Perceptual variation in grading hand, hip and knee radiographs: observations based on an Australian Twin Registry study of osteoarthritis

Nicholas Bellamy, Paul Tesar, Duncan Walker, Alexander Klestov, Kenneth Muirden, Petra Kuhnert, Kim-Anh Do, Louise O'Gorman, Nicholas Martin

Abstract

***Objective*—The radiographic diagnosis of osteoarthritis (OA) in the peripheral skeleton is dependent on the skilled examination of several morphological characteristics of the condition as visualised on plain radiographs. However, the process is perceptual and generally enhanced by comparison against photographic standards. This study assessed the intra-rater and inter-rater reliability of radiologists experienced in reporting hand, hip and knee films derived from a community-based sample when using the photographic atlas recently developed by Burnett *et al*.**

***Methods*—This study was part of a multifaceted diagnostics protocol, evaluating methodological issues, in the conduct of genetic research in osteoarthritis. From a cohort of 118 twin pairs, registered with the Australian Twins Registry (ATR), standard clinical examinations were performed on 74 complete and 11 incomplete pairs of twins over age 50 years, followed by standard AP hand, AP pelvis and AP standing radiographs of the knees. The pairs were selected both to represent twin pairs who had previously self reported a diagnosis of OA, as well as those who had not. Radiologists read the films blind to the original self reported diagnosis and without reference to their pairing. The films were read by comparison against photographic standards and were scored according to specific features. All films were read independently by two consultant radiologists blind to one another's assessments, and selected films were thereafter assigned for rereading. Inter-rater and intra-rater agreement were different for different features, different anatomic areas, and, for the former, were different for the two radiologists.**

***Results*—Inter-rater agreement was different for different anatomic areas, different radiographic features, and the two radiologists. Intra-rater agreement for the presence or absence of OA was as follows: actual observed agreement = 0.79 to 0.97 and 0.83 to 0.98; adjusted κ statistic = 0.58 to 0.94 and 0.67 to 0.96; inter-rater agreement was as follows: actual observed agreement = 0.77 to 0.97; adjusted κ statistic = 0.54 to 0.94. Agreement was generally high in most of the principal target joints for OA: DIP, PIP, 1st CMC, hip and knee.**

***Conclusions*—Although assessor agreement was not perfect, it is concluded that for genetic epidemiology purposes, while duplicate assessments may be advantageous, it is possible for radiographs to be examined accurately by a single experienced assessor. However, for less experienced assessors independent examinations should be made by at least two assessors and either a consensus reached on disparate examinations or an algorithm developed to adjudicate any discrepancies.**

(*Ann Rheum Dis* 1999;**58**:766–769)

University of Queensland, Brisbane, Australia
N Bellamy
P Tesar
D Walker
A Klestov

Queensland Institute of Medical Research, Brisbane, Australia
P Kuhnert
K-A Do
L O'Gorman
N Martin

Melbourne University, Melbourne, Australia
K Muirden

Correspondence to:
Professor N Bellamy, Director of CONROD and Chair of Rehabilitation Medicine, Department of Medicine, University of Queensland, "C" Floor, Clinical Sciences Building, Royal Brisbane Hospital, Herston, Brisbane, Qld 4029, Australia.

Accepted for publication
3 August 1999

There are several standard atlases of radiographs for assessing osteoarthritis (OA) in the peripheral skeleton.[1–4] Each extensively relies on the presence of varying levels of joint space narrowing and osteophyte formation for categorisation. Reading plain radiographs is a perceptual process, the accuracy of which can be enhanced by comparing test films against photographic standards. Even with such standards, there is still some degree of variability between, and, even, within individual observers.[5–7] Variability in reading plain radiographs has important implications for epidemiological research[8 9] as plain radiography is fast, inexpensive, widely available, and, as a consequence, is the most common imaging technique used in epidemiology research. When severe, misclassification can invalidate the results of epidemiological investigation. It is, therefore, essential to quantitate both intra-rater and inter-rater agreement in film reading before performing analyses on epidemiological data sets. We are currently investigating the genetic determination of OA. This study requires a radiographic as well as a clinical assessment of each participating twin pair. To rate radiographs of the hand, knees and hips, we used the standard atlas developed by Burnett *et al*,[3] against which to perform the ratings. Given that this is a relatively recent atlas, and has not been published in the peer review literature, we wished to examine the rater reliability of two experienced radiologists using the atlas in a community-based sample of people, some, but not all of, whom had OA.

Methods

SUBJECTS

The study used the Australian National Health and Medical Research Council Twin Registry.

The ATR is a volunteer registry started in 1978. Twins were recruited through schools, community groups and by media advertising throughout Australia. About 25 000 pairs of twins of all types and ages are registered. This represents approximately 10% of the expected number of twin pairs in Australia. From November 1993 to July 1995, a questionnaire was sent to a cohort of 1178 pairs over age 50 years. Of this group, 1533 people returned completed questionnaires, with complete data being obtained on 602 pairs. The questionnaire was extensive and included items on age, sex, zygosity, birth order, general health, attitudes to health, life events, coping, smoking, alcohol, exercise, emotions, personality, bones and joints, vitamins and sun exposure. In the bones and joints section, registrants were questioned separately about the following: pain, swelling, and stiffness in joints; prior diagnosis of OA (degenerative arthritis), RA, and other forms of arthritis or rheumatism; prior bone fracture or joint injury; radiographs of hands, knees or hips taken in the past five years. In addition, registrants indicated on a homunculus, any joints affected by pain or swelling. Registrants were asked to respond to these questions first considering themselves and then to provide information regarding their co-twin.

From a combination of self reported OA and involvement of target joints for OA without prior history of joint trauma, twins potentially affected by OA were identified. In contrast, those twins not identifying joint problems were categorised as being non-OA. Because it was necessary to examine, as well as take radiographs of, and blood from, study subjects, it was reasoned, given age and condition, that participants would be unlikely to travel more than 50 km from home. As the examinations were to be performed in Brisbane and Melbourne, we invited 63 OA pairs (41 discordant and 22 concordant pairs) and an additional 55 non-OA pairs who met the aforementioned selection criteria to participate. On the day of study, subjects were examined independently by two consultant rheumatologists, blood was taken by venipuncture, a skin mold was made and radiographs taken of hands, knees and hips. Not all twins attended in pairs although many did. They were not examined in any set order and clinical examinations were carried out in separate rooms. No discussion was allowed regarding individual examinations.

RADIOGRAPHS

The radiographs were sent to Royal Brisbane Hospital for central reading by two consultant radiologists (PT and DW). The atlas by Burnett *et al*[3] was used to compare study films against photographic standards. The features depicted in the atlas that were used in the study were as follows: DIP-joint space narrowing (JSN), osteophytes (OP); PIP-JSN, OP; MCP–JSN, OP; 1st CMC–JSN, OP; wrist–JSN, OP; knee JSN, OP, sclerosis (SCL), tibial spiking (SPK); hip–JSN, OP; SCL, cyst (CYS). The gradations permitted by the atlas were as follows: JSN 0–3, OP 0–3, SCL 0/1, SPK 0/1 and CYS 0/1. In addition, a global judgement was made by the radiologist for each joint as to whether there was evidence of OA. The left and right joints were rated separately. Radiographs were read against the Burnett *et al* atlas,[3] completely independently, by the two radiologists. After the initial reading by RAD_1, films were selected for repeat reading by both radiologists. Those films were selected because they represented a cross section of films from normal through mild or moderate to severe OA in the three areas of anatomic interest. Rereading was performed over a period of several months after the initial reading. The films from 70 subjects were assessed four times (that is, twice by RAD_1 and twice by RAD_2); the films from 30 subjects were assessed twice (that is, once by RAD_1 and once by RAD_2); and the films from 59 subjects were read once only by RAD_1.

STATISTICAL ANALYSES

Data analysis was conducted using S-PLUS.[10] Agreement statistics for presence versus absence of OA were calculated for each radiographic feature separately for each joint, as well as for composite features and also for the Radiologist's Global Impression (RGI). As we were interested in case detection, we collapsed all abnormal grades into a single category. However, because the distinction between 0 and 1 in JSN and OP assessment can be difficult, we performed replicate analyses based on splitting the absence versus presence at 0, 1 versus 2, 3 rather than at 0 versus 1, 2, 3 for these two variables. Cohen's κ[11] (unweighted), a statistic of agreement beyond chance has been used frequently in the measurement literature. However, Cohen's κ can be affected adversely by both bias and prevalence. We have, therefore, calculated the Bias Index (BI) and the Prevalence Index (PI), and expressed results using the adjusted κ (κ_{adj}), which takes into account both the BI and PI.[12] These indices respectively reflect the influence of observer bias (BI) and the prevalence of the radiographic feature being rated (PI). We have adopted the following convention when describing the magnitude of BI and PI values: small = <0.30, medium = 0.3–0.6, large = >0.6). The adjusted κ[12] is the preferred estimate of agreement beyond chance and closely parallels the variations in actual observed levels of agreement. For determining inter-rater agreements, the average of all possible RAD_1 *v* RAD_2 combinations was used.

Results

Radiographs of the hand, hips and knees from a total of 159 subjects (74 complete pairs, 11 incomplete pairs) were examined. The demographic characteristics of the members of the 74 complete pairs were as follows: mean age = 59 years (SD = 7), mean weight = 69 kg (SD = 14), mean height = 166 cm (SD = 9). There were 28 MZF (monozygotic females), 14 MZM (monozygotic males), 20 DZF (dizygotic females), three DZM (dizygotic males), four DZFM (dizygotic female 1st born male 2nd born) and five DZMF (dizygotic male 1st born female 2nd born) complete pairs. The corresponding values for the 11 incomplete

Table 1 Observed (OBS) and adjusted (κ_{adj}) agreements between radiologists in rating the presence of OA in 74 complete and 11 incomplete twin pairs participating in the ATR osteoarthritis study (split 0/1,2,3)

		Intra-rater (RAD_1)		Intra-rater (RAD_2)		Inter-rater	
Joint area	Specific features	OBS	κ_{adj}	OBS	κ_{adj}	OBS	κ_{adj}
DIP	JSN	0.88	0.75	0.86	0.72	0.86	0.73
	OP	0.82	0.65	0.91	0.82	0.80	0.60
	Global	0.81	0.62	0.83	0.67	0.78	0.55
PIP	JSN	0.92	0.84	0.91	0.83	0.93	0.85
	OP	0.89	0.77	0.95	0.89	0.89	0.79
	Global	0.85	0.69	0.91	0.81	0.86	0.73
MCP	JSN	1.00	0.99	0.99	0.99	1.00	0.99
	OP	0.97	0.94	0.98	0.96	0.97	0.93
	Global	0.97	0.94	0.98	0.96	0.97	0.94
1st IP	JSN	0.96	0.93	0.98	0.96	0.97	0.94
	OP	0.85	0.69	0.89	0.78	0.88	0.75
	Global	0.82	0.65	0.88	0.77	0.86	0.71
1st CMC	JSN	0.99	0.99	0.99	0.99	0.99	0.99
	OP	0.78	0.56	0.88	0.77	0.84	0.68
	Global	0.79	0.58	0.88	0.75	0.83	0.66
Wrist	JSN	0.95	0.91	0.96	0.93	0.94	0.89
	OP	0.92	0.85	0.96	0.93	0.95	0.91
	Global	0.92	0.85	0.94	0.88	0.93	0.85
Knee	JSN	0.79	0.58	0.97	0.94	0.79	0.58
	OP	0.88	0.77	0.94	0.89	0.91	0.82
	Sclerosis	0.92	0.85	0.96	0.93	0.94	0.89
	Global	0.79	0.58	0.92	0.84	0.77	0.55
Hip	JSN	0.91	0.82	0.94	0.88	0.94	0.87
	OP—acetabular	0.82	0.65	0.75	0.50	0.61	0.22
	OP—femoral	0.98	0.97	0.96	0.91	0.97	0.95
	Sclerosis	0.97	0.95	0.99	0.97	0.98	0.97
	Cysts	0.98	0.97	0.98	0.96	0.98	0.95
	Global	0.86	0.73	0.89	0.79	0.88	0.76

Table 2 Observes (OBS) and adjusted (κ_{adj}) agreement between radiologists in rating the presence of OA in 74 complete and 11 incomplete twin pairs participating in the ATR osteoarthritis study (split 0,1/2,3)

		Intra-rater (RAD_1)		Intra-rater (RAD_2)		Inter-rater	
Joint area	Specific features	OBS	κ_{adj}	OBS	κ_{adj}	OBS	κ_{adj}
DIP	JSN	0.97	0.93	0.99	0.98	0.98	0.96
	OP	0.95	0.91	0.98	0.96	0.97	0.93
PIP	JSN	0.98	0.95	0.99	0.99	0.98	0.97
	OP	0.98	0.96	0.99	0.97	0.99	0.97
MCP	JSN	1.00	1.00	1.00	1.00	1.00	1.00
	OP	1.00	1.00	1.00	0.99	1.00	1.00
1st IP	JSN	0.98	0.96	0.96	0.93	0.98	0.97
	OP	0.89	0.78	0.96	0.93	0.93	0.86
1st CMC	JSN	1.00	1.00	1.00	1.00	1.00	1.00
	OP	0.97	0.94	0.95	0.90	0.95	0.91
Wrist	JSN	0.97	0.94	0.99	0.97	0.98	0.96
	OP	0.98	0.95	1.00	1.00	0.99	0.98
Knee	JSN	0.95	0.90	0.99	0.97	0.97	0.94
	OP	0.94	0.88	0.94	0.89	0.94	0.88
Hip	JSN	0.98	0.97	1.00	1.00	1.00	0.99
	OP—acetabular	0.92	0.85	0.93	0.85	0.93	0.86
	Op—femoral	0.99	0.98	1.00	1.00	1.00	1.00

pairs (eight females and three males) were mean age 55 years (SD = 8), mean weight 74 kg (SD = 24) and mean height 166 cm (SD = 12). In reporting observer agreement, we have provided estimates of the actual observed agreement (OBS) and the adjusted κ (κ_{adj}) for each joint area, each radiographic feature, and for inter-rater agreement (that is, RAD_1 *v* RAD_2) as well as two separate (that is, RAD_1, RAD_2) intra-rater agreements (tables 1 and 2). In table 1, the absence versus presence split is between 0 versus 1, 2, 3, while in table 2 the split is between 0, 1 versus 2, 3. The majority of agreement coefficients are higher in table 2 than in table 1. It should be noted that the global rating is unaffected by the splitting procedure used and therefore is only illustrated in table 1. In addition, the DIP, PIP, IP thumb, 1st CMC, hip and knee joints are considered target joints for OA, while the MCP and wrist joints are non-target joints.

For epidemiological purposes, the consistency with which subjects are categorised as having or not having OA is important. The actual level of observed agreement, based on all coefficients, was as follows: <0.80 = 6%, ⩾ 0.80 = 94%. When based on global coefficients alone, the values were as follows: <0.80 = 17%, ⩾ 0.80 = 83%

Of the global observed values that were <0.80, all were very close to that value (that is, 0.77, 0.77, 0.79, 0.79). Finally, when based on non-global observed coefficients for the 1/2 split, all coefficients exceeded 0.91.

BI values, based on the 0/1 split, were small (1/2 split shown in parentheses): intra-raterRAD_1 −0.19 to 0.03 (−0.03 to 0.02); RAD_2 −0.09 to 0.10 (−0.02 to 0.04); inter-raterRAD_1 *v* RAD_2 −0.03 to 0.31 (−0.06 to 0.05). There was a broad range of corresponding PI values for the 0/1 split (1/2 split shown in parentheses): intra-rater RAD_1 −0.99 to 0.97 (−1.00 to −0.59); RAD_2 −0.99 to 0.99 (−1.00 to −0.69); inter-rater RAD_1 *v* RAD_2 −0.99 to 0.98 (−1.00 to −0.67).

Discussion

The correct categorisation of people is an essential component of genetic epidemiology research. Such categorisations may be based on clinical, radiographic or serological information or on a combination. In this study, we have assessed the extent to which two experienced radiologists agree in grading radiographic features of OA in hip, knee and hand radiographs. In doing so, it is important to recognise that while assessors may agree, they may, nevertheless, both be incorrect. Therefore, the clinical skill and experience of the participating radiologists is paramount. For this study, we chose two senior academic consultant radiologists, each with at least 15 years experience in reading musculoskeletal films. We believe, therefore, that statistics of agreement are appropriate in addressing the issue of the necessity for replicate ratings when performing genetic epidemiology studies. The issue of which statistic to use is contentious, as there are several reliability statistics, each of which examines a different aspect of reliability. For discrete variables, Cohen's κ statistic[11] may be appropriate and is commonly used. However, this statistic can be capricious, and some knowledge of factors that might affect κ is important. In this study, the BI was very small for both pairs of radiologists. This suggests that the effect of bias on κ values was small, different for different joints, and similar for the two pairs of assessors. Bias, therefore, is not a significant problem. In contrast, the PI was large for both pairs of assessors and for most joints. It was large because this was a community-based sample in which only some participants had OA, and was even greater for non-target joints because of the especially low prevalence in those joints, even in affected participants. As a consequence, we believe the adjusted κ[12] provides an estimate of agreement beyond chance that closely parallels the actual

observed agreement and takes into account BI and PI effects. Based on the observed agreement and adjusted κ values, we conclude that for non-target joints agreement is almost perfect.

The observed levels of actual agreements and the adjusted κ values suggest that inter-rater and intra-rater values are high for most radiographic features in all three anatomic areas and in global assessments of the perceived presence/absence of OA. As we were interested in case finding, we have examined closely agreements using a 0/1 split and also a 1/2 split in those features rated on 0–4 scales. The difference between grades 0 and 1 is subtle; it being frequently difficult to differentiate "normal" from "possibly abnormal". Indeed, the assessors may not be entirely convinced whether the 2D image of the joint is indicative of the presence of disease. In contrast, a 1/2 split differentiates the "possibly abnormal" from the "definitely abnormal", and, for case finding, represents a clinically more relevant end point. This perceptive difficulty in interpretation is the most plausible explanation for higher coefficients of agreement observed with the 1/2 split than with the 0/1 split. While the photographic standards were very useful, a number of problems were encountered as follows: (1) Penetration and projection were not uniform throughout the atlas, (2) not all joints of interest in this study were depicted, (3) arrows to a few features were not on the correct joints, (4) there was sometimes a conflict when an abnormality was observed on a radiograph seeming to indicate the presence of OA but not reaching the minimum requirement for being graded at least at grade 1, and (5) some higher grades seemed to be structurally similar.

Overall, we conclude that the agreement between experienced radiologists in rating features of OA in peripheral joints against photographic standards is excellent. These observations suggest that, for genetic epidemiology purposes, such ratings could be performed by a single similarly skilled individual. The requirement for only a single assessor has important time and cost implications for such studies. Whether a rheumatologist could perform the ratings at acceptable levels of intra-rater and inter-rater (rheumatologist/radiologist) agreement is the subject of an ongoing investigation.

1 Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthrosis. Ann Rheum Dis 1957;16:494–501.
2 Kellgren JH. *The epidemiology of chronic rheumatism: atlas of standard radiographs of arthritis*. Vol 2. Oxford: Blackwell Scientific, 1963.
3 Burnett S, Hart DJ, Cooper C, Spector TD. *A radiographic atlas of osteoarthritis.* London: Springer-Verlag, 1994.
4 Altman RD, Hochberg M, Murphy WA Jr, Wolf F, Lequesne M. Atlas of individual radiographic features in osteoarthritis. *Osteoarthritis Cartilage* 1995;3 (suppl A):3–70.
5 Scott WW Jr, Lethbridge-Cejku M, Reichle R, Wigley FM, Tobin JD, Hochberg MC. Reliability of grading scales for individual radiographic features of osteoarthritis of the knee. The Baltimore longitudinal study of aging atlas of knee osteoarthritis. Invest Radiol 1993;28:497–501.
6 Hart DJ, Spector TD, Brown P, Wilson P, Doyle DV. Silman AJ. Clinical signs of early osteoarthritis: reproducibility and relation to x-ray changes in 541 women in the general population. Ann Rheum Dis 1991;50:467–70.
7 Cooper C, Cushnaghan J, Kirwan J, Dieppe PA, Rogers J, McAlindon T, *et al*. Radiographic assessment of the knee in osteoarthritis. Ann Rheum Dis 1992;51:80–2.
8 Spector TD, Hart DJ, Byrne J, Harris PA, Dacre JE, Doyle DV. Definition of osteoarthritis of the knee for epidemiological studies. Ann Rheum Dis 1993;52:790–4.
9 Hart DJ, Spector TD. Radiographic criteria for epidemiologic studies of osteoarthritis. J Rheumatol 1995;22 (suppl 43):46–8.
10 *S-Plus guide to statistical and mathematical analysis.* Version 3.3. Seattle, USA StatSci, A division of Mathsoft Inc, 1995.
11 Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960;20:37–47.
12 Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993;46:423–9.