Relationship between Severity and Clinical Importance of Symptoms in Osteoarthritis

N. BELLAMY, G. WELLS*, J. CAMPBELL

Summary **Seventeen patients with primary osteoarthritis of the knee were evaluated with respect to the severity and clinical importance of pain, stiffness and physical function during the conduct of a double-blind randomized controlled trial of flurbiprofen SR versus diclofenac sodium SR using the WOMAC Osteoarthritis Index. Mean importance scores were similar for items within the same dimension as well as between items in different dimensions. In general, low levels of correlation were noted between the severity and importance of symptoms. Analysis of individual WOMAC items within a given subscale indicated that, although highly correlated, they differed from one another. Factor analysis further supported the contention that scores from items within a subscale could be summated into subscale scores. These observations are of importance in the weighting and aggregation of items within discrete dimensions and have the potential for reducing sample size requirements for clinical trials in osteoarthritis.**

Key words: **Clinical Metrology, Importance, Osteoarthritis.**

INTRODUCTION

The principal objective of outcome measurement procedures for therapeutic trials of nonsteroidal anti-inflammatory drugs (NSAIDs) is to detect statistically significant, clinically important, differences in health status between competing treatment programmes. Although much attention has been focused on outcome measures for rheumatoid arthritis (RA) (1,2), much less attention has been paid to the study of patients with osteoarthritis (OA). In general, clinical investigators and international agencies have recommended the use of multiple outcome measures for OA trials (3-6). The use of multiple outcome measures, however, necessitates a downward correction in the statistical p value which results in increased sample size requirements (7). Such problems can be overcome by weighting and aggregating different measures into a single composite index (8). Such a procedure, however, requires respect to relative clinical importance of different items, as well as differences in the lengths of the scales on which the different components are measured. Smythe et al have constructed a composite index, termed the Pooled Index, for application in patients with RA (9). However, their statistical techniques, while correcting for variability in scale length, do not respect the relative clinical importance of the component items (9). In contrast, Gade (10) and Freeman et al (11) have suggested diametrically opposed weighting and aggregation systems for assessing range of movement. We have recently conducted a series of studies (12-16) validating a tridimensional self-administered questionnaire probing pain, stiffness and physical function in patients with OA of the hip or knee. The resulting index is termed the Western Ontario and McMaster Universities (WOMAC) OA Index. An earlier study (13) had indicated that the twenty-four component questions of the Index were regarded by symptomatic patients as being of similar mean clinical importance. In that study (13) importance was measured on 5-point Likert (17) scales. We were unable to determine the extent to which the severity of the patients' symptoms influenced their determination of importance scores. Since we are now using a battery of 10 cm visual analogue (VA) (18) scales as the scaling base for the WOMAC Index, and since we wish to aggregate the component questions within, and, if possible, across the three subscales, we have administered the WOMAC instrument during the conduct of a double-blind randomized controlled multi-centre trial of flurbiprofen (Ansaid-SR) versus diclofenac (Voltaren SR) in OA knee. The study had two major objectives: 1) To examine the relationship between importance and severity scores using WOMAC; 2) To examine whether

Department of Rheumatology, University of Western Ontario, London, Ontario, Canada; *Division of Biometrics, Health and Welfare Canada, Ottawa, Ontario, Canada.

Table I : *Pain : Mean severity score (standard deviation) with blind and informed administration ; mean importance score (standard deviation); Pearson correlation coefficient and p-value of severity and importance with blind administration*

Item	Score-Blind		Score-Informed		p-value	Correlation	Importance		Correlation-Blind		p-value t-test
	Mean	s.d.	Mean	s.d.	paired t-test	Blind/In-formed	Mean	s.d.	r	p	from average
Walking on a flat surface	41.35	27.21	44.24	30.18	.37	.91	77.88	15.24	.23	.37	.54
Going up or down stairs	52.18	26.67	54.59	28.19	.43	.90	81.35	11.78	.05	.85	.00
At night while in bed	37.18	27.86	38.94	30.00	.34	.97	71.18	21.76	.31	.23	.02
Sitting or lying	37.53	29.84	40.94	28.96	.14	.95	71.18	22.92	.17	.52	.05
Standing upright	45.06	30.21	47.12	28.14	.35	.96	67.77	26.77	.60	.01	.38

a simple addition of component item scores into three separate subscale scores was justified. In order to ensure generalizability of our observations to different modes of index presentation, we compared severity scores at termination with prior scores both unavailable (blind presentation) and available (informed administration).

MATERIALS AND METHODS

Seventeen patients attending the rheumatology outpatient clinic at Victoria Hospital, London, with definite radiographic and clinical evidence of primary OA knee were entered as part of a multi-centre double-blind randomized controlled trial comparing Ansaid-SR with Voltaren SR. To be eligible patients had to fulfil the following criteria : Inclusion criteria : symptoms requiring NSAID medication, age ⩾18 years, symptoms ⩾2 months, informed consent obtained ; Exclusion criteria : Gastrointestinal ulceration or bleeding, NSAID hypersensitivity, significant uncontrolled impairment of major organ function, pregnancy or lactation, concomitant use of lithium or anticoagulants, clinically significant abnormalities in haematology or biochemistry. Following enrollment, patients underwent a 3-7 day washout period during which only acetaminophen was allowed. Subsequently, patients were randomly allocated to receive either Ansaid-SR (200 mg po once daily) or Voltaren SR (100 mg po once daily) for six weeks. The medications were identical in appearance thus maintaining physician and patient blinding. Patients were assessed at the end of Week 3 and Week 6. In addition to the WOMAC instrument, data were collected on several other variables. It should be noted, however, the WOMAC Index was only applied in our centre and that this report is confined to severity versus importance issues of the WOMAC instrument in our 17 patients. Data collected from other locations in this multi-centre study, as well as comparison of the two drugs for efficacy and tolerability, will be reported in a separate publication by the other investigators. At the end of the study patients completed all three subscales of the WOMAC Index rating the severity of their symptoms on 10 cm horizontal VA scales (terminal descriptors : None, Extreme) first without their prior WOMAC scores being available (i.e. blind), and again, some five minutes later with their prior scores available (i.e. informed) (13). The reliability, face, content validity, construct validity, and responsiveness of each of the 24 questions posed have been previously defined, verified and reported (14,15). After another five minutes, patients were shown an alternate form of WOMAC, in which they were asked to separately rate on 10 cm horizontal VA scales (terminal descriptors : None, Extreme) the importance, which they attached to being completely symptom free of each of those 24 symptoms. From these data, the mean and standard deviation for severity and importance scores of each item at study termination were calculated. For severity scores, these parameters were calculated for both blind and informed assessments. To examine the relationship between severity and importance, Pearson correlation matrices were constructed for each individual item and the level of correlation and statistical significance determined. The effect of administering the WOMAC Index under blind and informed conditions (i.e., prior score availability) was examined by comparing the mean severity scores under both types of administration using Student's t-test. The issue of whether a simple addition of component items to form subscale totals was examined using Student's t-test, correlation coefficients and factor analysis techniques. Since the 24 component items of WOMAC were considered in each of the three different statistical analyses, the p value, defining statistical significance, was corrected downward by a factor of 24 resulting in a value of $<.002$ (7).

RESULTS

The results are summarized in Tables I-III. The mean age of the study population was 60.24 years (range = 52 - 65) and the mean disease duration 8.57 years (range = 8 months - 20 years). There were 7 male and 10 female subjects. Their radiographic ratings according to

Table II : *Stiffness : Mean severity score (standard deviation) with blind and informed administration ; mean importance score (standard deviation); Pearson correlation coefficient and p-value of severity and importance with blind administration.*

Item	Score-Blind		Score-Informed		p-value	Correlation	Importance		Correlation-Blind		p-value t-test
	Mean	s.d.	Mean	s.d.	paired t-test	Blind/In-formed	Mean	s.d.	r	p	from average
Morning	42.47	30.01	42.65	33.05	.94	.96	66.71	25.01	.41	.11	. 74
Gelling	43.88	28.62	43.94	30.70	.98	.93	66.88	22.20	.43	.09	. 74

the Atlas of Standard Radiographs (19) were as follows : Grade I = 3, Grade II = 4, Grade III = 6, Grade IV = 4. The functional status ratings according to the Steinbrocker classification (20) were as follows : Grade II = 13, Grade III = 4. Of the seventeen patients, 8 received Ansaid-SR and 9 received Voltaren SR. The range of possible values for severity and importance on the VA scales was 0-100 mm.

Blind versus informed administration

For *blind* administration the range of severity scores calculated from the component questions of each dimension was as follows : Pain = 37.18 - 52.18 (Table I); Stiffness = 42.47 - 43.88 (Table II); Physical Function = 34.24 - 60.82 (Table III). For *informed* administration, the range of severity scores was as follows : pain = 38.94 - 54.59 (Table I); Stiffness = 42.65 - 43.94 (Table II); Physical Function = 35.77 - 63.65 (Table III). No statistically significant differences were noted between the severity scores at termination under blind versus informed administration (Tables I - III). The item scores for the two forms of administration were highly correlated (Tables I - III). Since there was no difference between severity scores obtained by blind and informed administration, we have reported the importance issue only with respect to blind administration.

Importance scores

Mean importance scores for component items were as follows: Pain = 67.77-81.35 (Table I), Stiffness = 66.71-67.88 (Table II), Physical Function = 56.29-76.24 (Table III). In all but one instance (bending to floor), the standard deviation for importance scores was less than for the corresponding severity scores.

Severity versus importance scores

We examined the relationship between severity scores and importance scores using correlation coefficients.

Table III : *Physical Function : Mean severity score (standard deviation) with blind and informed administration ; mean importance score (standard deviation); Pearson correlation coefficient and p-value of severity and importance with blind administration*

Item	Score-Blind		Score-Informed		p-value	Correlation	Importance		Correlation-Blind		p-value t-test
	Mean	s.d.	Mean	s.d.	paired t-test	Blind/In-formed	Mean	s.d.	r	p	from average
Descending stairs	49.53	28.56	49.29	28.70	.96	.83	71.77	18.36	.30	.24	.19
Ascending stairs	51.29	25.39	54.41	25.64	.29	.89	71.24	17.53	.20	.43	.18
Rising from sitting	46.77	26.94	50.06	31.16	.55	.71	66.29	24.93	.22	.39	.21
Standing	43.18	28.31	46.29	28.34	.24	.93	68.82	27.35	.53	.03	.47
Bending to floor	48.24	30.22	48.41	32.22	.94	.95	59.65	30.75	.64	.01	.25
Walking on flat surface	44.94	28.18	46.35	28.14	.49	.93	76.24	21.93	.34	.20	.63
Getting in/out of car	49.12	26.76	48.94	27.15	.94	.93	72.06	18.38	.36	.15	.16
Going shopping	48.77	30.33	53.35	29.24	.04	.96	68.77	25.84	.15	.56	.12
Putting on socks	42.77	31.44	42.88	32.68	.94	.98	61.41	30.06	.54	.03	.69
Rising from bed	36.53	31.18	40.65	32.31	.03	.98	56.29	29.10	.62	.01	.01
Taking off socks	41.41	31.97	41.47	31.35	.98	.96	58.53	27.69	.49	.05	.49
Lying in bed	34.24	28.31	38.53	29.69	.07	.95	60.35	20.89	.59	.01	.01
Getting in/out of bath	53.12	30.90	53.82	31.60	.71	.97	71.65	21.57	.47	.06	.09
Sitting	35.59	29.32	35.77	29.08	.94	.94	59.24	23.51	.69	.00	.00
Getting on/off toilet	37.65	29.89	41.35	30.37	.08	.97	63.65	28.01	.39	.13	.03
Heavy domestic duties	60.82	27.91	63.65	28.72	.30	.93	75.47	21.66	.48	.05	.00
Light domestic duties	38.88	29.61	43.35	31.72	.11	.94	71.59	27.56	.23	.38	.02

The following guidelines were used to interpret correlation coefficients: poor correlation = 0 < 0.3; moderate correlation 0.3 < 0.6; good correlation 0.6 < 0.8; excellent correlation ⩾0.8. Seven coefficients were poor, 13 moderate, 4 good but none were excellent. No statistically significant correlation was noted between the importance and severity scores for any of the 24 WOMAC items.

Item aggregation

Using Student's t-test no significant difference was detected between the scores of individual items and the average score for the subscale to which that item belonged, except in two instances of physical function (sitting, heavy domestic duties) (Tables I - III). The level of interitem correlation for components of each of the three subscales was high: pain = 0.79-0.96, stiffness =.83, physical function = 0.52-0.98. Most correlation coefficients were ⩽0.80. Principal component analysis was not performed for stiffness because the subscale contains only two items. However, analysis of the pain and physical function subscales showed that Factor I accounted for 88% of the variance in pain and 83% of the variance in physical function. The factor loading was high on each individual pain item (0.92-0.95) and each individual physical function item (0.70-0.97). There was relatively little additional variance accounted for by Factor II (pain = 7%, physical function = 6%).

DISCUSSION

In this study we have defined the severity and importance of 24 different symptoms of knee OA using the WOMAC Osteoarthritis Index. There is controversy in the literature as to whether serial questionnaires should be administered with or without access to prior scores (21,22). The conservative view prefers blind administration. Since we detected no difference between severity scores obtained by blind versus informed administration of the index in this study, we have based our report on blind administration. However, we have performed parallel analyses using informed scores, obtaining similar results and no interpretative differences.

In a previous study using a Likert-scaled version of WOMAC, we noted that the importance ascribed to symptoms was similar for different items in the same dimensions, as well as for symptoms in different dimensions. If direct comparison can be drawn between Likert and VA scaled responses, then it is of note that the mean importance scores on VA scales, in this study, 56.29 - 81.35 (i.e., 56%-81% along the length of the scale), were similar to mean importance scores reported 2.26 - 2.69 (i.e. 57%-66% along the length of the scale) on the 5-point Likert scales (0 = none, 1 = slight, 2 = moderate, 3 = very important, 4 = extremely important) for the 24 items in our previous study (13). Thus, given that 2 is the mid-point of the Likert scale, and 50 the mid point of the VA scale, we interpret our data as indicating that the majority of patients rate their symptoms somewhere between moderate and very important, and that there is a relatively narrow range for such values. These data support the contention that symptomatic patients regard their own particular symptoms of similar importance to those of other patients.

We had originally considered the possibility of using differences in importance scores as a method of weighting subscale items in the WOMAC Index. From the correlation analysis of severity versus importance, it can be seen that these two elements are distinct and require separate consideration. From a conceptual standpoint, the similarity in importance scores would suggest that items could be simply added together. We wish, however, to explore the statistical justification for such a system of weighting and aggregation. The fact that several items differed significantly from the subscale average, suggests that the items measure different aspects of the dimension and that all were relevant in aggregation. Likewise, although the factor analysis was only conducted on 17 subjects, the high percent of variability accounted for by Factor I and the very high Factor loading on each individual component item, further supports the contention that there are no redundant items in the WOMAC inventory. The high interitem correlation noted within each subscale and the fact that every single item had a high factor loading support the assumption that WOMAC subscale scores for pain, stiffness and physical function can be derived by the simple process of addition. The practical applications of our observations are as follows: 1) The fact that individual patients ascribe moderate levels of importance to each of the 24 WOMAC symptoms provides adequate justification for routinely measuring these symptoms as outcomes in clinical trials provided they fall within the dimensionality of the potential response to the intervention (i.e. NSAID therapy). Other investigators have suggested that the measurement process should focus on clinically relevant outcomes (23), and, indeed in this, as well as our previous study (13), we have demonstrated the clinical relevance of the WOMAC question inventory; 2) The different methods of analysis employed suggest that each item carries the same weight and that the three subscale scores can be derived by the process of simple addition; 3) If some dimensions carry consistently greater importance, then they should also carry more weight in the construction of any composite index; 4) Aggregation has important implications for sample size requirements for clinical trials. For example, without aggregation 24 statistical tests of independent variables would necessitate a reduction in the Type I error from .05 by 24 fold (i.e. $p < .002$). However, aggregation of the WOMAC inventory within each of its three dimensions would necessitate as maximum Type I error correction from .05 by only a factor of 3 (i.e. $p \leqslant .017$). Furthermore, the construction of a composite index, which combined the pain, stiffness and physical function subscales into a single value, would result in only a single statistical comparison and obviate the need for any Type I error correction below .05. The standard formula for calculating sample size for clinical trials is as follows: n per group = $2\ [\frac{(Z_\alpha + Z_\beta)\sigma}{\Delta}]^2$, where σ = standard deviation, and Δ = the change the investigator is interested in detecting (24). As the value for the Type I error is reduced, the value of Z_α increases and the sample size requirements for a proposed trial increase correspondingly.

In this study we have demonstrated that there is no difference in termination scores between blind and informed methods of administration of the WOMAC Osteoarthritis Index. Although symptomatic patients regard their symptoms of similar importance regardless of severity, our observations suggest that importance and severity are little associated. We have also shown that there are no redundant items in the WOMAC Index and demonstrated a justification for deriving subscale scores by the simple addition of component items. We did not address the issue of whether the three separate dimensions can be aggregated into a single total score in this study. This issue is the subject of a current study. At present we recommend that WOMAC subscale scores be constructed by the simple aggregation of items within each of the three different dimensions and that any comparative analysis treats each dimension as a separate entity. For the definitive studies we recommend setting the Type I error at $\leqslant .017$ to make adequate correction for multiple comparisons. When the instrument is being used for pilot studies, however, we do not recommend any correction and prefer to set the Type I error at $\leqslant 0.05$. We make this differentiation to respect the scientific rigour of a definitive study and to reflect the view of Dr. A. Feinstein that the purpose of a fishing expedition is to catch fish.

Acknowledgments: This work was supported in part by the Arthritis Society of Canada, and in part by a grant from the Upjohn Company of Canada.

REFERENCES

1. Bombardier, C., Tugwell, P., Sinclair, A., Dok, C., Anderson, G., Buchanan, W.W. Preference for endpoint measures in clinical trials: results of structured workshops. J Rheumatol 1982, 9, 5, 798-801.
2. Spector, T.D. Epidemiological aspects of studying outcome in rheumatoid arthritis. Br J Rheumatol 1988, 27 (Suppl. I), 5-11.
3. Altman, R.D., Meenan, R.F., Hochberg, M.C., Bole, G.G.Jr., Brandt, K., Cooke, T.D.V., Greenwald, R.A., Howell, D.S., Kaplan, D., Koopman, W.J., Mankin, H., Mikkelsen, W.M., Moskowitz, R., Sokoloff, L. An approach to developing criteria for the clinical diagnosis and classification of osteoarthritis: A status report of the American Rheumatism Association Diagnostic Subcommittee on Osteoarthritis. J Rheumatol 1983, 10, 2, 180-183.
4. Bellamy, N., Buchanan, W.W. Outcome measurement in osteoarthritis clinical trials: the case for standardisation. Clin Rheumatol 1984, 3, 3, 293-303.
5. Guidelines for the clinical evaluation of anti-inflammatory and antirheumatic drugs (adults and children). US Department of Health and Human Services, Public Health Service, Food and Drug Administration, April 1988, 12-15.
6. Guidelines for the clinical investigation of drugs used in rheumatic diseases. World Health Organization Regional Office for Europe, European League Against Rheumatism, 1983, 1-33.
7. Feinstein, A.R. Scientific decisions for data and hypotheses. In: Clinical Epidemiology - The Architecture of Clinical Research. Philadelphia, W.B. Saunders Company, 1985, 515-517.
8. Stewart, A.L., Ware, J.E.Jr., Brook, R.H. Advances in the measurement of functional status: Construction of aggregate indexes. Med Care 1981, XIX(5), 473-488.
9. Symthe, H.A., Helewa, A., Goldsmith, C.H. "Independent assessor" and "Pooled Index" as techniques for measuring treatment effects in rheumatoid arthritis. J Rheumatol 1977, 4, 2, 144-152.
10. Gade, H.G. A contribution to the surgical treatment of osteoarthritis of the hip joint: A clinical study. III. Comments on the follow-up examinations and the evaluation of the therapeutic results. Acta Chir Scand (Suppl) 1947, 120, 37-45.
11. Freeman, M.A.R., Swanson, S.A.V., Todd, R.C. Total replacement of the knee using the Freeman-Swanson knee prothesis. Orthop Related Res 1973, 94, 153-170.
12. Bellamy, N. Osteoarthritis-An Evaluative Index for Clinical Trials. M. Sc. Thesis, McMaster University, Hamilton, Canada, 1982.
13. Bellamy, N., Buchanan, W.W. A preliminary evaluation of the dimensionality and clinical importance of pain and disability in osteoarthritis of the hip and knee. Clin Rheumatol 1986, 5, 2, 231-241.
14. Bellamy, N., Buchanan, W.W., Goldsmith, C.H., Campbell, J., Stitt, L. Validation study of WOMAC: A health status instrument for measuring clinically-important patient-relevant outcomes following total hip or knee arthroplasty in osteoarthritis. J Orthop Rheumatol 1988, 1, 95-108.
15. Bellamy, N., Buchanan, W.W., Goldsmith, C.H., Campbell, J., Stitt, L.W. Validation study of WOMAC: A health status instru-

ment for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rheumatol 1988, 15, 12, 1833-1840.

16. Bellamy, N., Goldsmith, C.H., Buchanan, W.W., Campbell, J., Duku, E. Prior score availability: Observations using the WOMAC Osteoarthritis Index. Br J Rheumatol 1991, 30, 150-151.
17. Likert, R., A technic for the measurement of arthritis. Arch Psychol 1932, 140, 44-60.
18. Scott, J., Huskisson, E.C. Graphic representation of pain. Pain, 1976, 2, 175-184.
19. Kellgren, J.H., Lawrence, J.S. Radiological assessment of osteoarthrosis. Ann Rheum Dis 1957, 16, 494-502.
20. The Cooperating Clinics Committee of the American Rheumatism Association. A seven-day variability study of 499 patients with peripheral rheumatoid arthritis. Arthritis Rheum 1965, 8, 2, 302-334.
21. Guyatt, G.H., Berman, L.B., Townsend, M., Taylor, D.W. Should study subjects see their previous responses ? J Chron Dis 1985, 38, 12, 1003-1007.
22. Scott, J., Huskisson, E.C. Accuracy of subjective measurements made with or without previous scores : an important source of error in serial measurement of subjective states. Ann Rheum Dis 1979, 38, 558-559.
23. Tugwell, P., Bombardier, C., Buchanan, W.W., Goldsmith, C.H., Grace, E., Hanna, B. The MACTAR patient preference disability questionnaire - an individualized functional priority approach for assessing improvement in physical disability in clinical trials in rheumatoid arthritis. J Rheumatol 1987, 14, 3, 446-451.
24. Colton, T. Inference on Means. Statistics in Medicine, Boston. Little, Brown and Company, 1974, p. 142.

Received: 5 May 1989; Revision-accepted: 12 November 1990

Correspondence to: Dr. N. BELLAMY,
Suite 402A, Westminster Tower, Victoria Hospital, 800 Commissioners Road East, London, Ontario, Canada N6A 4G5.