correlate with the MHIQ social component, and although some items were reliable and responsive, this dimension was excluded from the Index. Moreover, in spite of the emotional component fulfilling construct validity criteria, and most items being reliable and responsive, we elected to withdraw the component pending a reevaluation of the social dimension. The final Index, therefore, utilizes the pain (5 items), stiffness (2 items), and physical (17 items) function subscales only (Table 1).

Relative efficiency. To be useful, a new health status measure should offer advantages over existing indices. In this respect, WOMAC offers 2 advantages. First, WOMAC and its subscales offer superior efficiency (as measured by relative efficiency scores) over selected traditional measures in assessing the efficacy of antirheumatic drugs. Such a measure, therefore, has potential for reducing sample size requirements for clinical trials using WOMAC as the primary outcome measure. Secondly, traditional measures often lack patient relevance. In contrast, WOMAC probes patient relevant outcomes, the clinical importance (Table 1) of which have been documented[2].

We believe WOMAC to be a reliable, valid, and responsive multidimensional, self-administered outcome measure designed specifically to evaluate patients with OA of the hip or knee. We are currently conducting further studies on aggregating scores across different dimensions, on the relative responsiveness of Likert and VAS scales, and the relative efficiency of WOMAC against several other health status instruments.

ACKNOWLEDGMENT

We wish to thank the following for their contribution to the development of the WOMAC: S. O'Leary, E. Grace and E. Duku.

REFERENCES

1. Bellamy N, Buchanan WW: Outcome measurement in osteoarthritis clinical trials: The case for standardization. *Clin Rheum 1984;*3:293–305.
2. Bellamy N, Buchanan WW: A preliminary evaluation of the dimensionality and clinical importance of pain and disability in osteoarthritis of the hip and knee. *Clin Rheum 1986;*5:231–41.
3. Bellamy N, Buchanan WW: A health status instrument for osteoarthritis of hip and knee: Stage 2 analysis of a self-administered questionnaire (abstr). In: *Abstracts: Australian Rheumatism Association Annual Scientific Meeting,* 1987:32.
4. Bellamy N, Buchanan WW, Grace E: Double-blind randomized controlled trial of isoxicam vs piroxicam in elderly patients with osteoarthritis of the hip and knee. *Br J Clin Pharmacol 1986;*22:149S–55S.
5. Likert R: A technique for measurement of attitudes. *Arch Psychol 1932;*140:44–60.
6. Huskisson EC: Measurement of pain. *J Rheumatol 1982;*9:768–9.
7. Doyle DV, Dieppe PA, Scott J, *et al:* An articular index for the assessment of osteoarthritis. *Ann Rheum Dis 1981;*40:75–8.
8. Lequesne M: European guidelines for clinical trials of new antirheumatic drugs. *EULAR Bull 1980;*(suppl 6)9:171–5.
9. Bradburn NM: *The Structure of Psychological Well-Being.* Chicago: Aldine Publishing, 1969.
10. Chambers LW: *McMaster Health Index Questionnaire (MHIQ).* Hamilton: McMaster University, Department of Clinical Epidemiology and Biostatistics, 1980.
11. Colton T: Inference on means. In: *Statistics in Medicine.* Boston: Little, Brown, 1974:99–150.
12. Armitage P: Distribution-free methods. In: *Statistical Methods in Medical Research.* New York: John Wiley, 1977:394–407.
13. Cronbach LJ: Coefficient alpha and the internal structure of tests. *Psychometrika 1951;*16:297–334.
14. Conover WJ: Some methods based on ranks. In: *Practical Nonparametric Statistics.* 2nd Ed. New York: John Wiley, 1980:256–61.
15. Colton T: Regression and correlation. In: *Statistics in Medicine.* Boston: Little, Brown, 1974:189–217.
16. 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.
17. George CF: Editorial. *Br J Clin Pharmacol 1986;*22:107S.
18. Feinstein AR: Scientific decisions for data and hypotheses. In: *Clinical Epidemiology — The Architecture of Clinical Research.* Philadelphia: WB Saunders, 1985:515–7.
19. Kendall M, Stuart A: Categorized Data. In: *The Advanced Theory of Statistics,* Vol. 2. *Inference and Relationship.* 4th Ed. High Wycombe: Charles Griffin, 1979:566–615.
20. Bombardier C, Tugwell P, Sinclair A, Dok C, Anderson G, Buchanan WW: Preference for endpoint measures in clinical trials: Results of structured workshops. *J Rheumatol 1982;*9:798–801.