Ann. rheum. Dis. (1957), **16**, 494.

RADIOLOGICAL ASSESSMENT OF OSTEO-ARTHROSIS

BY

J. H. KELLGREN AND J. S. LAWRENCE

From the Rheumatism Research Centre, University of Manchester, and Empire Rheumatism Council Field Unit

In a previous study (Kellgren and Bier, 1956), three sets of *x* rays of the hands were used to assess inter- and intra-observer differences in interpreting changes of rheumatoid arthritis. Wide disagreement between observers was found, and it was concluded that, to ensure maximum uniformity in grading *x* rays in field surveys and therapeutic trials, all readings should be made by the same observer, preferably at a single session. The advisability of having a set of standard reference films was also considered.

In the present study an attempt has been made to assess observer difference in reading *x* rays for osteo-arthrosis.

The term osteo-arthrosis, as used in this paper, is synonymous with osteo-arthritis and with degenerative joint disease affecting diarthrodial joints. Degenerative change in the synchondroses has not been included. Thus, in the spine, degenerative disease of the apophyseal joints is considered, but not spur formation on the bodies of the vertebrae or narrowing of the intervertebral disks. No separation has been made between local and primary generalized forms of osteo-arthrosis in this study, nor have those forms of osteo-arthrosis secondary to trauma been considered separately.

The following radiological features were considered evidence of osteo-arthrosis:

(1) The formation of osteophytes on the joint margins or, in the case of the knee joint, on the tibial spines.
(2) Periarticular ossicles; these were found chiefly in relation to the distal and proximal interphalangeal joints.
(3) Narrowing of joint cartilage associated with sclerosis of subchondral bone.
(4) Small pseudocystic areas with sclerotic walls situated usually in the subchondral bone.
(5) Altered shape of the bone ends, particularly in the head of femur.

This study is part of a survey of rheumatic disease in the population of Leigh in Lancashire. A series of *x* rays of eleven joints in 85 persons chosen at random from those between the ages of 55 and 64 were read for osteo-arthrosis by two observers: first together so that agreed standards for grading could be determined, and then separately after an interval of time.

The data so obtained have first been used to determine inter-observer difference. Subsequently one observer read the sample a third time to assess intra-observer difference. The interval between the combined and the independent readings was 2 years and the third reading was made one month after the independent readings. All observations were made without knowledge of the symptoms or clinical state or of the previous readings.

As in earlier surveys (Kellgren and Lawrence, 1952; Lawrence, 1955), osteo-arthrosis was divided into five grades as follows:

None	(0)
Doubtful	(1)
Minimal	(2)
Moderate	(3)
Severe	(4)

Grade 0 thus indicated a definite absence of *x*-ray changes of osteo-arthrosis, and Grade 2 that osteo-arthrosis was in our opinion definitely present though of minimal severity. The grading for groups of joints, as for example the distal interphalangeal joints of the hands, indicated the severity in the most affected joint; similarly, though a separate grading was sometimes given for each knee, the worst affected knee was subsequently used for the compilation of results.

In survey work only a limited number of films can be taken, partly for financial reasons and partly

because of the importance of avoiding radiation hazard. For this reason it is seldom possible to take more than one view of each joint. The gradings for each joint in this study are, therefore, based on a single *x* ray view, as follows:

Hands - -	Postero-anterior
Cervical spine -	Lateral
Lumbar spine -	Lateral
Hips - -	Antero-posterior
Knees - -	Antero-posterior
Feet - - -	Antero-posterior

Figs 1-8 show standard examples of Grade 1-4 severity for the distal interphalangeal, proximal interphalangeal, metacarpophalangeal, and first carpometacarpal joints of the hands, and for the wrists, cervical spine, hips, and knees. The lumbar spine has been omitted because of difficulties of reproduction. Copies of these standard sets of radiographs are in preparation and will be available for the use of others. Eight groups of joints (distal interphalangeal, metacarpophalangeal, first carpometacarpal, wrist, cervical spine, lumbar spine, hip, and knee) have been chosen for detailed study.

Grade 1. Grade 2. Grade 3. Grade 4.

Fig. 1.—Osteo-arthrosis of distal interphalangeal joint.

Grade 1. Grade 2. Grade 3. Grade 4.

Fig. 2.—Osteo-arthrosis of proximal interphalangeal joint.

Fig. 3.—Osteo-arthrosis of metacarpophalangeal joint.

Inter-Observer Error

In Table I the readings of the two observers on the distal interphalangeal joints of the fingers are compared. It will be noted that Observer A grades more films as showing moderate or severe changes, fewer as minimal or doubtful. The numbers read as having no osteo-arthrosis are almost identical, but in a third of the cases the films so graded do not relate to the same persons. Thus, although the numbers in which definite osteo-arthrosis (Grades 2-4) is diagnosed are closely similar (55 for Observer A and 54 for Observer B), there is considerable disagreement as to detail.

TABLE I

GRADING OF OSTEO-ARTHROSIS IN DISTAL INTERPHALANGEAL JOINTS BY OBSERVERS A AND B

	B1					
Total	20	11	42	10	2	85
4	—	—	1	2	2	5
3	—	—	9	7	—	16
2	3	6	24	1	—	34
1	4	2	3	—	—	9
0	13	3	5	—	—	21
Score	0	1	2	3	4	Total

A

Correlation coefficient r = 0·73
Standard error = 0·11

Table II shows the readings by each observer of osteo-arthrosis in the metacarpophalangeal joints of the hands, and here again Observer A grades the changes higher than Observer B. This applies in all grades, so that the conclusions on prevalence are very different, definite osteo-arthrosis being diagnosed twice as often by Observer A.

TABLE II

GRADING OF OSTEO-ARTHROSIS IN METACARPO-PHALANGEAL JOINTS BY OBSERVERS A AND B

	B1					
Total	51	20	12	2	—	85
4	—	—	—	1	—	1
3	—	1	1	1	—	3
2	6	8	9	—	—	23
1	16	7	2	—	—	25
0	29	4	—	—	—	33
Score	0	1	2	3	4	Total

A

Correlation coefficient r = 0·66
Standard error = 0·11

Data for the first carpometacarpal joints, wrists, cervical spine, dorso-lumbar spine, hips, and knees were studied in the same way. In each of these, except the wrist, Observer A read higher than Observer B in Grades 2-4. In the wrist, the agreement was so slight that it might well have been due to

Fig. 4.—Osteo-arthrosis of first carpometacarpal joint.

chance, and it was evident that for this joint the observers used quite different criteria. The difference was found to arise chiefly when severe rheumatoid arthritis was also present. The correlation coefficient for the readings on these joints was as follows:

First carpometacarpal joint	0·78
Wrist	0·10
Cervical spine	0·57
Dorso-lumbar spine	0·52
Hips	0·40
Knees	0·83

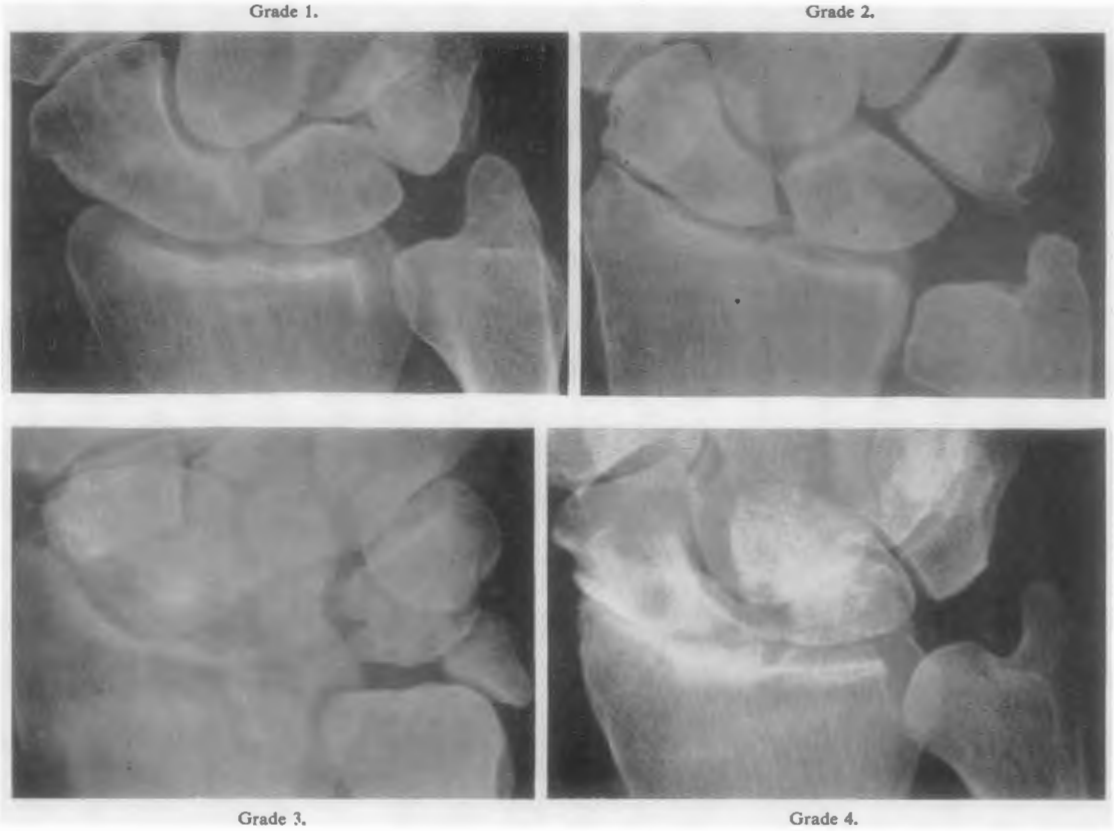

Fig. 5.—Osteo-arthrosis of wrist

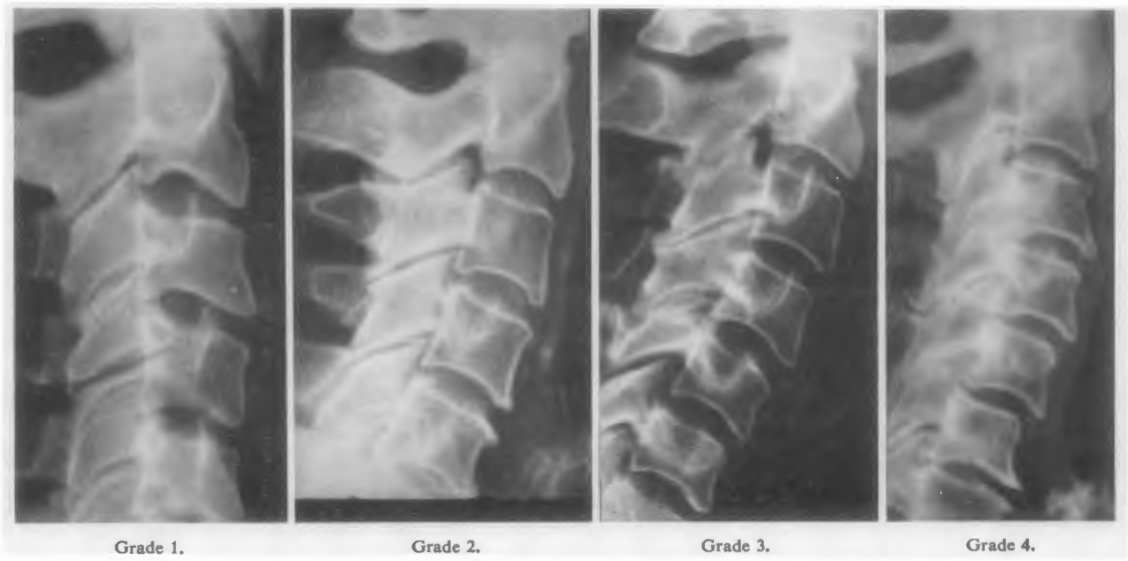

Fig. 6.—Osteo-arthrosis of cervical spine

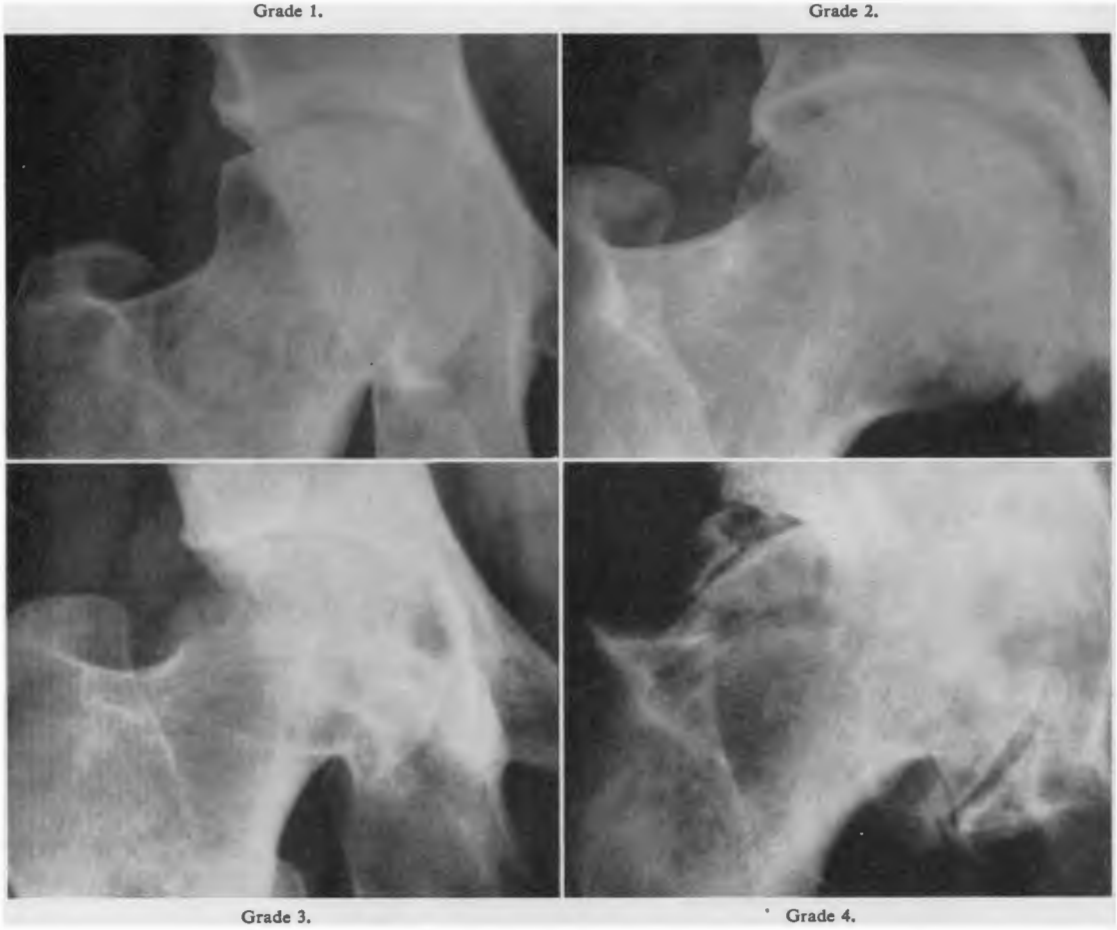

Fig. 7.—Osteo-arthrosis of hip.

Intra-Observer Error

A comparison was made between two readings made by the same observers on the same group of joints. In the distal interphalangeal joints of the fingers there was a greater agreement ($r=0{\cdot}81$) than between the two observers on the same series of joints, particularly in the more severe grades. More joints, however, were graded "1" (minimal or doubtful) in the second reading and fewer were graded "0". Similarly, in the other joints, there was much closer agreement between two readings by the same observer than between two readings by separate observers, and though there was some disagreement on individual films read twice by the same observer the estimated prevalences did not differ significantly. The correlation coefficients were as follows:

Metacarpophalangeal joints	0·88
First carpometacarpal joint	0·81
Wrist	0·62
Cervical spine	0·66
Dorso-lumbar spine	0·42
Hips	0·75
Knees	0·83

Combined Readings

The combined readings made in 1954 by Observers A and B in consultation are compared with the individual readings in Table III (overleaf). In all joints, except the distal interphalangeal joints, wrists, and lumbar spine, the combined reading gives prevalence values between those of the individual readings. In the distal interphalangeal joints and wrists the individual readings show higher values than the combined reading, and in the lumbar

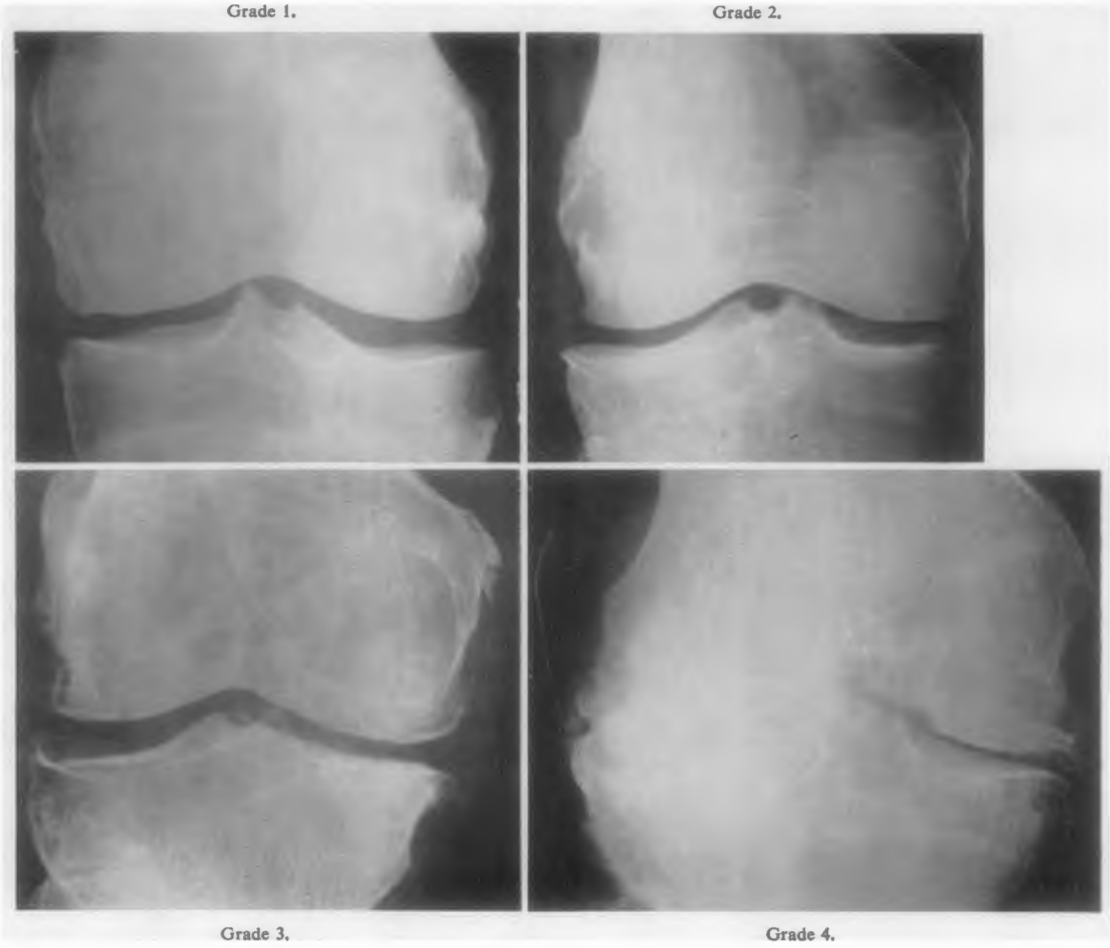

Fig. 8.—Osteo-arthrosis of knee.

spine Observer A's reading is identical and Observer B's reading is lower. Observer A's readings in all instances correlate better with the combined reading than Observer B's, and indeed in all joints, except the hips, Observer A's readings correlate with the combined readings more closely than those of Observer B with his own (B1 and B2). Owing, however, to the fact that Observer A tends to read higher than the

TABLE III
CORRELATION OF FOUR OBSERVER DIAGNOSIS OF OSTEO-ARTHROSIS (GRADES 2-4)

Observer		D.I.Ps.			M.Ps.			C.M.Cs.			Wrists			Cervical Spine			Dorso-Lumbar Spine			Hips			Knees			All Joints			
1	2	1	2	r	1	2	r	1	2	r	1	2	r	1	2	r	1	2	r	1	2	r	1	2	r	1	2	Mean	Deviation from Mean (per cent.)
A	B1	55	54	0 73	27	14	0·66	36	19	0·78	4	5	0·10	36	12	0·57	20	14	0·52	13	3	0·40	47	32	0·83	288	153	221	±31
B1	B2	54	57	0·77	14	15	0·73	19	25	0·77	5	3	0·62	12	9	0·66	14	16	0·42	3	4	0·75	32	40	0·83	153	169	161	± 5
A	A+B	55	48	0·85	27	17	0·75	36	30	0·88	4	1	0·74	36	20	0·67	20	20	0·47	13	6	0·66	47	37	0·87	288	179	234	±23
B1	A+B	54	48	0·81	14	17	0·64	19	30	0·80	5	1	0·18	12	20	0·57	14	20	0·42	3	6	0·47	32	37	0·86	153	179	166	± 8

combined reading the prevalence values show greater differences between Observer A and the combined reading than between B1 and B2. If all joints in which osteo-arthrosis (Grades 2-4) was diagnosed are considered together, the combined reading with a total of 179 joints is closer to the mean value of 195 joints than any of the individual readings.

Conclusions

It is clear that, although significant agreement was found between two observers for the grading of osteo-arthrosis in all joints except the wrist, the influence of personal bias may result in a very different assessment of severity and prevalence in group studies, so that in certain joints, for example the hip joint, one observer may read four times as much definite osteo-arthrosis. In other joints, *e.g.* the distal interphalangeal joints of the fingers, much closer agreement is likely to be found, but it would appear that deviations of ±31 per cent. from the average value may be expected in osteo-arthrosis in general. In individual joints, differences of two to three times in the prevalence of osteo-arthrosis between population groups assessed by different observers might well be due entirely to the observer effect on grading.

If, on the other hand, the *x* rays from population groups are all read by the same observer, so that only intra-observer error is involved, the differences in prevalence from this cause are likely to be less striking, the error in this series being of the order of ±5 per cent. It should be observed, however, that these two sets of readings were made at an interval of only one month. Had they been made at a greater interval, greater differences would no doubt have been encountered. When readings are made by one observer they will of course be subject to observer bias which may result in a high or low prevalence in all groups of *x* rays read by that observer. This will not interfere with the comparison between population groups but may give a generally distorted view of prevalences. This distortion may be overcome by a combined reading made by two observers in consultation. In this series the combined reading approximated more closely to the mean value for all readings of all joints than any of the individual readings, and this, when practicable, would appear to be the method of choice. Unfortunately it is seldom possible to arrange for the same two observers to be available to read all the *x* rays, and it is thus necessary to fall back on the plan of having one observer read every *x* ray from all groups which it is desired to compare. Probably the most satisfactory working arrangement, where, for example, it is desired to compare the prevalence of osteo-arthrosis in several localities studied by different research groups, is to exchange *x* rays, so that an observer from each team may read all the *x* rays taken at every locality. The final result could then be expressed as an average of all the readings.

Summary

A series of 510 *x* rays from 85 persons in the age group 55-64 chosen at random from an urban population was graded for osteo-arthrosis by two observers on four occasions to determine the extent of observer difference.

Standard films for four grades of osteo-arthrosis for each of eleven joints were chosen.

A significant correlation between the two observers was obtained for all joints except the wrist. The estimates of prevalence, however, varied widely because of the cumulative effect of observer bias (±31 per cent.). It is concluded that comparison of prevalence estimates by different observers could have little value in population studies.

Two readings by the same observer gave only a slightly better correlation on the reading of individual *x* rays, but by excluding observer bias they gave a much closer estimate of prevalence (±5 per cent.). These two readings, however, differed substantially from the mean value for all readings (−8 per cent. and −17 per cent.).

A combined reading by two observers reduced the influence of personal bias and differed little from the mean value (−3 per cent.).

It is suggested that, where possible, in all population studies which it is desired to compare, the *x* rays should be read by the same observer or preferably by two observers in consultation.

REFERENCES

Kellgren, J. H., and Bier, F. (1956). *Annals of the Rheumatic Diseases*, **15**, 55.
—— and Lawrence, J. S. (1952). *Brit. J. industr. Med.*, **9**, 197.
Lawrence, J. S. (1955). *Ibid.*, **12**, 249

Evaluation radiologique de l'ostéo-arthrose

RÉSUMÉ

Une série de 510 clichés radiographiques de 85 personnes âgées de 55 à 64 ans, choisies au hasard dans une population urbaine a été classée au point de vue d'ostéo-arthrose par deux observateurs, en quatre occasions, pour déterminer l'écart inter-observateur d'interprétation.

On choisit des clichés standard de quatre degrés d'ostéo-arthrose pour chacune de onze articulations.

Une corrélation significative entre les deux observateurs fut obtenue pour toutes les articulations, sauf le poignet. Les determinations de fréquence, cependant, variaient considérablement en raison de l'effet cumulatif de l'écart inter-observateur (±31%). On conclut que la comparaison des fréquences determinées par des observateurs différents ne pouvait avoir que peu de valeur dans les études de population.

Deux lectures par le même observateur donnaient une corrélation de clichés individuels à peine meilleure, mais en éliminant l'écart inter-observateur on obtenait une fréquence plus précise (±5%). Ces deux lectures différaient cependant largement de la valeur moyenne de toutes les lectures (−8% et −17%).

La lecture combinée par deux observateurs réduisait l'influence des facteurs personnels et différait peu de la valeur moyenne (−3%).

On suggère que, quand cela est possible, dans toutes les études portant sur une population, dans un but de comparaison, les clichés radiographiques soient lus par le même observateur, ou mieux, par deux observateurs travaillant ensemble.

Valoración radiológica de la ósteo-artrosis

SUMARIO

Una serie de 510 radiografías de 85 sujetos de 55 a 64 años de edad, tomados al azar en una población urbana, fué clasificada desde el punto de vista de ósteo-artrosis por dos observadores, en cuatro ocasiones, para determinar la divergencia de interpretación entre observadores.

Se eligieron clisés standard de cuatro grados de ósteo-artrosis para cada de once articulaciones.

Se obtuvo una correlación significativa entre los dos observadores para todas las articulaciones, excepto la muñeca. Las determinaciones de frecuencia, sin embargo, variaban considerablemente a causa del efecto cumulativo de la divergencia entre observadores (±31%). Se concluye que la comparación de frecuencias determinadas por observadores diferentes tiene poca valor en los estudios de población.

Dos lecturas por el mismo observador daban una correlación de clisés individuales un poquito mejor, pero al eliminar la deferencia entre observadores, se obtenía una frecuencia mucho más precisa (±5%). Sin embargo ambas lecturas diferían considerablemente del valor medio de todas las lecturas (−8% y −17%).

La lectura junta por dos observadores reducía la influencia de los factores personales y se distinguía poco del valor medio (−3%).

Se sugiere que, siempre que eso es posible, en todas las investigaciones de una población, con el fin de comparar, las radiografías sean leídas por el mismo observador o, mejor aún, por dos observadores juntos.