

Validity and reliability of Swedish WOMAC osteoarthritis index

A self-administered disease-specific questionnaire (WOMAC) versus generic instruments (SF-36 and NHP)

Peter Söderman and Henrik Malchau

Department of Orthopedics, Sahlgrenska University Hospital/Sahlgrenska, SE-413 45 Göteborg, Sweden
Tel +46 31-342 18 39. Fax -82 55 99.
Submitted 99-03-03. Accepted 99-10-20

ABSTRACT – We assessed whether a disease-specific, self-administered questionnaire could replace a generic instrument as an outcome tool after total hip replacement, and tested the validity and reliability of the Swedish WOMAC osteoarthritis index. 58 patients operated on with total hip arthroplasty 2–10 years ago were randomized to the study. All patients were asked to answer one disease-specific questionnaire (WOMAC osteoarthritis index) and two generic instruments (NHP and SF-36) in the same week. The procedure was repeated after 4 weeks. We tested content validity, construct validity, criterion validity, test-retest reliability and internal consistency reliability according to total score, domains and items.

We found that both the generic instruments (NHP and SF-36) and the disease-specific test (WOMAC) had a high validity, i.e., measuring what they were supposed to measure (high validity). The questionnaires were also reproducible over time (high reliability). We suggest the Swedish WOMAC to be used after total hip arthroplasty in future studies.

To provide generally applicable systems for various hip disabilities and methods of treatment, Harris (1969) introduced a new rating scale, which took into consideration pain, function, deformity and motion. The Harris Hip Score is the most frequently used disease-specific, but not self-administered hip score system. In recent years, there has been an increased demand for high quality (high validity, reliability and responsiveness) outcomes after medical and surgical interventions for as low

cost as possible. Several new self-administered questionnaires have been developed. These tests are: a) generic, such as the Medical Outcomes Study 36-Item Short-Form Health Survey (SF-36) with questions about patients' health in general, influenced by comorbidity, b) disease-specific, such as the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), and c) patient-specific, such as the McMaster Toronto Arthritis Patient Preference Disability Questionnaire (MACTAR). General health questionnaires allow the effect of treatment to be compared in various medical conditions (Bombardier et al. 1995).

Few reports exist concerning the validity and reliability of the Harris Hip Score. We found that the total score and the domains of pain and function had high validity (unpublished observation). The Harris Hip Score also had high reliability in the pain and function domains, except for the item about sitting. Deformity and range of motion, except hip flexion, had lower validity and reliability than pain and function.

The modern questionnaires, such as WOMAC, SF-36 and the general Nottingham Health Profile (NHP), have been tested for validity, reliability and responsiveness. These instruments have also been translated into many languages. The Swedish versions of NHP and SF-36 have high reliability and validity. WOMAC has been recommended for use as a primary efficacy measure in osteoarthritis treatment studies (Sun et al. 1997). For hip disabilities, there has been no specific test for the Swedish WOMAC's validity and reliability.

We determined whether a disease-specific, self-administered questionnaire could replace a general instrument as an outcome tool after total hip replacement, and tested the validity and reliability of the Swedish WOMAC osteoarthritis index.

Methods

Scales

The Western Ontario and McMaster University Osteoarthritis Index (WOMAC) is a disease-specific, self-administered, health measure developed to study patients with osteoarthritis in the hip or knee. The original version contained 41 items in 5 domains. The domains were pain, stiffness, physical function, social function and emotional function. Internal consistency (Cronbach's alpha) and test-retest reliability (Kendall's tau c statistic) were moderate-to-excellent for pain (0.86 and 0.68, respectively), stiffness (0.90 and 0.48), physical function (0.95 and 0.68), social function (0.89 and 0.61) and emotional function (0.96 and 0.72). The Doyle Index, Lequesne Index and Bradburn Index have been used to show significant correlations with WOMAC items in the same (convergent construct validity) and other dimensions (divergent construct validity). Construct validity was high for the domains of pain, stiffness and physical function. Social and emotional domains had lower validity. Hence the final index utilized the pain (5 questions), stiffness (2 questions) and physical function (17 questions) domains (Bellamy et al. 1988, Sun et al. 1997).

There is one computerized version of the WOMAC and four basic versions with different types of scales on which the response is scored (LK = 5 point Likert scale, VA = 10 cm horizontal visual analog scale) (Bellamy et al. 1992, Bellamy et al. 1997). In the present study, the Likert scale format was used. This questionnaire is reliable, valid and sensitive for detecting clinically important changes in health status following surgical interventions (responsiveness) (Bellamy et al. 1988, Bellamy et al. 1991, McGrory and Harris 1996, Martin et al. 1997). The WOMAC was translated into Swedish using Guillemin's procedure (Roos et al. 1998). The three domains in WOMAC can be analyzed separately or with a single score. Ev-

ery question can be given five alternative answers, which means a total of 0–4 points. The maximum score in the Likert version is 20 points for pain, 8 for stiffness and 68 points for physical function. In this study the score in each domain, was converted into a 0–100 point scale. To compare the results with SF-36 and NHP, the score was inverted. Therefore, a maximum score of 100 points occurred when the patient had minimum pain, stiffness and optimal function.

The Medical Outcomes Study 36-Item Short-Form Health Survey (SF-36) is a general self-administered questionnaire developed for applications in psychometric theory. It was based on the Rand Corporation's health insurance experiment. The original measure was long, containing 108 items. Later on, this test was used to study the effects, which seemed to be a direct function of disease and treatment, that is health-related quality of life (HRQL) (Brazier et al. 1992, Ware and Sherbourne 1992). SF-36 consists of 36 questions divided into eight domains: physical function, social function, role-emotional, role-physical, bodily pain, general health, mental health and vitality. The SF-36 is the most frequently used health-status measure in North America and is of value after total joint replacement in the hip and knee (Ware and Sherbourne 1992, Ritter et al. 1995, Martin et al. 1997). Sullivan et al. (1995) translated the test into Swedish and tested its validity and reliability. The raw score was converted into a 0–100 scale (transformed scale score) as recommended by the Swedish manual. A high value indicated a better result.

The Nottingham Health Profile (NHP) is a self-administered general instrument used to study the quality of life after medical and/or surgical treatments such as total hip replacement (Wiklund and Romanus 1988). As with SF-36 and WOMAC, the NHP has high validity and reliability (Hunt et al. 1980, 1981a, b). The Swedish version was developed by Wiklund and Romanus in the same way as SF-36. The test consists of two parts with 45 yes or no answers. There are 38 questions in part one concerning the patient's difficulties and it is divided into 6 domains: emotional reaction, sleep, energy, pain, physical mobility and social isolation. The items are weighted and each dimension yields a value between 0 and 100. Part two con-

sists of 7 statements that reflect the frequency of problems in various areas of life: occupation, housework, social life, family life, sexual function, hobbies and holidays (McKenna et al. 1981). To avoid negative values in the comparison with WOMAC and SF-36, the score was inverted so that maximum health yielded 100 points.

Statistics

We used SPSS for Windows (version 7.5.1, Chicago) for the statistical analyses.

Questionnaires with good content validity are expected to have fewer categories with ceiling or floor effects. A floor effect occurs when the patient reports the poorest function for all or almost all items and receives the worst possible score, which prevents demonstration of deterioration in function over time for such a patient with use of the questionnaire. A ceiling effect occurs when the patient reports excellent function and receives the best score, making it impossible to show improvement in function over time (Martin et al. 1997). Content validity was tested by directly comparing NHP, WOMAC and SF-36, and by studying the floor effects, ceiling effects, mean, median and standard deviations in each domain.

Construct validity can be evaluated by correlating the questionnaire scores with the characteristics of the patients (number of comorbid conditions, perceived overall health status, and changes in activities). Patients who have more comorbid conditions, poorer perceived overall health status, or changes in activities are expected to receive poorer scores, this is especially true of general questionnaires. Pearson's correlation between the total score of the tested instrument, such as WOMAC, and the domain of interest in other questionnaires, such as function in NHP and SF-36, should be significant at a level of 0.01 (Martin et al. 1997).

Total scores for NHP, WOMAC and SF-36 were calculated in all patients. Male, female, 2 and 10 years postoperatively, age more or less than 70 years old were also analyzed using the total score. The patients were classified according to the Charnley category, that is (A) one hip affected, (B) both hips affected, and (C) multiple-joint disease or other disabilities leading to difficulties in ambulation. Differences between the mentioned

groups were calculated and the significance was determined with the Mann-Whitney U-test. Divergent construct validity and convergent construct validity in the domains of pain and physical function, and total score were tested using Pearson's and Spearman's correlation coefficients. The hypothesis was that the same domains should have higher correlations between each other, for example, pain in WOMAC, SF-36 and NHP, than with other domains such as function.

Criterion validity is present when the scores correlate with an accepted measure (gold standard) of the condition being evaluated. For criterion validity, an acceptable level for Spearman's rho is greater than 0.40 and $p < 0.001$. SF-36 has been extensively validated and is a commonly used score system, for example, after total hip arthroplasty. We used it as the gold standard in our study (Chetter et al. 1997, Martin et al. 1997).

To study test-retest reliability, 58 patients answered the 3 questionnaires twice at an interval of four weeks. Total, domain and item scores were calculated using Pearson's and Spearman's correlation coefficients.

Internal consistency reliability was tested for the questionnaires and in the different domains in WOMAC, SF-36 and NHP. Cronbach's alpha coefficient was used for this purpose.

Patients

We randomly selected 62 patients from Sahlgrenska University Hospital for inclusion in the study. 4 patients were excluded because of mental deficiency. 18 patients had had a total hip arthroplasty 10 years ago, using Charnley (DePuy Olmed), Lubinus IP (Link), Christiansen (Howmedica) or Exeter polished (Stryker Howmedica) prostheses. 40 patients had been operated on 2 years earlier with the Spectron prosthesis (Smith & Nephew). All patients filled in the NHP, WOMAC and SF-36 during 1 week and again after 4 weeks. The mean age of the 58 patients was 71 (52–86) years, 38 were women. The 2-year and 10-year groups had the same distribution in gender and the mean age was 73 and 70 years, respectively. One third of the total group was affected in one hip (Charnley category A) and one-third in both hips (Charnley category B). The remaining third had general disease or another disease that impaired gait (Charnley

Table 1. Results of domain scores for NHP, SF-36 and WOMAC

Test	Domain	N	Mean	Median	SD	Range	Minimum (no. of floor)	Maximum (no. of ceiling)
NHP	Physical motion	58	79	88	22	82	18	100 (18)
	Pain	58	79	92	27	100	0 (1)	100 (27)
	Energy	58	80	100	33	100	0 (5)	100 (39)
	Social isolation	58	93	100	16	71	29	100 (47)
	Emotional reaction	58	90	100	19	85	15	100 (36)
SF-36	Physical function	58	65	70	25	95	5	100 (4)
	Bodily pain	58	71	74	28	100	0 (1)	100 (21)
	Vitality	56	67	70	26	100	0 (1)	100 (10)
	Social function	58	87	100	20	88	12	100 (34)
	Mental health	57	81	84	21	76	24	100 (16)
WOMAC	Pain	58	86	95	20	85	15	100 (28)
	Function	58	77	81	21	87	13	100 (5)
	Stiffness	58	77	75	22	100	0 (1)	100 (17)

Table 2. Total scores for WOMAC, NHP and SF-36. All patients (All), Charnley categories A, B, C, difference (Diff1) between Charnley categories A and C, gender, and difference (Diff2) between gender

	All n 58	ChA n 21	ChB n 15	ChC n 22	Diff1	Male n 20	Female n 38	Diff2
WOMAC	89	93	88	86	7 ^a	91	88	3 NS
NHP	83	90	87	74	16 ^a	92	78	14 ^b
SF-36	73	87	72	63	24 ^a	84	67	17 ^a

^a p = 0.01
^b p = 0.05 (2-tailed)
NS not significant

category C). Similar distributions have been reported in the literature (Garellick et al. 1998). Power analysis predicted that, with 35 patients, there should be at least a 70% chance of detecting a correlation between health status instrument score differences if one existed, assuming a Pearson's r-value of 0.40 (McGrory and Harris 1996).

Results

All 58 patients answered the questionnaire fully. A few patients did not answer some of the individual items.

Validity

Content validity. The content of NHP, SF-36 and WOMAC was related (pain, function and stiffness), but SF-36 and NHP also contained addi-

tional domains (vitality, social function, mental health, sleep and role-function). 1 floor value was seen in WOMAC, 2 in SF-36 and 6 in NHP. There were fewer ceiling values for function in WOMAC and SF-36 than in NHP, while the pain domain scored equally. The median in the domains of energy, social isolation and emotional reactions in NHP, and social function in SF-36 was 100 points (Table 1).

Construct validity. SF-36 and NHP scored a higher difference between gender and the Charnley categories than the disease-specific score system, WOMAC (Table 2). There were no statistically significant differences between the three questionnaires with respect to age and follow-up time (Table 3).

The pain domain in WOMAC correlated better with pain in SF-36 (Pearson's $r = 0.62$, Spearman's $\rho = 0.59$) and NHP ($r = 0.71$, $\rho = 0.62$)

Table 3. Total scores WOMAC, NHP and SF-36. 2 and 10 years postoperatively, difference (Diff3) between follow-up scores, patients younger or older than 70 years, and difference (Diff4) between younger and older patients

	2-year n 40	10-year n 18	Diff3	Age < 70 n 24	Age > 70 n 34	Diff4
WOMAC	90	87	3 NS	91	88	3 NS
NHP	84	80	4 NS	87	80	7 NS
SF-36	75	70	5 NS	78	69	9 NS

p < 0.01 (2-tailed), NS not significant

Table 4. Correlations between domains and total scores in SF-36, NHP and WOMAC, determined with Spearman's correlation

Domain	Score	Pain			Function			Total		
		SF-36	NHP	WOMAC	SF-36	NHP	WOMAC	SF-36	NHP	WOMAC
Pain	SF-36	1								
	NHP	0.69	1							
	WOMAC	0.59	0.62	1						
Function	SF-36	0.57	0.54	0.45	1					
	NHP	0.56	0.59	0.55	0.73	1				
	WOMAC	0.71	0.71	0.76	0.67	0.82	1			
Total	SF-36	0.81	0.66	0.58	0.75	0.76	0.83	1		
	NHP	0.7	0.79	0.66	0.59	0.74	0.78	0.8	1	
	WOMAC	0.66	0.67	0.83	0.59	0.7	0.93	0.73	0.71	1

Correlation is significant at the 0.01 level (2-tailed), n 52–58

than with function in SF-36 ($r = 0.56$, $\rho = 0.45$) and NHP ($r = 0.59$, $\rho = 0.55$) (Table 4). The same results were obtained when comparing the function domain in the WOMAC and NHP scores with pain domains, but not between WOMAC and SF-36.

The correlation between domains in one questionnaire and its total score was higher than with the other questionnaires. For example, there was a better correlation between the WOMAC score for pain and the total WOMAC score than the WOMAC score for pain and the NHP total score.

Our study showed high convergent and divergent validity for the disease-specific test.

Criterion validity. The Spearman's rho was acceptable when the total score for SF-36 and WOMAC (0.73) or NHP (0.80) were correlated (Table 4). This was also true of the correlation between SF-36 domains and the same domains in NHP and WOMAC (Spearman's rho range 0.59–0.82).

Reliability

Test-retest reliability. Total score, domains, and items were each tested with Pearson's and Spearman's correlation coefficients between two examinations, which were 3–4 weeks apart. The reliability of the total score was excellent in every questionnaire (WOMAC $r = 0.90$, SF-36 $r = 0.92$, and NHP $r = 0.94$). The reliability of the domains was also high, and the disease-specific test had the highest value (Table 5).

The test-retest reliability for each item was moderate-to-excellent (Pearson's r range 0.47–0.81). The correlations were significant at the 0.01 level (2-tailed).

Internal consistency reliability. Cronbach's alpha coefficient showed a high correlation between each domain (0.60–0.96). The WOMAC items received higher values (0.88–0.96) than the items in the NHP (0.65–0.83) and the SF-36 (0.75–0.94). WOMAC scored higher reliability for pain and function than SF-36 and NHP.

Table 5. Test-retest reliability for WOMAC, SF-36 and NHP. Pearson's (r) and Spearman's (rho) correlations between domains and total score

Test	WOMAC		SF-36		NHP	
	r	rho	r	rho	r	rho
Total	0.9	0.87	0.92	0.91	0.94	0.88
WOMAC						
pain	0.88	0.78				
stiffness	0.76	0.69				
function	0.91	0.91				
SF-36						
physical function			0.78	0.78		
role-function			0.74	0.72		
bodily pain			0.85	0.83		
general health			0.84	0.81		
vitality			0.92	0.92		
social function			0.7	0.65		
role-emotional			0.39	0.4		
mental health			0.8	0.69		
NHP						
emotional reaction					0.9	0.81
sleep					0.82	0.86
energy					0.78	0.87
pain					0.89	0.83
physical motion					0.88	0.83
social isolation					0.82	0.8

Discussion

We found that both the general instruments (NHP and SF-36) and the disease-specific test have high validity and reliability. Earlier studies have reported moderate-to-good correlation between pain, physical function and overall score in WOMAC and SF-36 (0.50–0.73) (Bombardier et al. 1995, Wright and Young 1997a). In our study there was a good correlation between pain, function and total score (0.59, 0.67 and 0.73, respectively).

Criterion validity with SF-36 as the gold standard seems to be acceptable, that is over 0.40, but some authors have proposed a 0.80 level of correlation when an instrument is to be considered as the gold standard (Bellamy et al. 1988).

There is no single ideal measure of reliability and each reliability measure has potential shortcomings and must be interpreted carefully. Spearman's correlation coefficient is familiar to most readers and is easily interpreted. It can be used to calculate validity and reliability (Katz et al. 1996). Pearson's and Spearman's rank-correlation coefficients are the parameters most frequently used in analyzing parallel data. We used both of these co-

efficient to make it easier for others to compare their results (Krabbe et al. 1996). We also tested the linearity between the correlations and found that one could use both Pearson's and Spearman's correlation coefficients because there is a high linearity with no skewed correlations.

According to Bombardier et al. (1995), disease-specific measures focus on the disorder under consideration and the patient's problems related to it and therefore may be more relevant to the patient and the physician than general instruments, as well as better at detecting the effect of the treatment. General measures detect complications or side-effects in areas of function or organ systems not specifically related to the disease under consideration. General instruments also allow the impact of treatment to be compared in various medical conditions (Bombardier et al. 1995). The

American Academy of Orthopaedic Surgeons and the Société Internationale de Chirurgie Orthopédique et de Traumatologie recommend that outcome studies include a disease-specific measure, such as WOMAC, in all studies of the outcome of hip arthroplasty (Laupacis et al. 1993). Disease-specific scales are more sensitive than general health status measures for evaluation of the outcomes of orthopedic procedures. However, a complete evaluation of an operation such as hip arthroplasty requires the use of both a specific and a general questionnaire (Rorabeck et al. 1994, Wright and Young 1997b, Bellamy et al. 1988). Both the Likert score and Visual analog scale are available for WOMAC (Bellamy et al. 1992). Visual analog scales are reliable but difficult to score and cannot be administered by telephone. They are not understood by 7% of the population (Katz et al. 1996). With telephone follow-up evaluation, one can complete a missing answer in a single Likert question and therefore increase the response rate (McGrory et al. 1997).

In conclusion, the Swedish WOMAC osteoar-

thrititis index measured what it was supposed to measure very well (high validity) and was also reproducible (high reliability). As regards the present study and others, we recommend the Swedish WOMAC Likert score for studies after total hip arthroplasty, but it would be better with a shorter questionnaire. To study results after hip arthroplasty, and to make the results comparable with other diseases, the short questionnaire should contain both a disease-specific part (with items such as walking, hip flexion and pain) and a general part, including the Charnley classification (Callaghan et al. 1990, Bryant et al. 1993, Garellick et al. 1994).

The authors thank the Volvo Research Foundation, the Swedish National Board of Health and Welfare, the Gothenburg Medical Association, the Dr. Felix Neuberger Foundation, the Greta and Einar Asker Foundation and the Hjalmar Svensson Foundation for financial support.

- Bellamy N, Buchanan W, Goldsmith C H, Campbell J, Stitt L W. 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 and the knee. *J Rheumatol* 1988; 15: 1833-40.
- Bellamy N, Wells G, Campbell J. Relationship between severity and clinical importance of symptoms in osteoarthritis. *Clin Rheumatol* 1991; 10 (2): 138-43.
- Bellamy N, Kean W F, Buchanan W W, Gercz-Simon E, Campbell J. Double-blind randomized controlled trial of sodium meclofenamate (Meclomen) and diclofenac sodium (Voltaren): Postvalidation reapplication of the WOMAC osteoarthritis index. *J Rheumatol* 1992; 19: 153-9.
- Bellamy N, Campbell J, Stevens J, Pilch L, Stewart C, Mahmood Z. Validation study of a computerized version of the Western Ontario and McMaster Universities VA3.0 Osteoarthritis Index. *J Rheumatol* 1997; 24 (12): 2413-5.
- Bombardier C, Melfi C A, Paul J, Green R, Hawker G, Wright J, Coyte P. Comparison of a generic and a disease-specific measure of pain and physical function after knee replacement surgery. *Med Care (Suppl)* 1995; 33 (4): AS131-AS44.
- Brazier J E, Harper R, Jones N M B, O' Cathain A, Thomas K J, Usherwood T, Westlake L. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *BMJ* 1992; 305: 160-4.
- Bryant M J, Kernohan J R, Nixon J R, Mollan R A B. A statistical analysis of hip scores. *J Bone Joint Surg (Br)* 1993; 75: 705-9.
- Callaghan J J, Dysart S H, Savory C F, Hopkinson W J. Assessing the results of hip replacement. A comparison of five different rating systems. *J Bone Joint Surg (Br)* 1990; 72: 1008-9.
- Chetter I C, Spark J I, Dolan P, Scott D J A, Kester R C. Quality of life analysis in patients with lower limb ischaemia: suggestion for European standardization. *Eur J Vasc Surg* 1997; 13: 597-604.
- Garellick G, Herberts P, Strömberg C, Malchau H. Long-term results of Charnley arthroplasty. A 12-16-year follow-up study. *J Arthroplasty* 1994; 9: 333-40.
- Garellick G, Malchau H, Herberts P. Disease-specific or general health outcome measure in evaluation of total hip replacement. A comparison between Harris Hip Score and Nottingham Health Profile. 1998; *J Bone Joint Surg (Br)*. Accepted
- Harris W H. Traumatic arthritis of the hip after dislocation and acetabular fractures: treatment by mold arthroplasty. *J Bone Joint Surg (Am)* 1969; 51: 737-55.
- Hunt S M, McKenna S P, McEwen J, Backett E M, Williams J, Papp E. A quantitative approach to perceived health status: a validation study. *J Epidemiol Community Health* 1980; 34: 281-6.
- Hunt S M, McKenna S P, McEwen J, Williams J, Papp E. The Nottingham health profile: subjective health status and medical consultations. *Soc Sci Med* 1981a; 15A: 221-9.
- Hunt S M, McKenna S P, Williams J. Reliability of a population survey tool for measuring perceived health problems: a study of patients with osteoarthritis. *J Epidemiol Commun Health* 1981b; 35: 297-300.
- Katz J N, Phillips C R, Poss R, Harrast J J, Fossel A H, Liang M H, Sledge C B. Correspondence. *J Bone Joint Surg (Am)* 1996; 78: 1445-8.
- Krabbe P F M, Essink-Bot M-L, Bonsel G J. On the equivalence of collectively- and individually-collected responses. *Med Decis Making* 1996; 16: 120-32.
- Laupacis A, Bourne R, Rorabeck C, Feeny D, Wong C, Tugwell P, Leslie K, Bullas R. The effect of elective total hip replacement on health-related quality of life. *J Bone Joint Surg (Am)* 1993; 75: 1619-26.
- Martin D P, Engelberg R, Agel J, Swiontkowski F. Comparison of the musculoskeletal function assessment questionnaire with the Short form-36, the Western Ontario and McMaster Universities Osteoarthritis index, and the Sickness Impact Profile Health-Status Measures. *J Bone Joint Surg (Am)* 1997; 79: 1323-35.
- McGrory B J, Harris W H. Can the Western Ontario and McMaster Universities (WOMAC) Osteoarthritis Index be used to evaluate different hip joints in the same patients? *J Arthroplasty* 1996; 11: 841-4.
- McGrory B J, Shinar A A, Freiberg A A, Harris W H. Enhancement of the value of hip questionnaires by telephone follow-up evaluation. *J Arthroplasty* 1997; 12: 340-3.
- McKenna S P, Hunt S M, McEwen J. Weighting the seriousness of perceived health problems using Thurstone's method of paired comparisons. *J Epidemiology* 1981; 10: 93-7.

- Ritter M A, Albohm M J, Keating M, Faris P M, Meding J B. Comparative outcomes of total joint arthroplasty. *J Arthroplasty* 1995; 10: 737-41.
- Roos E M, Roos H P, Lohmander L S, Ekdahl C, Beynon B D. Knee injury and osteoarthritis outcome score (KOOS)-development of a self-administered outcome measure. *J Orthop Sports Phys Ther* 1998; 28: 88-96.
- Rorabeck C H, Bourne R B, Laupacis A, Feeny D, Wong C, Tugwell P, Leslit K, Bullas R. A double-blind study of 250 cases comparing cemented with cementless total hip arthroplasty. *Clin Orthop* 1994; 298: 156-64.
- Sullivan M, Karlsson J, Ware J R. The Swedish SF-36 health survey. I. Evaluation of data quality, scaling assumptions, reliability and construct validity across general populations in Sweden. *Soc Sci Med* 1995; 41: 1349-358.
- Sun Y, Sturmer T, Gunther K P, Brenner H. Reliability and validity of clinical outcome measurements of osteoarthritis of the hip and knee. A review of the literature. *Clin. Rheumatol* 1997; 16: 185-98.
- Ware J E, Sherbourne C D. The MOS 36-Item Short-Form Health Survey (SF-36). *Med Care* 1992; 30: 473-83.
- Wiklund I, Romanus B. Nottingham health profile. Livskvalitetsbedömning hjälp vid operationsprioritering. *Läkartidningen* 1988; 85: 3060-1.
- Wright J G, Young N L. A comparison of different indices of responsiveness. *J. Clin Epidemiol* 1997a; 50: 39-246.
- Wright J G, Young N L. The Patient-Specific Index: Asking patients what they want. *J Bone Joint Surg (Am)* 1997b; 79: 974-83.