Osteoarthritis and Cartilage (2009) **17,** 182–187

doi:10.1016/j.joca.2008.06.009

Osteoarthritis and Cartilage

International Cartilage Repair Society

OARSI OSTEOARTHRITIS RESEARCH SOCIETY INTERNATIONAL

Brief report

Comparative evaluation of three semi-quantitative radiographic grading techniques for hip osteoarthritis in terms of validity and reproducibility in 1404 radiographs: report of the OARSI-OMERACT Task Force

L. Gossec M.D.†*, J. M. Jordan M.D., Ph.D.‡, M.-A. Lam M.Sc.§, F. Fang M.Sc.‡, J. B. Renner M.D., Ph.D.‡, A. Davis Ph.D., Senior Scientist, Associate Professor#††, G. A. Hawker M.D., Ph.D.‖¶, M. Dougados M.D.†, J. F. Maillefert M.D., Ph.D.‡‡§§‖‖‖ and For the OARSI-OMERACT task force "total articular replacement as outcome measure in OA"

† *Paris Descartes University, Medicine Faculty, UPRES-EA 4058, APHP, Rheumatology B Department, Cochin Hospital, Paris, France*

‡ *Thurston Arthritis Research Center, University of North Carolina, Chapel Hill, NC, USA*

§ *Department of Medicine, Women's College Hospital, University of Toronto, Toronto, Canada*

‖ *Division of Rheumatology, Department of Medicine, Women's College Hospital, University of Toronto, Toronto, Canada*

¶ *Department of Health Policy, Management and Evaluation, University of Toronto, Toronto, Canada*

Division of Health Care and Outcomes Research, Arthritis and Community Research and Evaluation Unit, Toronto Western Research Institute, Toronto, Ontario, Canada

†† *Departments of Rehabilitation Science and Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada*

‡‡ *Department of Rheumatology, Dijon University Hospital, Dijon F-21078, France*

§§ *University of Burgondy, Dijon F-21079, France*

‖‖‖ *INSERM U887, Dijon F-21079, France*

Summary

Objective: The objective of this work was to compare the measurement properties of three categorical X-ray scoring methods for hip osteoarthritis (OA).

Methods: In data obtained from trials and cohorts, radiographs were evaluated using the Kellgren and Lawrence (KL) grading system, the Osteoarthritis Research Society International (OARSI) joint space narrowing score, and quantitative measurement of joint space width (JSW), analysed as a categorical variable according to Croft and Lane's cutoffs (1.5, 2.5 and 3 mm).
Predictive validity was assessed through logistic regression to predict joint replacement in one database. Construct validity was assessed through logistic regression between pain and function and X-ray stages. Inter-observer and intra-observer reliability were assessed in 50 subjects by weighted kappa. Sensitivity to change was assessed in 50 patients over a 24-month interval, by standardized response mean (SRM).

Results: Radiographs were available from one trial and two cohorts (1404 X-rays). All three methods predicted joint replacement in the trial. Correlation with clinical parameters was low for the three scoring methods, except for the single community-based cohort. Interrater reliability was higher for categorical JSW (kappa, 0.71 vs 0.44 and 0.47 for KL and OARSI, respectively). Intrarater reliability was similar for the three methods (0.79 vs 0.69 and 0.81). Sensitivity to change was higher for categorical JSW than KL and OARSI (SRM, 0.77 vs 0.28 and 0.35).

Conclusion: Categorical JSW has similar validity and higher sensitivity to change than the other categorical scoring techniques in hip OA. These results indicate categorical JSW may be the preferred method to evaluate structural severity in hip OA clinical trials.

Key words: Hip, Osteoarthritis, Progression, Radiography, Validity, Reproducibility, Kellgren–Lawrence, Joint space width.

Introduction

In 2004, an international working group was created under the auspices of Osteoarthritis Research Society International (OARSI) and Outcome Measures in Rheumatology Clinical Trials (OMERACT) in order to elaborate a set of criteria defining theoretical requirement for total joint replacement in knee and hip osteoarthritis (OA), for use in clinical

*Address correspondence and reprint requests to: Dr Laure Gossec, Service de Rhumatologie B, Hôpital Cochin, 27, rue du faubourg Saint-Jacques, 75014 Paris, France. Tel: 33-158412606; Fax: 33-143549256; E-mail: laure.gossec@cch.aphp.fr
Received 12 March 2008; revision accepted 18 June 2008.

trials evaluating potential disease-modifying drugs in OA. It was decided that the domains of pain, physical function and joint structure on radiographs[1] would be combined as a surrogate measure of outcome. As a first step, three working subgroups were constituted, to determine which instrument should be used to evaluate these domains. The work of the "structure group" on knee OA has been reported elsewhere[2]. This article presents the work of the "structure group" on hip OA.

The objective was to examine categorical measures of OA defining structural severity in hip OA in terms of suitability to enter into a composite criterion representing a dichotomous indicator for joint replacement. At the present time, there is no consensus on how to evaluate structural severity on hip radiographs. Indeed, although there is agreement on the radiographic technique (i.e., antero-posterior hip X-ray with feet in internal rotation of 10°)[3–5], the unresolved issue relates to data interpretation, i.e., how to analyze the radiographs. At this time, evaluation of structural degradation in trials uses a quantitative measurement of joint space width (JSW), either at the narrowest point or by mean JSW, by precise measurement using a ruler or caliper, or through computer assisted techniques[6]. However, JSW measurement provides a continuous variable, while the working group aimed to establish a dichotomized outcome (virtual indication for joint replacement yes–no). To this end, it is necessary to categorize or dichotomize the continuous variable JSW, or change in JSW, and cutoffs have been proposed in the literature[7]; or to evaluate the domain structure using a widely used categorical instrument, such as the Kellgren and Lawrence (KL) scoring method[8], or the OARSI joint space narrowing stages[9].

As part of the process of this OARSI-OMERACT initiative, it was necessary to determine the most appropriate categorical measure of structural severity in hip OA. To assess a potential outcome measure, it is necessary to assess its psychometric properties, as defined by the OMERACT filter. The OMERACT filter[10] checks that a potential outcome measure is truthful, i.e., reflects what it is supposed to reflect, and is discriminant, which includes reproducibility, and sensitivity to change, over time, and between different severity stages. The last element in the OMERACT filter refers to feasibility, which relates to time, cost, availability and is not assessed through statistics.

There are some published data regarding the psychometric properties of the different scoring techniques in hip OA[11–13]. However, to our knowledge, studies in which all the properties of the OMERACT filter are compared head-to-head are lacking, making it difficult to choose the most effective scoring system for our purposes.

The objective of this work was to provide such a head-to-head comparison of the different semi-quantitative scoring techniques in hip OA, by assessing their psychometric properties[10].

Patients and methods

DATA SOURCES

A call for data (available databases of hip OA with pelvic X-rays either from trials or cohorts) was sent out to the OARSI-OMERACT group and to leading experts, in order to evaluate predictive and construct validity of hip radiographs[2]. The databases could be trials or cohorts. Ideally, the radiographs had been analysed with all three scoring methods (this was actually the case in the Johnston County database); in all other cases the radiographs had to be available for interpretation. It was also necessary to obtain some clinical data (see below). All available radiographs were analysed; for reproducibility ($n = 50$) and sensitivity to change ($n = 50$) a random sample was analysed.

Ethics approval was obtained where necessary for reinterpretation of the radiographs.

SCORING OF RADIOGRAPHS

For the purposes of this study, at least one radiographic view was analysed for each patient. All available radiographs were analyzed, the radiographs were required to be of sufficient quality to allow interpretation of the joint region. If the magnification was reported JSW could be measured, otherwise only OARSI and KL scores were assessed. In trials, only one hip per individual was analysed, the "index hip" (symptomatic or most symptomatic hip). In cohorts, both hips were analysed and the most severe radiographically was used for analysis.

X-rays were scored three times each[2]: (1) radiologic grade according to the KL classification[8] based on Kellgren's original written description, (2) OARSI grade for joint space narrowing[9], and (3) JSW as a continuous variable (quantitative measurement, in millimeters). Joint space was measured manually at the narrowest point, with a magnifying lens fitted with a graticule. JSW was changed into a categorical variable using previously published values (Croft's cutoffs modified by Lane *et al*)[7]: cutoffs were 1.5, 2.5, and 3 mm.

The reading of the films was performed by rheumatologists who had all undergone training to standardize the X-ray scoring[2]. The readers were blinded to all clinical and questionnaire data.

CLINICAL DATA COLLECTION

Demographic data comprising age, sex, race/ethnicity, and body mass index were recorded. Clinical severity (i.e., pain and functional status) was evaluated in the two cohorts through the subscales of the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)[14], and in the clinical trial by pain visual analog scale (VAS) and Lequesne's algo-functional index[15]. Results were presented after a linear transformation as range 0–100 score for all measures, with higher scores indicating worse pain or function[2].

DATA ANALYSIS

Validity results are presented in the order of the original OMERACT filter[10]. Predictive validity, i.e., which radiographic score best predicts later total joint replacement was assessed in the ECHODIAH database on 484 radiographs, through logistic regression to explain the event "joint arthroplasty" over 3 years (yes/no), by radiographic severity at baseline (X-rays grades, adjusted for age, sex, and BMI). Validity was assessed through the cross-sectional relationship between X-ray stages and pain and function (WOMAC or VAS and Lequesne's index), by logistic regression. In the database obtained from a trial, the baseline data were used for analysis. This assessment of validity will be termed here "construct validity"[2]. Logistic regression analyses were carried out to model symptomatic severity (pain and function), categorised into quartiles, by X-rays grades, adjusted on age, race, sex, and BMI (proportional odds model). Associations between symptomatic severity on each item and radiographic severity for KL, OARSI joint space narrowing and JSW (categorised) were expressed as adjusted odds ratios with 95% confidence intervals (CIs). The lowest radiographic category was used as the reference category.

Intra-observer and inter-observer reliability were assessed in 50 randomly selected subjects issued from the ECHODIAH trial[16] by one (JFM) and two readers (JFM and LG), respectively. Weighted kappas were calculated, as well as intra-class correlation coefficients (ICCs)[17] although ICCs are best adapted to continuous data. Sensitivity to change was assessed on 50 pairs of radiographs issued from the ECHODIAH trial[16] (in different patients from reliability) over a 24-month interval. The films were read with knowledge of the order. Sensitivity to change was assessed by standardized response mean (SRM): mean (month 24–month 0) X-ray scoring change/standard deviation of change. Although SRM was not developed as a measure for semi-quantitative data, it was used here since the assumption of calculations of mean and standard deviation regarding equal intervals was violated to the same extent by each of the outcomes. Statistical significance was set at 0.05.

Statistical analysis was performed using SAS statistical software (SAS Institute Inc., Cary, NC).

Results

DATABASES

In total, three databases were available for analysis. Two databases were issued from cohorts and one from a trial, i.e., a total of 1404 radiographs. In the trial and the Toronto cohort, the diagnosis of OA was based on the American

Table I
Description of databases and patients' characteristics

Origin of data	Study design	Number of patients	Age, years	Gender % female	BMI, kg/m^2	Pain*	Function†	KL (% of patients)			OARSI (% of patients)			Categorical JSW (% of patients)		
			Mean (SD)		Mean (SD)	Median (Range)	Median (Range)	0–1	2	3–4	0–1	2	3	0–1	2	3
Paris, France	ECHODIAH trial, diacerein vs placebo	507	63.0 (7.0)	59.6	25.8 (3.5)	47 (0–96)	47 (6–100)	3	70	27	31	61	8	38	39	23
Johnston County, US	Population-based cohort	735	67.2 (9.5)	65.7	30.4 (6.5)	15 (0–100)	19 (0–100)	62	34	4	97	2	1	88	10	3
Toronto, Canada	Population-based symptomatic cohort	162	69.9 (8.4)	71.0	28.2 (5.3)	40 (0–95)	42 (0–84)	73	14	13	62	33	5	NA		

WOMAC scores are normalized 0–100. Grade 0: ≥3 mm, grade 1: 2.5–3 mm, grade 2: 2.5–1.5 mm, grade 3: <1.5 mm. NA: not available.
*Pain was assessed by WOMAC pain in both cohorts and by VAS in the trial.
†Function was assessed by WOMAC function in both cohorts and by Lequesne's algo-functional index in the trial. All scores are normalised to 0–100.

College of Rheumatology criteria[18]. Characteristics of populations are shown in Table I.

The US Johnston County database is a community-based cohort with symptomatic and asymptomatic OA and individuals without OA[19]. The French ECHODIAH database was obtained during a randomized placebo-controlled trial of diacerein in 507 patients with symptomatic hip OA over a 36-month period[16]. The Canadian Toronto database is issued from a population-based cohort of symptomatic OA patients[20]. Because the magnification factor for the Toronto radiographs was not available, they were analysed according to KL and OARSI scores but JSW could not be calculated.

PREDICTIVE VALIDITY

Prediction of total hip replacement (THR) over 3 years was assessed in the French trial as the data regarding THR were available only in the trial at the time of collection. Of 484 available data, 117 patients (24%) had THR over the 3 years of follow-up. Predictive validity is shown in Table II. All three radiographic scores were associated with later joint replacement, i.e., more patients with the highest radiographic grade underwent THR. There was no superiority in prediction of a particular grading technique.

CONSTRUCT VALIDITY

Construct validity (cross-sectional association with symptoms) is shown in Table III. The analysis of the Johnston County database did not include the more severe radiographic stages, which were observed in too few patients. For the same reason, the analysis of the Paris database did not include the KL and OARSI grades of 0.

No association was observed between X-ray grades and symptoms in the Paris and Toronto databases, except for one paradoxical association in the Toronto database in which patients with KL score of 2 presented with less pain than those with KL of 0. On the contrary, in the Johnston County database, relationships were observed between pain, function and radiographic stages for all scoring systems, the patients with more severe structural degradation suffering from more pain and functional disability. In the three databases, there was no apparent superiority of one scoring system.

RELIABILITY

Reproducibility of readings both intra-reader and inter-reader is shown in Table IV. Interrater reliability tended to be higher for categorical JSW (weighted kappa, 0.71 for categorical JSW vs 0.44 and 0.47 for KL and OARSI, respectively). Intrarater reliability was similar for the three methods (0.79 vs 0.69 and 0.81 for categorical JSW, KL and OARSI, respectively).

SENSITIVITY TO CHANGE

Sensitivity to change was higher for categorical JSW, with the following results expressed as SRMs: 0.77 vs 0.28 and 0.35, for categorical JSW, KL and OARSI, respectively. Sensitivity to change was therefore assessed as high for categorical JSW, and medium for the other techniques, in the context of the trial from which the radiographs were extracted.

Table II
Predictive validity of radiographic scoring methods at baseline for THR in the following 3 years in the ECHODIAH trial

X-ray grade	KL	OARSI	Categorical JSW
Adjusted odds ratio			
			1 vs *0:* 0.95 (0.41−2.20) $P = 0.91$
		2 vs *1:* 1.56 (0.94−2.57) $P = 0.08$	*2* vs *0:* 1.78 (0.86−3.69) $P = 0.12$
	***3* vs *2:* 2.62 (1.67−4.13) $P < 0.0001$**	***3* vs *1:* 4.47 (1.98−10.05) $P = 0.0003$**	***3* vs *0:* 4.02 (1.88−8.56) $P = 0.0003$**
% THR (N/N patients exposed)			
0			15.0 (12/80)
1	0 (0/12)	17.9 (27/150)	14.1 (15/106)
2	20.1 (69/343)	24.3 (73/300)	24.0 (46/192)
3	37.8 (48/128)	50.0 (17/34)	41.1 (44/107)
4	0 (0/1)		

Prediction of THR during the 3-year follow-up, according to baseline X-ray scoring. Grade 0: ≥3 mm, grade 1: 2.5−3 mm, grade 2: 2.5−1.5 mm, grade 3: <1.5 mm. Analyses are adjusted for age, sex and body mass index. Statistically significant results are shown in bold characters. Analyses were not performed for KL grade 0 and 1 and OARSI grade 0 because of an insufficient number of observations.

Discussion

This large study of X-ray grading in hip OA allows us to conclude that although KL, OARSI stages and categorisation of JSW all have similar predictive and construct validity, it appears that categorical JSW is more reproducible and more sensitive to change.

To our knowledge, this is the first study to compare in a large number of X-rays from different patient samples, the psychometric properties for all available categorical scoring techniques. Acknowledged methodology, as recommended by the OMERACT group of experts[10], was applied to compare the radiographic scoring techniques. Three databases were used in order to evaluate construct validity since construct validity is a critical point with conflicting results published in the literature. Thus it was important (1) to increase the power of the evaluation and (2) to obtain data from studies with different designs, different populations, and performed in different countries to allow for better representativity.

However, there are some limits to this work including selection bias (the trial is not population-based) and the somewhat low intereader reproducibility for KL and OARSI stages; however, these data are comparable to published results[12]. Another limitation is related to the noted differences in results across the three databases. In fact, one of the cohorts was not a population-based sample and the trial was of course not population-based. However, this diversity of data sources is both a limitation (heterogeneity) and a strength. Another limitation is that radiographs for sensitivity to change were read without blinding for chronological order; however, the effect of blinding for order is a subject of debate. Finally, the objective of this study was to compare the psychometric properties of the three scoring systems; therefore we did not analyse the data to evaluate the usefulness of hip X-ray in hip OA (e.g., comparison between OA and non-OA patients).

In this study, there was no significant difference between the three radiographic scoring techniques with regards to construct validity, which was low in the trial and the symptomatic Canadian cohort. The observed discordance between clinical and radiographic data is a subject of debate[12,19,21,22]. Several hypotheses can be proposed for explanation, as discussed elsewhere[2]. It should be noted in particular that these results may be explained by recruitment bias[2], as the association does exist in the population-based Johnston County cohort.

Predictive validity (prediction of future joint replacement) was evidenced for all radiographic scoring techniques in the present study (for data based on the clinical trial ECHODIAH only as this was the only database with THR data). Therefore the conclusion regarding prediction of joint replacement is only applicable to patients with OA included in a trial. As reported by Lievense *et al.* in a systematic literature review[23], and confirmed by other studies, several studies on predictive factors of THR in hip OA indicate that baseline radiographic grade is an important predictive factor[24–27]. Rapid radiographic change is also probably important but was not studied here. To our knowledge no comparative study has been performed in the same patients to assess the radiographic scoring techniques comparatively as regards predictive validity.

Interrater reproducibility tended to be higher for categorical JSW. Thus categorical JSW appears more reproducible, including for less experienced readers, which further supports interest in this scoring technique. However, the outcome based on categorical JSW measurement was the only one derived from a continuous measure of JSW, which axiomatically can be expected to be more reproducible than semi-quantitative ratings. Importantly, a distinct finding of the present study is that sensitivity to change was higher for categorical JSW than for the other scoring techniques. This cannot in this case be explained by the reproducibility results. This is an essential aspect of a criterion to be entered into a composite score for use in clinical trials evaluating potential disease-modifying drugs.

The construct and predictive validity results might be considered as validating the inclusion of a structural domain in a set of criteria defining theoretical requirement of total joint replacement in hip OA. The weak correlation between pain and function on the one hand, and structure on the other hand, suggests that the domain structure captures another phenomenon than symptomatic assessment. The satisfying predictive validity of structural assessment suggests that this domain effectively captures a clinically relevant hip OA pathologic feature.

In conclusion, this large study indicates that categorical JSW measurement should be the preferred instrument to evaluate structure in a set of criteria defining theoretical requirement for total joint replacement in hip OA, for use in clinical trials evaluating potential disease-modifying

Table III
Construct validity of the three hip X-ray grading systems: adjusted odds ratios to explain pain and functional status (in quartiles) by X-ray grade

Database	X-ray grade	Pain			Functional status		
		KL	OARSI	Categorical JSW	KL	OARSI	Categorical JSW
Paris trial, $N=507$	2 vs 1	0.87 (0.33–2.28)	1.09 (0.77–1.54)	*1* vs *0:* 0.92 (0.55–1.53)	1.31 (0.50–3.43)	1.14 (0.81–1.62)	*1* vs *0:* 0.77 (0.46–1.29)
	3 vs 1	0.88 (0.33–2.39)	0.85 (0.45–1.61)	*2* vs *0:* 1.14 (0.72–1.82)	1.73 (0.64–4.68)	0.96 (0.51–1.82)	*2* vs *0:* 1.47 (0.92–2.36)
	4 vs 1	0.25 (0.01–4.66)		*3* vs *0:* 1.00 (0.60–1.68)	1.45 (0.10–21.07)		*3* vs *0:* 1.21 (0.72–2.03)
Johnston County, $N=735$	≥1 vs 0	1.62 (0.88–2.98)	**1.37 (1.01–1.87)***	1.05 (0.73–1.52)	1.52 (0.83–2.78)	**1.37 (1.01–1.86)***	1.07 (0.74–1.54)
	≥2 vs 0	**2.22 (1.17–4.20)***		1.57 (0.95–2.60)°	**2.13 (1.13–4.01)***		**1.67 (1.00–2.78)***
Toronto, $N=162$	1 vs 0	1.10 (0.55–2.21)	1.03 (0.29–3.61)	NA	0.72 (0.36–1.43)	1.40 (0.40–4.91)	NA
	2 vs 0	**0.35 (0.13–0.93)***	0.99 (0.26–3.69)		0.66 (0.26–1.71)	1.29 (0.35–4.81)	
	3 vs 0	1.08 (0.31–3.70)	0.55 (0.09–3.36)		1.02 (0.29–3.48)	1.45 (0.24–8.83)	
	4 vs 0	0.54 (0.14–2.11)			1.22 (0.32–4.73)		

X-ray grade is applicable for KL and OARSI, and cat JSW. NA: not available. *, *P* value 0.01–0.05. °, *P* value 0.06–0.08. Pain was assessed by WOMAC pain in both cohorts and by VAS in the trial. Function was assessed by WOMAC function in both cohorts and by Lequesne's algo-functional index in the trial. Associations between symptomatic severity on each item and radiographic severity were evaluated using logistic regression analyses (proportional odds model), with adjustment for age, sex and body mass index. Symptomatic severity (pain and function) was categorised into quartiles. Three radiographic scoring systems were evaluated, i.e., KL, OARSI joint space narrowing and categorized measured JSW (grade 0: ≥3 mm, grade 1: 2.5–3 mm, grade 2: 2.5–1.5 mm, grade 3: <1.5 mm). Associations between symptomatic severity on each item and radiographic severity were expressed as adjusted odds ratios with 95% confidence intervals (CIs). The lowest radiographic category was used as the reference category. The analysis of the Johnston County database did not include the more severe radiographic stages, which were observed in too few patients. The analysis of the Paris database did not include the KL and OARSI grades of 0, due to an insufficient number of patients. Statistically significant results are shown in bold characters.

Table IV
Reproducibility of the three hip X-ray grading systems

	KL		OARSI		Categorical JSW	
	Weighted kappa (95% CI)	ICC (95% CI)	Weighted kappa (95% CI)	ICC (95% CI)	Weighted kappa (95% CI)	ICC (95% CI)
Inter-reader	0.44 (0.26–0.61)	0.64 (0.44–0.78)	0.47 (0.29–0.66)	0.61 (0.41–0.76)	0.71 (0.54–0.87)	0.75 (0.61–0.85)
Intra-reader	0.69 (0.48–0.89)	0.81 (0.69–0.89)	0.81 (0.69–0.94)	0.89 (0.82–0.94)	0.79 (0.67–0.91)	0.90 (0.83–0.94)

Scoring was performed once by two readers (inter-reader reproducibility) and twice by one reader (intra-reader reproducibility) on 50 randomly chosen radiographs.

drugs in OA. Further studies are needed to assess whether a categorical or a dichotomized JSW measurement should be used in such set of criteria, and to establish the threshold(s).

Conflict of interest

The authors have declared no conflicts of interest.

Acknowledgments

This work was supported by unrestricted grants from both scientific societies (OARSI and OMERACT) and pharmaceutical companies: Pfizer, Astra Zeneca, Novartis, and Negma Lerads laboratories. The ECHODIAH trial database was made available by Negma. The Johnston County Osteoarthritis Project data collection was supported by the Centers for Disease Control and Prevention/Association of Schools of Public Health S043 and S3486, and the National Institute of Arthritis, Musculoskeletal and Skin Diseases Multipurpose Arthritis and Musculoskeletal Disease Center 5-P60-AR30701 and Multidisciplinary Clinical Research Center 5-P60-AR49465. The Canadian cohort data collection was supported by grants from the Canadian Institutes of Health Research (formerly, the Medical Research Council of Canada) (MT-12919).
The study sponsors played no role in the study design, in the collection, analysis and interpretation of data; in the writing of the manuscript; or in the decision to submit the manuscript for publication.

References

1. Gossec L, Hawker G, Davis AM, Maillefert JF, Lohmander LS, Altman R, *et al*. OMERACT/OARSI initiative to define states of severity and indication for joint replacement in hip and knee osteoarthritis. J Rheumatol 2007;34:1432–5.
2. Gossec L, Jordan JM, Mazzuca S, Lam M, Suarez-Almazor M, Renner JB, *et al*. For the OARSI-OMERACT task force "total articular replacement as outcome measure in OA". Comparative evaluation of three semi-quantitative radiographic grading techniques for knee osteoarthritis in terms of validity and reproducibility in 1759 X-rays: report of the OARSI-OMERACT task force. Osteoarthritis Cartilage 2008 Jul; 16(7):742–8 [Epub 2008 Apr 15].
3. Altman RD, Fries JF, Bloch DA, Carstens J, Cooke TD, Genant H, *et al*. Radiographic assessment of progression in osteoarthritis. Arthritis Rheum 1987;30:1214–25.
4. Conrozier T, Lequesne MG, Tron AM, Mathieu P, Berdah L, Vignon E. The effects of position on the radiographic joint space in osteoarthritis of the hip. Osteoarthritis Cartilage 1997;5:17–22.
5. Dieppe PA. Recommended methodology for assessing the progression of osteoarthritis of the hip and knee joints. Osteoarthritis Cartilage 1995;3:73–7.
6. http://www.fda.gov/cder/guidance/2199dft.pdf; 1995.
7. Lane NE, Nevitt MC, Genant HK, Hochberg MC. Reliability of new indices of radiographic osteoarthritis of the hand and hip and lumbar disc degeneration. J Rheumatol 1993;20:1911–8.
8. Kellgren JH, Lawrence JS. Radiological assessment of osteoarthrosis. Ann Rheum Dis 1957;16:494–502.
9. Altman RD, Hochberg M, Murphy WA, Wolfe F, Lequesne M. Radiographic atlas for osteoarthritis of the hand, hip and knee. Osteoarthritis Cartilage 1995;3(Suppl A):3–19.
10. Boers M, Brooks P, Strand CV, Tugwell P. The OMERACT filter for outcome measures in rheumatology. J Rheumatol 1998;25:198–9.
11. Reijman M, Hazes JM, Pols HA, Bernsen RM, Koes BW, Bierma-Zeinstra SM. Validity and reliability of three definitions of hip osteoarthritis: cross sectional and longitudinal approach. Ann Rheum Dis 2004;63:1427–33.
12. Reijman M, Hazes JM, Koes BW, Verhagen AP, Bierma-Zeinstra SM. Validity, reliability, and applicability of seven definitions of hip osteoarthritis used in epidemiological studies: a systematic appraisal. Ann Rheum Dis 2004;63:226–32.
13. Ingvarsson T, Hagglund G, Lindberg H, Lohmander LS. Assessment of primary hip osteoarthritis: comparison of radiographic methods using colon radiographs. Ann Rheum Dis 2000;59:650–3.
14. Bellamy N. WOMAC Osteoarthritis Index. A User's Guide. Ontario: London Health Services Centre; McMaster University; 1996.
15. Lequesne M, Samson M, Gérard P, Mery C. Pain-function indices for the follow-up of osteoarthritis of the hip and the knee. Rev Rhum Mal Osteoartic 1990;57: 32S–6.
16. Dougados M, Nguyen M, Berdah L, Mazieres B, Vignon E, Lequesne M, *et al*. Evaluation of the structure-modifying effects of diacerein in hip osteoarthritis: ECHODIAH, a three-year, placebo-controlled trial. Evaluation of the chondromodulating effect of diacerein in OA of the hip. Arthritis Rheum 2001;44:2539–47.
17. Kendall MC, Stuart AIn: The Advanced Theory of Statistics. New York: Hafner Publishing; 1967;Volume 2. pp. 302–4.
18. Altman R, Alarcon G, Appelrouth D, Bloch D, Borenstein D, Brandt K, *et al*. The American College of Rheumatology criteria for the classification and reporting of osteoarthritis of the hip. Arthritis Rheum 1991; 34:505–14.
19. Jordan JM, Helmick CG, Renner JB, Luta G, Dragomir AD, Woodard J, *et al*. Prevalence of knee symptoms and radiographic and symptomatic knee osteoarthritis in African Americans and Caucasians: the Johnston County Osteoarthritis Project. J Rheumatol 2007;34:172–80.
20. Hawker GA, Wright JG, Coyte PC, Williams JI, Harvey B, Glazier R, *et al*. Differences between men and women in the rate of use of hip and knee arthroplasty. N Engl J Med 2000;342:1016–22.
21. Szebenyi B, Hollander AP, Dieppe P, Quilty B, Duddy J, Clarke S, *et al*. Associations between pain, function, and radiographic features in osteoarthritis of the knee. Arthritis Rheum 2006;54:230–5.
22. Duncan R, Peat G, Thomas E, Hay E, McCall I, Croft P. Symptoms and radiographic osteoarthritis: not as discordant as they are made out to be? Ann Rheum Dis 2007;66:86–91.
23. Lievense AM, Bierma-Zeinstra SM, Verhagen AP, Verhaar JA, Koes BW. Prognostic factors of hip osteoarthritis: a systematic review. Arthritis Rheum 2002;47:556–62.
24. Birrell F, Afzal C, Nahit E, Lunt M, Macfarlane GJ, Cooper C, *et al*. Predictors of hip joint replacement in new attenders in primary care with hip pain. Br J Gen Pract 2003;53:26–30.
25. Dougados M, Gueguen A, Nguyen M, Berdah L, Lequesne M, Mazieres B, *et al*. Requirement for total hip arthroplasty: an outcome measure of hip osteoarthritis? J Rheumatol 1999;26:855–61.
26. Hawker GA, Guan J, Croxford R, Coyte PC, Glazier RH, Harvey BJ, *et al*. A prospective population-based study of the predictors of undergoing total joint arthroplasty. Arthritis Rheum 2006;54:3212–20.
27. Gossec L, Tubach F, Baron G, Ravaud P, Logeart I, Dougados M. Predictive factors of total hip replacement due to primary osteoarthritis: a prospective 2 year study of 505 patients. Ann Rheum Dis 2005; 64:1028–32.