RADIOLOGICAL SIGNS OF RHEUMATOID ARTHRITIS

A STUDY OF OBSERVER DIFFERENCES IN THE READING OF HAND FILMS

BY

J. H. KELLGREN

WITH THE ASSISTANCE OF FRIDA BIER

From the Rheumatism Research Centre, University of Manchester

(RECEIVED FOR PUBLICATION JANUARY 2, 1956)

Radiological changes in the hands and feet have long been considered a useful aid in the diagnosis of rheumatoid arthritis. In a recent study, Fletcher and Rowley (1952) defined these changes in great detail. They confirmed the generally accepted opinion as to the value of surface and pocketed articular erosions, but found osteoporosis difficult to assess even with the help of a control bone as advocated by Steven (1947). Alterations in joint space, though common, may also be difficult to assess.

Defined gradings of these x-ray changes are being used as diagnostic aids in epidemiological studies of the prevalence of rheumatoid arthritis in different populations (Miall, Caplan, Cochrane, Kilpatrick, and Oldham, 1953; Miall, 1955) and similar radiological gradings have been used in therapeutic trials.*

It has therefore become necessary to know more about the reproducibility of such x-ray gradings by any given individual, and also the degree of inter-observer variance in the interpretation of individual films, since these factors have been shown to be important in the interpretation of chest films (Fletcher and Oldham, 1949; Yerushalmy, Harkness, Cope, and Kennedy, 1950; Yerushalmy, Garland, Harkness, Hirschshaw, Miller, Shipman, and Zwerling, 1951; Cochrane, Campbell, and Stein, 1949; Cochrane and Garland, 1952). The results reported by Miall (1953, 1955), however, suggest that the observer error in reading x rays of the hands and feet for rheumatoid arthritis may not be very great.

* Two trials comparing cortisone and aspirin therapy in rheumatoid arthritis, organized by the Joint Committee of the Medical Research Council and Nuffield Foundation on Clinical Trials of Cortisone, ACTH, and Other Therapeutic Measures in Chronic Rheumatic Diseases (1954, 1955), and by the Empire Rheumatism Council (1955) respectively.

Materials and Method

Three sets of x rays have been studied. The first comprises 192 films of the hands obtained from sixty patients with early rheumatoid arthritis who were participating in a comparative study of the value of cortisone and aspirin therapy organized by the Joint Committee of the Medical Research Council and Nuffield Foundation (1954, 1955). In this group there were 24 males and 36 females with ages ranging from 17 to 59 years. The differential sheep cell agglutination test (D.A.T.), done by the method of Rose, Ragan, Pearce, and Lipman (1948), or that of Ball (1950), was positive in nearly three-quarters of the patients and all films were taken while the duration of the disease was less than 3 years. These films were read independently by six different observers with considerable experience in the diagnosis of rheumatoid arthritis: four clinicians (Observers I, II, III, and IV) and two radiologists (Observers V and VI). Osteoporosis and erosions which included narrowing of joint space were recorded separately and graded as follows:

- None (0),
- Doubtful (1),
- Slight (2),
- Moderate (3),
- Severe (4).

In addition, one of the clinicians (Observer I) read the films a second time, using a single but similar grading for rheumatoid change as a whole, incorporating porosis, erosions, and alterations in joint space.

The second set of films was taken from a sample of the town population of Leigh, Lancs, including only those in the age group 55-64 years, who were x-rayed irrespective of symptoms. There were originally 173 men and 207 women in the sample, but only those with some evidence suggestive of rheumatoid disease either clinical, radiological, or serological were included in this study of x-ray readings. This gave 126 films of the hands from 41 men and 85 women, the D.A.T. being positive in nineteen of these individuals. These films were read

three times by the clinician (Observer I) who carried out the double reading in the first set of films.

The third set of films comprised 259 hands taken from a sample of an urban population in South Wales in which only individuals with clinical polyarthritis were x-rayed. This sample included all age groups and was derived from 88 males and 171 females. These films were read twice by Observer I, but there was an interval of over one year between the first and second reading.

Observer I has so far been mainly responsible for reading films in the rheumatological survey work being carried out by the Pneumoconiosis Research Unit and the Manchester Rheumatism Centre. It was therefore particularly desirable to compare this observer's gradings with those of his clinical and radiological colleagues, and the material has been analysed mainly from this point of view.

Results

Therapeutic Trial.—In this study 192 hand films from sixty patients with early rheumatoid arthritis were read independently by six different observers. The total scores (Table I) show that there are considerable discrepancies. The radiologists recorded more films as normal, while the clinicians recorded more films as abnormal and tended more often to rate these changes as moderate or severe in degree, particularly as regards porosis. This last discrepancy may be partly due to the radiologists' greater awareness of the technical causes of apparent porosis.

In recording erosions, the main differences appear to be due to language—thus Observer IV graded erosions as moderate or severe in 53 instances, whereas Observer V graded erosions as moderate in only thirteen instances and severe in none. On the other hand, the total number of films considered abnormal on account of erosions varies less, the range being from 81 for Observer V to 105 for Observer III.

In recording porosis, the differences are even greater and are not reduced by combining all grades of abnormality. The total number of films considered abnormal on account of porosis range from 153 for Observer IV to 59 for Observer V.

These films were taken by various techniques and only a proportion of them included a control bone to facilitate estimates of porosis, but separate consideration of the gradings of the 55 films with control bones showed practically as much disagreement; the assessment of porosis was exceptionally difficult in many of these films owing to great variations in radiographic technique.

The differences of opinion over the assessment of porosis in individual films were also very great, as illustrated in Table II where Observers IV and V, the two who disagree most, are compared.

TABLE II
THERAPEUTIC TRIAL—POROSIS

Observer V							Observer IV
Total	72	61	51	6	2	192	
4	—	2	8	3	2	15	
3	10	31	37	3	—	81	
2	29	22	6	—	—	57	
1	18	3	—	—	—	21	
0	15	3	—	—	—	18	
Score	0	1	2	3	4	Total	

It will be noted that disagreement of three grades occurred in 12 and of two grades in 68 of the 192 films. It is therefore apparent that recording estimates of porosis based upon inspection of films can be of only limited value.

TABLE I
THERAPEUTIC TRIAL

Sign	Grade	Observers					
		Clinicians				Radiologists	
		I	II	III	IV	V	VI
Erosion	0 = None	62	50	39	49	59	56
	1 = Doubtful	41	50	48	39	52	50
	2 = Slight	51	43	53	51	68	57
	3 = Moderate	32	36	47	45	13	25
	4 = Severe	6	13	5	8	0	4
	Total Abnormal ..	89	92	105	104	81	86
Porosis	0 = None	30	34	28	18	72	85
	1 = Doubtful	45	77	73	21	61	29
	2 = Slight	75	56	60	57	51	50
	3 = Moderate	32	21	28	81	6	23
	4 = Severe	10	4	3	15	2	5
	Total Abnormal ..	117	81	91	153	59	78

The differences of opinion over the assessment of erosion in individual films is illustrated in Tables III to V, in which are compared the readings of Observers V and VI (the two radiologists), Observers III and V (the pair with greatest differences), and Observers II and IV (two typical clinicians).

TABLE III
THERAPEUTIC TRIAL—EROSION

Observer V

Total	59	52	68	13	0	192
4	—	—	1	3	—	4
3	—	1	19	5	—	25
2	1	16	36	4	—	57
1	17	20	12	1	—	50
0	41	15	—	—	—	56
Score	0	1	2	3	4	Total

Observer VI

Although it will be seen that the distribution as shown in the Tables is not strictly normal, it is considered that the deviation is not so great as to preclude the use of correlation coefficients in analysis of the data. The results for the three pairs of observers illustrated are as follows:

Observers V and VI: $r = +0.77$ (Table III)

Observers III and V: $r = +0.86$ (Table IV)

Observers II and IV: $r = +0.83$ (Table V)

Observer I had had more experience of grading rheumatoid changes in hand films than the other observers, since he had been reading films from field surveys for over 2 years, and it is interesting to note that his total scores for erosion (Table I) fell very near the mean of the other five observers.

This observer's readings of individual films were compared with a composite score of the other five observers, obtained by adding together the individual scores of each film recorded by these observers and dividing the total by 5. The result is shown in Table VI. Here the correlation is: $r = +0.91$.

TABLE IV
THERAPEUTIC TRIAL—EROSION

Observer V

Total	59	52	68	13	0	192
4	—	—	—	5	—	5
3	—	1	39	7	—	47
2	2	22	28	1	—	53
1	22	25	1	—	—	48
0	35	4	—	—	—	39
Score	0	1	2	3	4	Total

Observer III

TABLE VI
THERAPEUTIC TRIAL—EROSION

Observer I

Total	62	41	51	32	6	192
4	—	—	—	—	—	0
3.8	—	—	—	—	3	3
3.6	—	—	—	—	2	2
3.4	—	—	—	—	—	0
3.2	—	—	—	2	—	2
3	—	—	—	8	1	9
2.8	—	—	1	7	—	8
2.6	—	—	10	4	—	14
2.4	—	—	5	7	—	12
2.2	—	3	7	3	—	13
2	—	—	10	1	—	11
1.8	—	4	7	—	—	11
1.6	—	3	5	—	—	8
1.4	1	3	1	—	—	5
1.2	4	5	1	—	—	10
1	2	7	2	—	—	11
0.8	4	7	1	—	—	12
0.6	8	4	1	—	—	13
0.4	12	3	—	—	—	15
0.2	18	2	—	—	—	20
0	13	—	—	—	—	13
Score	0	1	2	3	4	Total

Average of Observers II to VI

TABLE V
THERAPEUTIC TRIAL—EROSION

Observer IV

Total	49	39	51	45	8	192
4	—	—	—	6	7	13
3	—	—	8	27	1	36
2	2	4	28	9	—	43
1	15	19	13	3	—	50
0	32	16	2	—	—	50
Score	0	1	2	3	4	Total

Observer II

The individual readings of the other observers have been similarly compared with the composite scores of their colleagues, the results being as follows:

Observer II: $r = +0.88$;

Observer III: $r = +0.88$;

Observer IV: $r = +0.87$;

Observer V: $r = +0.89$;

Observer VI: $r = +0.83$.

There is thus no great difference between the performance of the various observers.

The same 192 films were re-read by Observer I using a single unified grading for rheumatoid changes instead of the separate gradings for porosis and erosion. The results are shown in Table VII.

TABLE VII
THERAPEUTIC TRIAL

Grade	Erosion	Porosis	Rheumatoid Arthritis
0 = None ..	62	30	55
1 = Doubtful ..	41	45	36
2 = Slight ..	51	75	65
3 = Moderate ..	32	32	29
4 = Severe ..	6	10	7
Total Abnormal ..	89	117	101

In comparing individual films the following results were obtained:

Rheumatoid arthritis and Erosion: $r = +0.82$;
 Rheumatoid arthritis and Porosis: $r = +0.72$;
 Rheumatoid arthritis and a combined factor of
 $\frac{\text{Erosion} + \text{Porosis}}{2}$: $r = +0.88$.

Thus, although porosis is difficult to estimate and may be unreliable, it has played a considerable part in the overall diagnosis of rheumatoid arthritis by Observer I in this sample of films.

Leigh Survey.—The 126 films of the hands in this series were all taken by the same technique and nearly all included a control bone. Unlike the first set of films, they were obtained from a population sample in the 55-64 age group which included very few individuals with early active rheumatoid arthritis. A substantial number, however, had old or inactive rheumatoid arthritis, many had advanced osteo-arthritic changes, and quite a number of the women also had osteoporosis due to conditions other than rheumatoid arthritis.

These 126 films were read by Observer I on three

occasions, the first and third readings being done jointly with an additional clinician (Observer VII). In the first and second readings only the combined grading for rheumatoid arthritis was recorded, but in the third reading porosis and erosions were graded separately and recorded irrespective of the apparent cause of these changes. The findings are shown in Table VIII, together with a composite score obtained by adding the individual gradings of the first two readings to the score for erosion and/or porosis, whichever was the highest, and dividing the total by 3.

Comparison of the gradings of the individual films in the first and second reading is shown in Table IX. The gradings were reasonably reproducible: $r = +0.80$. More films were recorded as normal or doubtful by the one observer reading alone, but in no instance did the single observer grade a film as abnormal that had been considered normal or doubtful in the first reading.

TABLE IX
LEIGH SURVEY—RHEUMATOID ARTHRITIS

First Reading							Second Reading
Total	48	45	27	2	4	126	
4	—	—	—	1	4	5	
3	—	—	—	1	—	1	
2	—	—	16	—	—	16	
1	4	20	5	—	—	29	
0	44	25	6	—	—	75	
Score	0	1	2	3	4	Total	

Comparison of the rheumatoid gradings with those for porosis showed wide disagreement: $r = +0.34$ for the first and $+0.36$ for the second rheumatoid arthritis reading. This disagreement is

TABLE VIII
LEIGH SURVEY

Grade	Rheumatoid Arthritis		Third Reading		Composite Score
	First Reading	Second Reading	Porosis	Erosion	
0 = None	48	75	84	74	$0 - \frac{1}{3}$ 53
1 = Doubtful	45	29	24	28	$\frac{2}{3} - 1\frac{1}{3}$ 48
2 = Slight	27	16	15	17	$1\frac{1}{3} - 2\frac{1}{3}$ 19
3 = Moderate	2	1	1	3	$2\frac{1}{3} - 3\frac{1}{3}$ 1
4 = Severe	4	5	2	4	$3\frac{1}{3} - 4$ 5
Total Abnormal	33	22	18	24	$\geq 1\frac{1}{3}$ 25

largely due to the almost complete absence of porosis in the 41 films from males in this series. In females the total distribution of the various grades of porosis and of rheumatoid arthritis hardly differ, but, even so, the correlation in females between porosis and the first and second rheumatoid arthritis readings is not more than $+0.54$ and $+0.58$. Comparison of the rheumatoid gradings with those of erosion, on the other hand, showed good agreement: r for the two sexes combined = $+0.81$ and $+0.84$, with no conspicuous difference between males and females.

Osteoporosis as seen in hand films is not always qualitatively the same. Thus, in old age and the post-menopausal state the porosis develops slowly over many years and the radiological appearance is one of thin and sparse trabeculation in the bone, the trabecular pattern being, if anything, more clearly defined than usual. In active early rheumatoid arthritis and in post-traumatic porosis, the bone is more rapidly removed over a period of weeks and months and the radiological appearances differ. Thus, the trabecular pattern as seen in hand films is "smudged" and "blurred" with spotty areas of translucency.

This appearance can to some extent be recognized irrespective of estimates of total bone density. Amongst the films obtained from the sample population aged 55-64 years were six showing porosis of the more chronic type and one showing acute porosis associated with a Colles fracture. These were not considered as evidence of rheumatoid arthritis and this accounts for the poor correlation between osteoporosis and radiological rheumatoid arthritis in this group.

The remaining 254 films from the Leigh Survey which were not considered to show any definite signs of rheumatoid arthritis on the first reading were read a second time, porosis and erosions being recorded separately irrespective of whether these changes were thought to be due to rheumatoid arthritis or not.

Minimal porosis was recorded in films from two men and fourteen women, and moderate porosis in films from three women.

Minimal erosions were recorded in films from three men and three women.

In only one film did porosis and erosions coincide, and this was the only film in which the changes might on re-reading have been considered to be due to rheumatoid arthritis. The porosis in the others was all of the chronic variety. Finally, six minimal erosions were associated with definite evidence of trauma or deformities to which they were thought to be related.

South Wales Survey.—The results of the two separate readings of these 259 hand films by Observer I are shown in Table X. Between the first and second reading there was an interval of over one year, and during that time most of the other studies reported here had been carried out. As a result, this observer's criteria for the radiological grading of rheumatoid arthritis had become stricter. This can be seen by the uniform shift in the diagram and by the fact that 27 films that were graded as positive on the first reading were graded as negative or doubtful on the second; whereas no films were graded as positive on the second reading which were considered negative or doubtful on the first. The correlation between first and second readings is in general good: $r = +0.88$.

TABLE X
SOUTH WALES SURVEY—RHEUMATOID ARTHRITIS

First Reading							Second Reading
Total	128	40	44	22	25	259	
4	—	—	—	—	17	17	
3	—	—	—	16	7	23	
2	—	—	17	5	1	23	
1	11	12	16	1	—	40	
0	117	28	11	—	—	156	
Score	0	1	2	3	4	Total	

Conclusions

- (1) Osteoporosis may be difficult to assess and there can be wide disagreement between different observers in recording even its presence or absence.
- (2) Erosion can be recognized with reasonable certainty, and fair agreement is obtained in recording the presence or absence of this change, but there is wide disagreement over degree of severity. This is presumably a language difficulty and might be partly overcome by the use of standard films.
- (3) In films from field surveys there is little correlation between erosions and porosis, because of the prevalence of non-rheumatoid porosis in older age groups. Erosions should therefore be mainly relied upon for grading survey films for rheumatoid arthritis.
- (4) The x-ray recordings of Observer I are about average and are not likely to differ greatly from

the mean readings obtained from a number of clinicians and radiologists.

- (5) In grading films for rheumatoid arthritis there is a lack of reproducibility, especially at the important division between doubtful and slight so that in survey work a composite score derived from several separate readings should be used.
- (6) As criteria for diagnosis may differ between observers and may change with time, particularly at the division between doubtful and slight, some agreed set of standard reference films should be employed if *x*-ray changes are to be used for assessing comparative prevalence rates of rheumatoid arthritis by different observers, or by the same observer at different times.
- (7) If *x*-rays of the hands are used to assess progress in therapeutic trials, it is essential that all films be read by the same observer, preferably at a single session.

Summary

Three sets of *x* rays of the hands have been used to estimate the degree of inter-observer and intra-observer differences in interpreting changes of rheumatoid arthritis.

Wide disagreement between observers was found in the assessment and even in the recognition of osteoporosis. Erosion was recognized with reasonable certainty, but its severity was variously interpreted.

The reassessment, after a year, of a set of 259 films by the same observer showed a small difference between the two readings; this difference was due to stricter criteria for the radiological grading of rheumatoid arthritis at the second reading.

Suggestions are made to ensure a greater uniformity in grading *x* rays in field surveys and therapeutic trials, such as an agreed set of standard reference films and the reading of all films by the same observer, preferably at a single session.

The films of the therapeutic trial were read by Drs. E. G. L. Bywaters, J. J. R. Duthie, D. E. Fletcher, K. A. Rowley, and H. F. West. Without their co-operation and helpful criticism this study would not have been possible and I am deeply indebted to them for their help. I should also like to thank Dr. W. E. Miall of the Pneumoconiosis Research Unit for providing the films of the South Wales Survey.

REFERENCES

- Ball, J. (1950). *Lancet*, 2, 520.
 Cochrane, A. L., Campbell, H. W., and Stein, S. C. (1949). *Amer. J. Roentgenol.*, 61, 153.
 —, and Garland, L. H. (1952). *Lancet*, 2, 505.
 Empire Rheumatism Council (1955). *Annals of the Rheumatic Diseases*, 14, 353.

- Fletcher, C. M., and Oldham, P. D. (1949). *Brit. J. industr. Med.*, 6, 168.
 Fletcher, D. E., and Rowley, K. A. (1952). *Brit. J. Radiol.*, 25, 282.
 Joint Committee of the Medical Research Council and Nuffield Foundation on Clinical Trials of Cortisone, A.C.T.H., and Other Therapeutic Measures in Chronic Rheumatic Diseases (1954). *Brit. med. J.*, 1, 1223.
 — (1955). *Ibid.*, 2, 695.
 Miall, W. E. (1955). *Annals of the Rheumatic Diseases*, 14, 150.
 —, Caplan, A., Cochrane, A. L., Kilpatrick, G. S., and Oldham, P. D. (1953). *Brit. med. J.*, 2, 1231.
 Rose, H. M., Ragan, C., Pearce, E., and Lipman, M. O. (1948). *Proc. Soc. exp. Biol. (N.Y.)*, 68, 1.
 Steven, G. D. (1947). *Annals of the Rheumatic Diseases*, 6, 1, 184.
 Yerushalmy, J., Garland, L. H., Harkness, J. T., Hinshaw, H. C., Miller, E. R., Shipman, S. J., and Zwerling, H. B. (1951). *Amer. Rev. Tuberc.*, 64, 225.
 —, Harkness, J. T., Cope, J. H., and Kennedy, B. R. (1950). *Ibid.*, 61, 443.

Signes radiologiques de l'arthrite rhumatismale. Une étude des différences d'interprétation des radiologies des mains selon l'observateur

RÉSUMÉ

A l'aide de trois séries des radiographies des mains on détermina l'amplitude des divergences entre plusieurs observateurs et chez le même observateur dans l'interprétation des lésions rhumatismales.

On trouva un désaccord considérable entre les observateurs en ce qui concerne l'évaluation et même la présence de l'ostéoporose. Le diagnostic de l'érosion fut plus facile, mais sa sévérité fut variablement interprétée.

Une seconde évaluation par le même observateur un an plus tard d'une série de 259 clichés révéla quelques différences dues au fait que ses critères pour l'appréciation radiologique de l'arthrite rhumatismale devinrent plus rigoureux.

Pour assurer une appréciation plus uniforme des radiologies au cours des enquêtes "de champ" et des essais thérapeutiques, on suggère quelques remèdes, tels qu'une série convenue de clichés-étalon et l'interprétation de tous les clichés par le même observateur, préférablement au cours d'une seule session.

Manifestaciones radiológicas de la artritis reumatoide. Un estudio de las diferencias entre observadores en la interpretación de la radiografías de manos

SUMARIO

Con la ayuda de tres series de radiografías de manos se determinó la amplitud de las diferencias entre varios observadores y en el mismo observador respecto a la interpretación de las lesiones reumáticas.

Un desacuerdo considerable entre varios observadores fué revelado en la valoración y hasta en la afirmación de la presencia de la osteoporosis. El diagnóstico de la erosión fué más fácil pero hubo mayores variaciones en la interpretación de su severidad.

La valoración repetida por el mismo observador un año después de 259 clisés resultó en pequeñas diferencias, debidas al hecho de que los criterios en la apreciación radiológica de la artritis reumatoide fueron más rigurosos la segunda vez.

Para asegurar la valoración más uniforme de radiologías durante investigaciones "de campo" y ensayos terapéuticos, se sugieren algunos remedios, tales como una serie convenida de clisés de norma y la interpretación de todos los clisés por el mismo observador, preferiblemente en el curso de una sesión.