Reliability and Validity of Clinical Outcome Measurements of Osteoarthritis of the Hip and Knee - A Review of the Literature

Y. SUN, T. STÜRMER, K.P. GÜNTHER*, H. BRENNER

Summary **High reliability and validity of clinical rating schemes is crucial for their use as outcome measurements of treatment of hip and knee osteoarthritis. In this paper, we review the empirical evidence on the reliability and validity of commonly used clinical scores. Clinical scores and related reliability and validity studies were identified by systematic literature search. Scores were classified according to the type and joint. Reliability and validity studies were characterized according to design, population, number and qualification of observers, number of measurements, time interval between repeat measurements and results. Reliability and validity studies were reported for only 6 and 15 of the 45 identified clinical scores, respectively. Although comparisons are difficult due to differences in study design, relatively high reliability was reported for most measurements of pain, stiffness, and physical function, while results are less conclusive for clinical signs. Most validity studies focused on the correlation between various scores. Correlation was generally found to be high for overall numerical ratings, but scores often differed with respect to the interpretation of these ratings. Validity has been more comprehensively studied for Lequesne's scores, WOMAC, and ILAS, and these scores have shown satisfactory responsiveness to different treatment effects. Overall, knowledge on reliability and validity of clinical scores of hip and knee osteoarthritis is limited, underlining the need for further properly designed and conducted studies.**

Key words **Osteoarthritis, Clinical Assessment, Outcome Measurement, Reliability and Validity**

INTRODUCTION

Osteoarthritis (OA) is the most common joint disease and a major public health problem throughout the world (1). OA of the hip and knee joints (cox- and gonarthrosis) is recognized as a major cause of pain, disability, and high social expenditure (1, 2). Treatment is usually aimed at reducing symptoms and preventing impairment and disability. Increasing importance is being placed on the monitoring of outcomes of treatment in clinical studies to investigate the possible therapeutic use of different therapies (such as surgical treatment, physical therapy, or drug therapy). Clinical assessment plays a central role for this purpose.

In the past few decades, a large number of clinical instruments for outcome measurements as well as severity ratings of hip and knee OA have been introduced. Knowledge is limited, however, on the appropriateness of various instruments for clinical and epidemiologic studies, which require a high level of reliability and validity of measurements.

In this paper, which was developed during the preparation of a multi-center study on the epidemiology of hip and knee OA in South Germany, we review the empirical evidence on the reliability and validity of commonly used clinical rating systems of hip and knee OA.

SCOPE OF THIS REVIEW

This review will focus on the following aspects of studies on the inter-rater, intra-rater and test-retest-reliability and of the content and construct validity of commonly used clinical rating systems:

1. General characteristics and the special use of the scores.

Department of Epidemiology and *Department of Orthopaedics, University of Ulm, Ulm, Germany.

2. Clinical items and their weighting included in the different clinical rating systems.
3. Setting and design of reliability studies, such as number and qualification of raters and number and spectrum of patients.
4. Intra-rater, inter-rater and test-retest-reliability for both single clinical items and overall scores.
5. Setting and design of validity studies, such as the qualification of observers and the spectrum of patients.
6. Content and construct validity.

Literature search

In order to comprehensively identify instruments of clinical outcome measurements of osteoarthritis of the hip and knee and studies on their reliability and validity, MEDLINE searches were performed for the time interval from 1984 to 1995, using the following controlled vocabulary: "osteoarthritis", "index of severity", "severity", "clinical rating", "clinical assessment", "outcome measurement", "reliability", "validity". Bibliographies and cross-referencing were used for identification of pre-1984 studies and complementation of the literature search.

Conceptual and statistical background for assessing reliability

Three types of reliability are commonly distinguished: 1) inter-rater-reliability indicates to which degree different observers, using a method to assess the same individual, obtain the same result (3). 2) intra-rater-reliability indicates to which degree the same result is obtained, if the measurement is applied more than once on the same individual, by the same observer (3). 3) test-retest-reliability indicates to which degree the same result is obtained in repeat applications of self-assessment instruments (3).

Statistical measures of reliability depend on the measurement scale: Pearson's correlation coefficient (r) (4) is commonly used to quantify correlation between repeat measurements of continuous variables. These variables should be normally distributed. Spearman's rank correlation coefficient (rs) (4) is often used to assess the reliability of variables that do not follow the normal distribution. An alternative to Spearman's rank correlation coefficient is Kendall's tau (τ_a, τ_b, τ_c), (5) which is somewhat less frequently used. All of these correlation coefficients can take values from –1 (maximum possible negative correlation) to +1 (perfect positive correlation). A limitation of these correlation coefficients for quantifying reliability is that they do not reflect systemic variation between observers or between measurements. Some authors proposed the use of t-tests for paired comparisons along with the correlation coefficient to reflect systematic variations. It should be noted, however, that the test statistic reflects the size of the sample in addition to the difference between ratings. Furthermore, none of the aforementioned approaches can be used to quantify reliability of more than two measurements per study participant.

An alternative measure of reliability that reflects both systematic and random variation between tests is the intraclass correlation coefficient (ICC) (6). This coefficient is based on the estimation of variance components in analysis of variance. ICC quantifies the proportion of overall variance of ratings that is due to between-subjects variability, and it can therefore take values from 0 (variance entirely due to imperfect reliability) to 1 (variance entirely due to between-subjects variability). ICC can be used for two or more measurements per study participant (6).

Reliability of categorical data is commonly quantified by kappa coefficients (κ) (7) which quantify the agreement of classification beyond chance agreement. Although primarily developed for dichotomous variables, kappa coefficients can also be applied to variables with more than two categories. For ordinal variables, weighted kappa coefficients are often used in which disagreements are weighted by the magnitude of the discrepancy between ratings. Weighted kappa coefficients are equivalent to Pearson's correlation coefficient and the intraclass correlation coefficient applied to the categorical data under certain conditions (7). Kappa coefficients have a maximum value of 1 when agreement is perfect. A value of 0 indicates no agreement beyond agreement by chanche, and a value below 0 is observed with less than chance agreement. Although the magnitude of kappa depends on a variety of factors other than reliability, such as the marginal distributions of ratings, values greater than 0.75 are generally considered to represent excellent agreement beyond chance, while values below 0.40 are considered to reflect poor agreement (7).

An alternative to the kappa coefficient is the Goodman-Kruskal's gamma coefficient (γ) (8), which is less frequently used. The gamma coefficient is defined as the difference between agreement and disagreement of paired observations divided by the sum of agreement and disagreement. It can take values from –1 (maximum possible disagreement) to +1 (perfect agreement). A limitation of the gamma coefficient for quantifying reliability is that it can only be used for two measurements per study participant. Furthermore, it does not take chance agreement into account.

Conceptual and statistical background for assessing validity

Validity is often defined as the extent to which an instrument measures what it purports to measure (3). There are different types of validity, neither of which is typically directly measurable. A distinction is commonly made between content validity and construct validity. Content validity raises the question how adequately the sampling of items reflects the aims of an instrument as specified in the conceptual definition of its scope (3). Construct validity indicates how properly an instrument reflects the theoretical construct behind the measurement (3).

The following approaches are commonly taken to assess validity of clinical scores in practice:

1. Relationship between the scores and a "goldstandard" or a measurement with known validity (3). This so-called "criterion" or "concurrent" validity is often quantified by Pearson's correlation coefficient or Spearman's rank correlation coefficient, depending on the measurement scale. Sometimes, two scores with unknown validity are compared in the same way to assess whether they measure the same construct ("correlational evidence" (3)).
2. Ability to discriminate between groups of patients, such as patients with or without effective treatment ("predictive validation"). Closely related to predictive validation is assessment of responsiveness to treatment, such as drug treatment or surgery. Student's t-test and Wilcoxon's test for paired observations are commonly used for that purpose. An alternative measure, called "responsiveness index" has been newly introduced by Guyatt et al. in 1987 (9). The responsiveness index is calculated by taking the ratio of a clinically important difference divided by the square root of twice the mean square error of repeated measurements in stable patients.
3. Application of factor analysis to identify whether an instrument measures only one major aspect or several independent aspects of the condition under investigation ("factorial validation"). Principal component analysis (PCA) is commonly used for that purpose (3).

In general, "criterion" and "predictive" validation are used to asses content validity, while "correlational evidence" and "factorial" validation are used to assess construct validity (3).

RESULTS

We identified 45 clinical rating systems that may be used for the outcome measurement of hip and knee OA (see Table I). They were developed between 1947 and 1994. In total, there are 18 scores for the hip, 24 scores for the knee, and 3 scores for both of them. Only 5 scores (Danielsson, Lequesne's L-ISH and L-ISK, Jones, WOMAC) were established especially for the assessment of OA of hip and/or knee joints. All other scores were introduced as comprehensive instruments not only for the assessment of patients with OA but also for patients with other disorders of the hip or knee (such as rheumatoid arthritis or post-traumatic sequelae). Most of the scores are primarily used for quantifying the clinical outcome after hip or knee arthroplasty. In general, two types of items contained in the scores can be distinguished: "subjective items" which essentially are self-reported by patients, such as pain, stiffness and items concerning the physical or social disability, and "objective items" which are based on medical examinations, such as clinical signs (including range of motion) and radiographic signs. As can be seen from Table I, most of the scores are predominantly based on subjective items.

More specifically, five major components are considered in the various rating scores; symptoms, clinical signs, physical and/or social function, radiographic signs, and the emotional status of patients. Within each component, items are often measured on an ordinal scale. Scores for each component are obtained by summing up ratings of single items. In most instruments, component scores are added to an overall rating index of severity of hip and knee OA. Such instruments are denoted "overall scores" in this paper. But in some of the rating scores (denoted "separated scores"), component scores are only used separately to characterize various aspects of OA. An overview on the contained components and their weightings in calculation of the overall scores is shown in Table II. Symptoms, clinical signs and physical function of lower extremities are included in most of the scores. The components of the clinical scores are weighted very differently, however, in various overall scores. For example, the score by Shepherd and Lequesne's scores emphasize functional status of patients, whereas some of the knee scores (Wilson, Hungerford and Hofmann) do not consider functional status at all. Radiographic signs are included in only three scores (HSS-2, Mayo, Lotke), and only the knee score by Baumgaertner et al. includes emotional status.

Overviews on the setting, design, statistical methods, and overall results of reliability and validity studies are given in Tables III, IV and V.

Reliability and validity studies have been reported for 6 and 15 of the 45 clinical scores, respectively. All of the studies were carried out after 1980. In general, patients included were either OA patients or patients who underwent arthroplasty (for various reasons including OA).

Table I: *Clinical scores for outcome measurement of hip and knee OA*

Joint	Name or Abbreviation	Author(s)	Year	Primary use	Proportion based on "subjective" items	Ref.
Hip	Gade	Gade	1947	Hip-Arthorplasty[11]	67%	10
	Judet	Judet et al.	1952	Hip-Arthroplasty[11]	67%	11
	MdA	Merle d'Aubigne et al.	1954	Hip-Arthroplasty[11]	67%	12
	Shepherd	Shepherd	1954	Hip-Arthroplasty[11]	89%	13
	Stinchfield	Stinchfield et al.	1957	Hip-Arthroplasty[11]	66%	14
	Larson	Larson	1963	Hip-Arthroplasty[11]	85%	15
	Danielsson	Danielsson	1964	OA[12]	67%	16
	Lazansky	Lazansky	1967	Hip-Arthroplasty[11]	50%	17
	Harris	Harris	1969	Hip-Arthroplasty[11]	91%	18
	HSS-1[1]	Wilson et al.	1972	Hip-Arthroplasty[11]	75%	19
	Andersson	Andersson	1972	Hip-Arthroplasty[11]	67%	20
	Charnley	Charnley	1972	Hip-Arthroplasty[11]	67%	21
	McKee	McKee et al.	1973	Hip-Arthroplasty[11]	100%	22
	L-ISH[2]	Lequesne	1980	OA[12]	100%	23
	UCLA[3]	Dutton et al.	1982	Hip-Arthroplasty[11]	100%	24
	HSS-2[4]	Pellicci et al.	1985	Hip-Arthroplasty[11]	50%	25
	Mayo[5]	Kavanagh et al.	1985	Hip-Arthroplasty[11]	80%	26
	JOA[6]	Yano et al.	1990	Hip-Arthroplasty[11]	80%	27
Knee	Potter	Potter et al.	1972	Knee-Arthroplasty[13]	30%	28
	Ranawat	Ranawat et al.	1973	Knee-Arthroplasty[13]	52%	29
	Larson	Larson et al.	1974	Knee-Arthroplasty[13]	80%	30
	Kettelkamp	Kettelkamp et al.	1975	Knee-Arthroplasty[13]	52%	31
	Wilson	Wilson et al.	1976	Knee-Arthroplasty[13]	40%	32
	Freeman	Freeman et al.	1977	Knee-Arthroplasty[13]	80%	33
	Lotke	Lotke et al.	1977	Knee-Arthroplasty[13]	70%	34
	Aichroth	Aichroth et al.	1978	Knee-Arthroplasty[13]	62%	35
	Ewald	Ewald	1979	Knee-Arthroplasty[13]	80%	36
	Goldberg	Goldberg et al.	1981	Knee-Arthroplasty[13]	81%	37
	Hungerford	Hungerford et al.	1982	Knee-Arthroplasty[13]	40%	38
	L-ISK[7]	Lequesne	1982	OA[12]	100%	39
	Wang	Wang	1984	Knee-Arthroplasty[13]	72%	40
	Weber	Weber et al.	1985	Knee-Arthroplasty[13]	40%	41
	Matthews	Matthews et al.	1986	Knee-Arthroplasty[13]	100%	42
	Merkel	Merkel et al.	1986	Knee-Arthroplasty[13]	68%	43
	Mackinnon	Mackinnon et al.	1988	Knee-Arthroplasty[13]	70%	44
	Hernigou	Hernigou et al.	1988	Knee-Arthroplasty[13]	40%	45
	KS[8]	Insall et al.	1989	Knee-Arthroplasty[13]	67%	46
	Baumgaertner	Baumgaertner et al.	1990	Knee-Arthroplasty[13]	100%	47
	Raunest	Raunest et al.	1990	Knee-Arthroplasty[13]	65%	48
	Larson 2	Larson et al.	1991	Knee-Arthroplasty[13]	50%	49
	Hofmann	Hofmann et al.	1991	Knee-Arthroplasty[13]	40%	50
	Jones	Jones et al.	1991	OA[12]	50%	51
Hip/ Knee	WOMAC[9]	Bellamy et al.	1982	OA[12]	100%	52
	Öberg	Öberg et al.	1994	H/K disorder[14]	60%	53
	ILAS[10]	Shields et al.	1994	H/K disorder[14]	0%	54

[1] Hipscore of the Hospital of Special Surgery, Nr. 1; [2] Lequesne's index of severity of hip osteoarthritis; [3] Hipscore of the University College of Los Angeles; [4] Hipscore of the Hospital of Special Surgery, Nr. 2; [5] Hipscore of the Mayo Clinic; [6] Hipscore of the Japanese Orthopedic Association; [7] Lequesne's index of severity of knee osteoarthritis; [8] Kneescore of the Knee Society; [9] The Western Ontario and McMaster Universities Osteoarthritis Index;[10]Lower Extremity Assistance Scale of the University of Iowa; [11,13] Outcome measurement of arthroplastic of hip and/or knee disorders; [12] Outcome measurement of treatment of hip and/or knee OA; [14] Outcome measurement of treatment of hip and/or knee disorders.

Table II: *Components contained and their weightings in different scores*

Type of score	Joint	Name or Abbreviation	Symptoms	Clinical signs	Function	Radiographic signs	Emotional status
Separate	Hip	Gade	x	x	x		
		Danielsson	x	x	x		
		Charnley	x	x	x		
		UCLA	x		x		
	Knee	Matthews	x		x		
		KS	x	x	x		
		Jones	x	x			
Overall	Hip	Judet	33%	33%	33%		
		MdA	33%	33%	33%		
		Shepherd		11%	89%		
		Stinchfield	33%	33%	33%		
		Larson	40%	15%	45%		
		Lazansky	27%	50%	23%		
		Harris	44%	9%	47%		
		HSS-1	25%	25%	50%		
		Andersson	33%	33%	33%		
		McKee	33%		67%		
		L-ISH	33%		67%		
		HSS-2	17%	17%	33%	33%	
		Mayo	40%		40%	20%	
		JOA	40%	20%	40%		
	Knee	Potter	19%	70%	11%		
		Ranawat	30%	48%	22%		
		Larson	30%	20%	50%		
		Kettelkamp	25%	48%	27%		
		Wilson	40%	60%			
		Freeman	50%	20%	30%		
		Lotke	36%	30%	34%	x[1]	
		Aichroth	12%	38%	50%		
		Ewald	50%	20%	30%		
		Goldberg	44%	19%	37%		
		Hungerford	40%	60%			
		L-ISK	33%		67%		
		Wang	36%	28%	36%		
		Weber	20%	60%	20%		
		Merkel	33%	33%	34%		
		Mackinnon	30%	30%	40%		
		Hernigou	20%	60%	20%		
		Baumgaertner	33%		33%		33%
		Raunest	40%	35%	25%		
		Larson 2	30%	50%	20%		
		Hofmann	40%	60%			
	Hip/Knee	WOMAC	29%		71%		
		Öberg	5%	40%	55%		
		ILAS			100%		

[1] only the component-scores of symptoms, signs and functions are added to an overall rating index.

Table III: *Characteristics and results of reliability studies*

Joint	Name or abbrev.	Author's	Year	Study population			Qualification of observers	Number of observers per patient	Number of ratings[2]	Time interval between repeat measures	Inter-rater reliability	Intra-rater reliability	Test-retest reliability	Ref.
				type[1]	size	age								
Hip	L-ISH	Lequesne	1987	OA pat.	38	not given	"trained"	2		not given	"no systematic difference between raters" (t-test)			56
Knee	L-ISK	Lequesne	1987	OA pat.	24	not given	"trained"	2		not given	"no systematic difference between raters" (t-test)			56
	Jones	Jones et al.	1991	OA pat.	49	50-92	consultant rheumatologist, consultant geriatrician, rheumatology senior registrar, rheumatology registrar, general medical regisrar	2	2	1-3 hours for intra-rater reliability, < 5 days for inter-rater reliability	overall agreement for various types of pain κ=0.53-0.72 various types of stiffness, κ=0.46-0.62 various clinical signs κ=0.09-0.35	overall agreement for various types of pain κ=0.76-0.86 various types of stiffness, κ=0,74-0.90 various clinical signs κ=0.54-0.78		51
Hip/Knee	WOMAC	Bellamy et al.	1988	OA pat.	57	55-82	not applicable (self-rating)		2	one week			for pain τ_c=0.680/0.64[3] for stiffness τ_c=0.48/0.61[3] for function τ_c0.68/0.72[3]	58
	Öberg	Öberg et al.	1994	OA pat.	42	46-91	physical therapist	2		not given	for pain γ=1.00 for functional items γ=0.99-1.00 for clinical signs γ=1.00			53
	ILAS	Shields et al.	1995	AP pat.	86	34-88	physical therapist	2	2	<2 days for inter-rater reliability Videotaped assessment 3-6 months apart for intra-rater reliability	for 5 functional items: supine to sit κ_w=0.66 sit to stand κ_w=0.53 ambulation κ_w=0.48 stair climbing κ_w=0.76 ambulation velocity κ_w=0.78 for overall score ICC=0.82	agreement of combined items for various observers κ_w=0.79-0.90		59

[1] OA pat. = osteoarthritis patients, AP pat. = patients who underwent arthroplasty; [2] Number of ratings per observer per patient (Jones, Iowa) or number of self-assessments per patient (WOMAC); [3] Values of test-retest reliability given for "Likert scale"/"visual analogue scale (VAS)"

Table IV: *Characteristics and results of validity studies*

Joint	Name or abbrev.	Author's	Year	Study population			Investigator	Content and/or construct validity	Ref.
				type[1]	size	age			
Hip	L-ISH	Lequesne	1987	OA pat.	55	not given	"trained	Responsiveness to NSAID treatment (versus placebo, t-test): $p < 0.001$ for L-ISH overall score $p < 0.001$ for investigator's overall opinion $p < 0.01$ for patient's overall opinion $p < 0.01$ for pain level (visual analogue scale) $p < 0.05$ for walking time not significant for abduction, flexion	56
	Mayo	Kavanagh et al.	1985	AP pat.	161	not given	not given	"correlational evidence" compared with Harris: $r > 0.99$	26
	Harris	Bryant et al.	1993	AP pat.	226	not given	not given	factor analysis 3 independent score factors were identified, describing functional activity, hip movement and deformity, and pain	60
	Judet MdA Shepherd Stinchfield Larson Harris Andersson McKee HSS-2 Mayo	Bryant et al.	1993	AP pat.	47	not given	not given	"correlational evidence" 10 hipscores compared with each other (results shown in Table V)	60
Knee	L-ISK	Lequesne	1987	OA pat.	27	not given	"trained"	Responsiveness to NSAID treatment (versus placebo, t-test): $p < 0.025$ for L-ISK overall score $p < 0.006$ for investigator's overall opinion $p < 0.014$ for patient's overall opinion $p < 0.004$ for pain level (visual analogue scale) $p < 0.05$ for time for going up and down a standard flight of stairs not significant for duration of morning stiffness, limitation of flexion, and pain on flexion and extension	56

Table IV (continued): *Characteristics and results of validity studies*

<table>
<tr><th rowspan="2">Joint</th><th rowspan="2">Name or abbrev.</th><th rowspan="2">Author's</th><th rowspan="2">Year</th><th colspan="3">Study population</th><th rowspan="2">Investigator</th><th rowspan="2">Content and/or construct validity</th><th rowspan="2">Ref.</th></tr>
<tr><th>type[1]</th><th>size</th><th>age</th></tr>
<tr><td>Hip/
Knee</td><td>WOMAC</td><td>Bellamy et al.</td><td>1988</td><td>OA pat.</td><td>57</td><td>55-82</td><td>not applicable (self-rating)</td><td>Responsiveness to NSAID treatment after washout period (in Wilcoxon's test): $p < 0.001/0.003^2$, $0.001/0.013^2$ and $0.003/0.002^2$ for pain, stiffness and function, respectively

"correlational evidence" (r):
WOMAC items of pain, stiffness and function (VAS) compared with corresponding Lesquesne items and tenderness according to Doyle[3]
<table>
<tr><th>WOMAC</th><th colspan="3">Lequesne</th><th>Doyle</th></tr>
<tr><th></th><th>pain</th><th>stiffness</th><th>function</th><th>tenderness</th></tr>
<tr><td>pain</td><td>0.39-0.62</td><td>0.04-0.24</td><td>0.36-0.59</td><td>0.36-0.57</td></tr>
<tr><td>stiffness</td><td>0.32</td><td>0.27</td><td>0.35</td><td>0.47</td></tr>
<tr><td>function</td><td>0.32-0.50</td><td>0.01-0.31</td><td>0.36-0.59</td><td>0.28-0.54</td></tr>
</table></td><td>58</td></tr>
<tr><td></td><td></td><td>Bellamy et al.</td><td>1988</td><td>OA pat.</td><td>30</td><td>54-83</td><td>not applicable (self-rating)</td><td>Responsiveness to arthroplasty (in Wilcoxon's test): $p < 0.001$ for all of the subscales of pain, stiffnesss and function

"correlational evidence" (r):
WOMAC items of pain, stiffness and function (VAS) compared with corresponding Lequesne items and tenderness according to Doyle[3]
<table>
<tr><th>WOMAC</th><th colspan="3">Lequesne</th><th>Doyle</th></tr>
<tr><th></th><th>pain</th><th>stiffness</th><th>function</th><th>tenderness</th></tr>
<tr><td>pain</td><td>0.47-0.65</td><td>0.21-0.38</td><td>0.49-0.78</td><td>0.29-0.43</td></tr>
<tr><td>stiffness</td><td>0.22-0.43</td><td>0.32-0.35</td><td>0.13-0.25</td><td>0.07-0.12</td></tr>
<tr><td>function</td><td>0.32-0.48</td><td>0.17-0.51</td><td>0.30-0.56</td><td>0.07-0.49</td></tr>
</table></td><td>61</td></tr>
<tr><td></td><td></td><td>Bellamy et al.</td><td>1991</td><td>OA pat.</td><td>17</td><td>52-65</td><td>not applicable (self-rating-</td><td>factor analysis (PCA) of pain and physical function subscales:
88% and 83% of variance accounted for by factor I, respectively; high factor loadings on each individual pain item (0.92-0.95) and each individual physical function item (0.70-0.97).</td><td>62</td></tr>
<tr><td></td><td>Öberg</td><td>Öberg et al.</td><td>1994</td><td>OA pat.</td><td>105</td><td>46-91</td><td>physical therapist</td><td>factor analysis (varimax-rotation)
5 independent factors identified which were in close agreement with the author's primary categorization (hip impairment, knee impairment, physical disability, social disability, and pain)</td><td>53</td></tr>
</table>

Table IV (continued): *Characteristics and results of validity studies*

Joint	Name or abbrev.	Author's	Year	Study population			Investigator	Content and/or construct validity	Ref.
				type[1]	size	age			
Hip/ Knee	ILAS	Shields et al.	1995	AP pat.	55	34-88	physical therapist	"Predictive validation": Discrimination between patients 2 and 6 days post-operatively (responsiveness index = 0.75) "correlational evidence": compared with Harris: r = –0.86	59

[1] OA pat. = osteoarthritis patients, AP pat. = patients who underwent arthroplasty
[2] Values of validity given for the subgroups of patients treated with isoxicam/piroxicam, respectively
[3] Intervals indicate the range of values for multiple pairwise comparison of items; VAS: visual analogue scale

Table V: *Correlational evidence of 10 hip scores (60).*

Spearman's rank correlation coefficient (rs)

1. Correlation of ordinal ratings[1]

	MdA	Shepherd	Stinchfield	Harris	Andersson	McKee
Judet	0.73	0.49	0.55	0.70	0.64	0.46
MdA		0.52	0.65	0.53	0.66	0.53
Shepherd			0.49	0.64	0.77	0.65
Stinchfield				0.45	0.67	0.66
Harris					0.64	0.49
Andersson						0.71

2. Correlation of numerical ratings[2]

	MdA	HSS-2	Stinchfield	Harris	Andersson	McKee	Mayo	Larson
Judet	0.94	0.83	0.89	0.82	0.76	0.67	0.79	0.83
MdA		0.78	0.85	0.82	0.71	0.73	0.82	0.82
HSS-2			0.93	0.87	0.91	0.79	0.86	0.94
Stinchfield				0.89	0.85	0.75	0.86	0.90
Harris					0.83	0.83	0.89	0.95
Andersson						0.84	0.75	0.89
McKee							0.71	0.84
Mayo								0.93

[1] Expressed in terms of excellent result, good result, or failure
[2] Numerical ratings expressed as the percentage of the maximum possible numerical score

The number of patients studied varied from 17 to 226. Where reported, study participants were between 46 and 91 years old. The clinical background of the investigators was given in only three studies. In the reliability and validity studies by Öberg et al. and Shields et al., observers were physical therapists. Observers with various backgrounds (rheumatologists, geriatrician, medical- and rheuma-registrar) were involved in the reliability study by Jones et al. The number of observers per patient for assessment of inter-rater-reliability was limited to two in all studies. The Jones score and the Iowa score are the only scores for which intra-rater-reliability has been assessed.

Lequesne's L-ISH and L-ISK are the first clinical scores for which reliability and validity studies have been reported (39, 55-57). The inter-rater reliability for overall scores of L-ISH and L-ISK reported by the author did not show any significant systematic variation between observers ($p > 0.05$ in t-tests). The mean deviation of observers was 0.55 and 0.146 for L-ISH and L-ISK on a 24-point scale. Content validity for overall scores of L-ISH and L-ISK has been assessed by predictive validation (measurement of responsiveness to therapy). In a double-blind crossover randomized short-term trial, the difference of patients' health status was measured with these clinical scores between the end of a week active NSAID (nonsteroidal anti-inflammatory drug) therapy and the end of a week placebo therapy. Satisfying responsiveness was reported for both scores ($p < 0.001$ for L-ISH and $p < 0.025$ for L-ISK in t-test for overall index of severity).

Jones' score is a clinical rating index which is mainly used for the assessment of joint inflammation and clinical status of knee OA. A validity study of this score has not been reported to date, and the intra- and inter-rater reliability were tested only for single items of the score. The time interval between repeated measures was one to three hours for intra-rater-reliability and up to five days for inter-rater reliability. In general, intra-rater reliability was found to be satisfactory for all items of the score (range of kappa statistics for various items: 0.54-0.90) with the highest value for inactivity stiffness and the lowest value for synovial swelling. In contrast, inter-rater-reliability was found to be satisfactory only for symptoms (range of kappa statistics: 0.46-0.72 for various symptoms, 0.09-0.35 for various clinical signs).

WOMAC (Western Ontario and McMaster Universities Osteoarthritis Index) is a self-rating score whose early version contains five clinical subscales: pain, stiffness, physical function, social function and emotional function, but the later version contains only three subscales (pain, stiffness and physical function). A test-retest-reliability study and two validity studies (pharmacologic and orthopaedic validation study) of the later version of WOMAC were reported by the authors in 1988 for its three subscales (58, 60). In all of these studies, items were measured on two types of scales: a five level ordinary "Likert scale" (levels 0, 1, 2, 3, 4) and a Visual Analogue Scale ("VA-scale") of length 100 mm with terminal descriptors "none" and "extreme". In the reliability study, the time interval between repeated measures was one week, and test-retest-reliability was higher for the VA-scale than for the Likert scale (except for pain), and for items of pain and physical function ($\tau_c = 0.64$ and 0.72 for VA-scale, respectively) than for items of stiffness ($\tau_c = 0.61$ for VA-scale). Overall the authors concluded that the test-retest-reliability of WOMAC was satisfying.

In the two validity studies of WOMAC subscales, content validity was assessed by predictive validation (measurement of responsiveness to therapy). In the pharmacologic validation study (58), patients' health status was compared before and after 6-week NSAID therapy (isoxicam or piroxicam) using the WOMAC subscale-scores. Adequate responsiveness was reported for both the Likert scale and VA-scale version of WOMAC (for the subscales of pain, stiffness, and physical function, $p < 0.001$, 0.001, and 0.003 in isoxicam subgroup and $p < 0.003$, 0.013, and 0,002 in piroxicam subgroup in Wilcoxon's test, respectively). In the orthopaedic validation study (61), a one group repeated-measures design was likewise employed in patients undergoing total joint arthroplasty for hip and knee OA. Difference of patients' health status was measured with WOMAC subscale-scores between the day before surgery and 6 weeks, 3 months and 6 months after surgical treatment. Here, we focus on comparisons of results before and 6 months after surgery. Satisfying responsiveness was reported in this study (p-values < 0.001 with Wilcoxon's test) for all of the subscales of pain, stiffness, and physical function with both Likert scale and VA-scale version of the test. Construct validity of WOMAC was measured in both the pharmacologic and orthopaedic validation study by way of comparing all items of WOMAC with the items of Lequesne's score and Doyle's tenderness score (comparisons were also made with Bradburn Index of Well Being (63) and the social component of the McMaster Health Index Questionnaire (64), but these results are not reported here). Overall, WOMAC subscale items showed relatively higher levels of correlation with Lesquesne's items probing the same dimensions of health (pain, stiffness and physical function) than with Doyle's items and with Lequesne's items probing different dimensions of health.

Application of principal component analysis to the pain and physical function subscales of WOMAC reported in 1991 (62), supported the contention that scores from items within subscales can be summated into subscale scores, and that there are no reduntant items in the WOMAC inventory.

The score proposed by Öberg et al. is not confined to outcome measurement of osteoarthritis. It was designed to measure lower-extremity dysfunction of any type. The inter-rater-reliability was evaluated separately for the different items of the score. The authors found extremely high inter-rater reliability between two independent physical therapists ($\gamma = 0.99$-1) for all items of the score (53). Validity of Öberg score was assessed by factor analysis (principal component analysis with varimax rotation). The authors found a factor solution which was very close to their primary subgrouping of variables according to clinical knowledge (categories: hip impairment, knee impairment, physical disability, social disability and pain).

Similar to the score proposed by Öberg et al., ILAS (the Iowa Level of Assistance Scale) is also designed to measure lower-extremity dysfunction of any type. In contrast to other scores, ILAS contains only one major clinical component (physical function). Four intensively trained physical therapists were involved in the reliability study, and the intra-rater-agreement was measured 3-6 months apart using videotaped assessment, Overall, good intra-rater reliability ($\kappa_w = 0.79$-0.90 for different observers) and moderate to good inter-rater-reliability ($\kappa_w = 0.48$-0.78 for different items, ICC = 0.82 for overall score) were reported for this score (54). Concurrent validity of ILAS was assessed by comparison to the Harris' hip score. A striking correlation ($r = -0.86$) was reported between the two overall scores, which are inversely coded. Furthermore, responsiveness was measured in a one-group repeated-measures design. The functional status among patients who underwent total joint arthroplasty was measured twice post-operatively. Responsiveness index was used to quantify the functional difference between day 2 and day 6 after surgical treatment. The responsiveness index of the total functional score was about 0.75 with a mean change of 7 points in an overall 30-point scale, indicating high responsiveness of the index for early postoperative changes.

L-ISH, L-ISK, Jones, WOMAC, Öberg and ILAS are the only scores for which reliability has been reported. The validity of another 10 hip scores was assessed by correlational evidence (see Table IV and V). Kavanagh et al. assessed the correlation of the Mayo clinic hip score with the Harris hip score in 1985 (26). A striking correlation ($r > 0.99$) was reported between the two point scores (0-100). But when scores were categorized (good-to-excellent, fair, or poor outcome), a lower proportion of patients were classified as having good-to-excellent outcomes with the Mayo clinic score than with the Harris score. Eight years later, Bryant et al. analyzed the Harris hip score with factor analysis (60). Three independent core factors were identified, describing functional activities, hip movement and deformity, and pain. The authors recommended separate recording of three essential variables (walking distance, hip flexion and pain) to describe three dimensions rather than the use of component indices. Bryant et al. also compared overall scoring of patients by different hip scores (part of the results related to this review are shown in Table V). A wide discrepancy was found between scores if classification by ordinal rating (excellent result, good result, or failure) was employed. But there was striking correlation among the numerical scores expressed as the percentage of the maximum possible value.

DISCUSSION

A large number of scores have been introduced in the past decades that may be used for the clinical outcome measurements of OA. While the majority of them have been developed to assess the outcome of surgical treatment, there is increasing interest in monitoring outcome of other therapeutic interventions, such as drug treatment. High levels of reliability and validity of measurements are basic prerequisites for that purpose. This paper provides a review of studies on the reliability and validity of clinical outcome measurement of hip and knee OA.

Reliability studies were only reported for 6 of the 45 identified clinical rating scores. All of these studies were carried out in the past fifteen years. Obviously, this review could only include published reliability studies. We suspect that additional reliability studies may have been carried out without appearing in the literature. In particular, studies with less favourable results may have remained undetected since investigators and editors may be more reluctant to publish such studies.

While the number of reliability studies reported in the literature is very limited, setting and design of these studies vary widely, making comparisons between results very difficult. Clinical experience and training of observers appear to be very important. In particular, interpretation of clinical signs may strongly depend on clinical specialty and training of observers. For example, very low levels of inter-rater-reliability of measurements of clinical signs were reported for the Jones' score. Five observers involved in the reliability study of the score (a consultant rheumatologist, a consultant geriatrician, a rheumatology senior registrar, a rheumatology registrar and a general medical registrar). The poor result of inter-rater-agreement for clinical signs in this study might reflect a different clinical background and limited training (only half an hour training period) of observers rather than poor reliability of the score itself. Interestingly, intra-rater-reliability, which can be supposed to be less affected by heterogeneity in clinical background, was much higher than inter-rater-reliability for clinical signs. Similarly, inter-rater-reliability of anamnestic evaluation of pain and stiffness which may be less dependent on the clinical background of the observer than clinical signs, was higher than inter-rater-reliability of clinical signs. Extremely high levels of inter-rater-reliability were found in the study by Öberg et al. for all kinds of items between two observers with the same clinical background. Unfortunately, there is only limited information on the conditions under which this excellent agreement has been achieved, such as the training of observers or the time interval between the ratings. Other important factors may be the very detailed and precise description of measurement procedures and categories.

Another difficulty in comparing results is the use of different statistical measures of reliability. Furthermore, the number of study participants was rather small in most reliability studies, and the number of observers per patient for the measurement of inter-rater-reliability was limited to 2 in all studies, leading to imprecise estimates of reliability. Three scores have been assessed for intra-rater-reliability or test-retest-reliability. It appears likely that these types of reliability strongly depend on the time interval between ratings since symptoms and clinical signs are known to vary over time. On the other hand, memorization and warming-up (e.g., for the measurement of the range of joint motion) have to be considered when the time interval is too short. Studies that allow quantitative assessment of those aspects have not been carried out to date.

Validity studies have been reported for 15 clinical scores. With the exception of Lequesne's scores, WOMAC score, Öberg score and ILAS, all of these scores were introduced for quantifying treatment effects of surgical therapies for patients with hip disorders (including osteoarthritis). Responsiveness to (the typically large) intervention effects of surgical treatment like total joint replacement should be a self-evident minimum requirement of such scores. Only the correlation with other scores has been assessed in the validity studies of those scores. The high correlation between the hip scores (with values expressed as the percentage of maximum possible numerical scores) reported by Kavanagh et al. and Bryant et al. demonstrates that these scores measure the same health aspects. On the other hand, the large discrepancy between quantitative judgements (such as excellent, good, or failure) derived from these scores indicates that interpretation of score results vary strongly between scores; such interpretations may partly reflect the specific clinical background and experience of the authors. Many of the scores are commonly used in clinical and epidemiological research, especially in the evaluation of patients after total hip replacement. Neither of them is internationally accepted as "goldstandard". This makes judgement of validity exclusively based on correlational evidence difficult. Similarly, exclusive validation by factor analysis as reported for the Harris score by Bryant et al. and for the Öberg score by its authors only reflects the relationship between theoretical background and the measurement itself.

More comprehensive assessment of validity has been reported for Lequesne's scores, WOMAC score and ILAS, which were introduced with different specific concepts of measurement. The responsiveness of Lequesne's scores and WOMAC score to both drug and surgical treatment effects was demonstrated by predictive validation. The limited correlation between Lequesne's score and WOMAC score reported by Bellamy et al. should not be regarded as evidence against their usefulness in clinical and epidemiological research of hip and knee OA, but as an indication that these two scores measure slightly different aspects of the same diseases. For example, Lequesne's score measures mainly the type of pain and the duration of stiffness. In contrast, WOMAC score which mainly measures the severity of pain and stiffness is more sensitive to change.

Both WOMAC and Lequesne's algofunctional indices are increasingly used as measures of disease activity and outcome in a number of treatment studies, as patient and physician global assessments of patient status and evaluations of activity-related pain and night pain show satisfying validity and reliability (65). Therefore, the proceedings of a consensus conference held under the auspices of the WHO and American Association for Orthopaedic Surgery recommend the use of WOMAC or Lequesne's scores as primary efficacy measures in osteoarthritis treatment studies (65).

ILAS was primarily introduced for outcome measures of physical therapy and found to be responsive to early postoperative changes (discrimination between patients 2 and 6 days post-operatively), which should typically be relatively large. Whether the instrument is responsive to more subtle treatment effects is yet to be determined.

CONCLUSION

Beause of the very limited number and heterogeneous design of the reported reliability and validity studies, it is very difficult to give a definitive answer as to the appropriateness of various scores for clinical and epidemiological research of hip and knee OA. Nevertheless, the following preliminary conclusion may be very cautiously drawn:

1. Despite the different specific concept of measurement of the scores for which reliability studies have been carried out, pain, stiffness and the physical function of lower extremities could be measured with relatively high reliability in all studies, while results are less conclusive for clinical signs.
2. Validity studies of Lequesne's scores and WOMAC score demonstrate that these scores are sufficiently responsive to both drug and surgical treatment effects, while ILAS has so far only been demonstrated to be responsive to early postoperative changes. The different aspects measured with Lequesne's scores and WOMAC score, which have been primarily introduced for assessing hip and knee OA, should be taken into account. But both algofunctional indices are recommended as primary efficacy measures in treatment studies.
3. The commonly used hip scores for assessing surgical treatment are comparable when the percentage of maximum possible numerical scores is used. The large discrepancies when using qualitative judgements (such as excellent, good, or poor outcome) indicate a lack of standardization and make these categories less suitable for clinical and epidemiologic studies.

The most intriguing result of this review, however, is probably the fact that little is known about the reliability and validity of many clinical rating schemes of hip and knee OA to date. This is disquieting since clinical assessment plays a key role in clinical and epidemiological research on these diseases. Properly designed reliability and validity studies are still needed for the majority of commonly employed scores in which much care is devoted to the choice, qualification, and training of observers, number and selection of patients, the time interval between repeat measurements, type of interventions assessed by predictive validation, appropriate statistical analysis, and reporting of results.

Acknowledgement: This work was partly funded by the German Federal Ministry of Research and Technology (BMBF, grand No. 01EF9406). We thank Mrs. R. Deckner and Ms. B. Mack for their help with preparation of this manuscript.

REFERENCES

1. Felson DT. Epidemiology of hip and knee osteoarthritis. Epidemiol Rev 1988; 10: 1-28.
2. Guccione AA, Felson DT, Anderson JJ/ Defining arthritis and measuring functional status in elders: methodological issues in the study of disease and physical disability. Am J Public Health 1990; 80: 945-9.
3. McDowell I, Newell C. Measuring Health: A guide to rating scales and questionnaires. New York, Oxford University Press, 1987.
4. Altman DG. Practical statistics for medical research. London, Chapman & Hall, 1991.
5. Brown MB, Benedetti JK. Asymptotic standard errors and their sampling behavior for measures of association and correlation in the two-way contingency table. Technical Report No. 23, Health Sciences Computing Facility, University of California, Los Angeles, 1976.
6. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol Rep 1996; 19: 3-11.
7. Fleiss JL. Statistical methods for rates and proportions. 2nd ed., New York, Wiley and Sons, 1981.
8. Goodman, LA, Kruskal WH. Measures of association for crossclassficiation. J Am Stat Assoc 1954; 49: 732-64.
9. Guyatt G, Walter S, Norman G. Measuring change over time: Assessing the usefulness of evaluative instruments. J Chron Dis 1987; 40(2): 171-8.
10. Gade HG. A contribution to the surgical treatment of osteoarthritis of the hip joint: a clinical study. Acta Chir Scand 1947; 120: 37-45.
11. Judet R, Judet J. Technique and results with acrylic femoral head prosthesis. J Bone Joint Surg 1952; 34-B: 173-80.
12. Merle d'Aubigné R, Postel M. Functional results of hip arthroplasty with acrylic prothesis. J Bone Joint Surg 1954; 36-A: 451-75.
13. Shepherd MM. Assessment of function after arthroplasty of the hip. J Bone Joint Surg 1954; 36-B: 354-63.
14. Stinchfield FE; Cooperman B, Shea CE. Replacement of the femoral head by Judet or Austin Moore prosthesis. J Bone Surg 1957; 39-A: 1043-58.
15. Larson CB. Rating scale for hip disabilities. Clin Orthop 1963; 31: 85-93.
16. Danielson LG. Incidence and prognosis of coxarthrosis. Acta Orthop Scand (Suppl) 1964; 66: 1-114.
17. Lanzansky MG. A method for grading hips. J Bone Surg 1967; 49-B: 644-51.
18. Harris WH. Traumatic arthritis of the hip after dislocation and acetabular fractures: treatment by mold arthroplasty. J Bone Joint Surg 1969; 51-A: 737-55.
19. Wilson PD. Amstutz HC, Czerniecki A, Salvati EA, Mendes DG. Total hip replacement with fixation by acrylic cement. A preliminary study of 100 consecutive McKee-Farrar prosthetic replacements. J Bone Joint Surg 1972; 54-A: 207-36.
20. Anderson G. Hip assessment: A comparison of nine different methods. J Bone Joint Surg 1972; 54-B: 621-25.
21. Charnley J. The long-term results of low-friction arthroplasty of the hip performed as a primary intervention. J Bone Joint Surg 1972; 54-B: 61-76.
22. McKee GK, Chen SC. The statistic of the McKee-Farrar method of total hip replacement. Clin Orthop 1974; 95-26-33.
23. Lequesne MG, Mery C. European guidelines for clinical trials of new antirheumatic drugs. Eular Bull 1980; 9: 171-5.
24. Dutton RO, Amstutz HC, Thomas BJ, Hedley AK. Tharies surface replacement for osteonecrosis of the femoral head. J Bone Joint Surg 1982; 64-A: 1225-37.

25. Pellici PM, Wilson PD, Sledge CB, Salvati EA, Ranawat CS, Poss R, Calaghan JJ. Long-term results of revision total hip replacement: a follow-up report. J Bone Joint Surg 1985; 67-A: 513-6.
26. Kavanagh BF, Fitzgerald RH. Clinical and roentgenographic assessment of total hip arthroplasty. Clin Orthop 1984; 193: 133-40.
27. Yano H, Sano S, Nagata Y, Tabuchi K, Okinaga S, Seki H, Suyama T. Modified rotational acetabular osteotomy (RAO) for advanced osteoarthritis of the hip joint in the middle-aged person. First report. Arch Orthop Trauma Surg 1990; 109: 121-5.
28. Potter TA, Weinfeld MS, Thomas WH. Arthroplasty of the knee in rheumatoid arthritis and osteoarthritis. As follow-up study after implantation of the McKeever and McIntosh prostheses. J Bone Joint Surg 1972; 54-A: 1-24.
29. Ranawat CS, Shine JJ. Doncondylar total knee arthroplasty. Clin Arthop 1973; 94: 185-95.
30. Larson RL. Rating sheet for knee function. In: Diseases of the Knee Joint. Eds.: Smili, I., Edingburg, Churchill Livingstone, 1974, 28-30.
31. Kettelkamp DB, Thompson C. Development of a knee scoring scale. Clin Orthop 1975; 107: 93-9.
32. Wilson FC, Venters GC. Results of knee replacement with the Walldius prosthesis. An interim report. Clin Orthop 1976; 120: 39-46.
33. Freeman MAR, Todd RC, Cundy AD. The presentation of the results of knee surgery. Clin Orthop 1977; 128: 222-7.
34. Lotke, PA, Ecker ML. Influence of positioning of prosthesis in total knee replacement. J Bone Joint Surg 1977; 59-A; 77-9.
35. Aichroth PM, Freeman MAR, Smillie IS, Souter WA. A knee function assessment chart. J Bone Joint Surg 1978; 60-B, 308-9.
36. Ewald FC. Rheumatoid arthritis. Part III. Surgery of the knee in rheumatoid arthritis. The American Academy of Orthopaedic Surgeons. Instr Course Lect 1979, 28: 285-98.
37. Goldberg VM, Heiple KG, Ratnoff OD, Kurczynski E, Arvan G. Total knee arthroplasty in classic hemophilia. J Bone Joint Surg 1981; 63-A: 695-701.
38. Hungerford DS, Kenna RV, Krackow KA. The porous-coated anatomic total knee. Orthop Clin North Am 1982; 13: 103-22.
39. Lequesne M. Clinical features, diagnostic criteria, functional assessments and radiological classifications of osteoarthritis. Rheumatol 1982; 7: 1-10.
40. Wang WN. Clinical observations of Blauth's total endoprosthesis of the knee joint. Arch Orthop Trauma Surg 1984; 103: 263-8.
41. Weber U, Hackenbroch MH, Zwingers Th. Ergebnisse In: Endoprothetik am Kniegelenk, Derzeitiger technischer und klinischer Stand. Arbeitstagung, Köln 1984. Eds.: Weber U, Hackenbroch MH. (in German) Stuttgart - New York, Thieme, 1985.
42. Matthews LS, Goldstein SA, Kolowich PA, Kaufer H. Spherocentric arthroplasty of the knee. A long-term and final follow-up evaluation. Clin Orthop 1986; 205: 58-66.
43. Merkel KD, Johnson EW. Supracondylar fracture of the femur after knee arthroplasty. J Bone Joint Surg 1986; 68-A: 29-43.
44. Mackinnon J, Young S, Baily RAJ. The St Georg Sledge for unicompartimental replacement of the knee. A prospective study of 115 cases. J Bone Joint Surg 1988; 70-B: 217-23.
45. Hernigou P, Goutallier D. GUEPAR unicopartmental lotus prosthesis for single compartment femorotibial arthrosis. A five to nine-year follow-up study. Clin Orthop 1988; 230: 186-95.
46. Insall JN, Dorr LD, Scott RD, ScottWN. Rationale of the Knee Society Clinical Rating System. Presented at the 4th open Scientific Meeting of the Knee Society, Las Vegas, Nevada, February 12, 1989.
47. Baumgaertner MR, Cannon WD, Vittori JM, Schmidt ES; Maurer RC. Arthroscopic debridement of the arthitic knee. Clin Orthop 1990; 253: 197-202.
48. Raunest J, Löhnert J. Arthroscopic cartilage debridement by excimer laser in chondromalacia of the knee joint. a prospective randomized clinical study. Arch Orthop Trauma Surg 1990; 109: 155-9.
49. Larson KR, Cracchiolo A, Dorey FJ, Finerman GAM. Total knee arthroplasty in patients after patellectomy. Clin Orthop 1991; 264: 243-54.
50. Hofmann AA, Murdock LE, Wyatt RWB, Alpert JP. Total knee arthroplasty Two to four-year experience using an asymmetric tibial tray and a deep trochleargrooved femoral component. Clin Orthop 1991; 269: 78-88.
51. Jones A, Hopkinson N, Pattrick M, Berman P, Doherty M. Evaluation of a method for clinically assessing osteoarthritis of the knee. Ann Rheum Dis 1991; 51: 243-45.
52. Bellamy N. Osteoarthritis - An evaluative index for clinical trials. Msc Thesis. McMaster University, Hamilton, Canada. 1982.
53. Öberg U, Öberg T. Validity and reliability of a new assessment of lower extremity dysfunction. Phys Ther 1994; 74: 861-71.
54. Shields RK, Leo KC, Miller B, Dostal WF, Barr R. An acute physical therapy clinical practice database of outcomes research. Phys Ther 1994; 74: 463-70.
55. Lequesne MG, Samson M. A functional index for hip disease. Reproducibility. Value for discriminating drug's efficacy. 15th International Congress of Rheumatology. Expansion Scientifique française, Paris, 1981: 778-9.
56. Lequesne MG. Indexes of severity for osteoarthritis of the hip and knee. Scand J Rheumatol (Suppl) 1987; 65: 85-9.
57. Lequesne MG, Samson M. Indices of severity in osteoarthritis for weight bearing joint. J Rheumatol (Suppl 27) 1991; 18: 16-8.
58. Bellamy N, Buchanan WW, Goldsmith H, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically-important patient relevant outcomes to anti-rheumatic drug therapy on patients with osteoarthritis of the hip or knee. J Rheumatol 1988; 15: 1833-40.
59. Shields RK, Enloe LJ, Evans RE, Smith KB, Steckel SD. Reliability, validity, and responsiveness of functional tests in patients with total joint replacement. Phys Ther 1995; 7(3): 169-79.
60. Bryant MJ, Kernohan WG, Nixon JR, Mollan RAB. A statistical analysis of hip scores. J Bone Joint Surg 1993; 75-B: 705-9.
61. Bellamy N, Buchannan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study WOMAC: a health status instrument for measuring clinically-important patient-relevant outcomes following total hip or knee arthroplasty in osteoarthritis. J Orthop Rheumatol 1988; 15: 95-108.
62. Bellamy N, Wells G, Campbell J. Relationship between severity and clinical importance of symptoms in osteoarthritis. Clin Rheumatol 1991; 10(2): 138-43.
63. Bradburn NM. The structure of psychological well being. Chicago, Aldine Publishing, 1969.
64. Chambers LW. McMaster health index questionnaire (MHIQ). Hamilton, McMaster University, Department of Clinical Epidemiology and Biostatistics, 1980.
65. Dieppe P. Recommended methodology for assessing the progression of osteoarthritis of the hip and knee joints. Osteoath Cat 1995; 3: 72-6.

Received: 25 March 1996.
Revision-accepted: 30 July 1996.

Correspondence to: Y. SUN, Department of Epidemiology, University of Ulm, Albert-Einstein-Allee 43, D-89081 Ulm, GERMANY.