

Inter-observer Variation in Coding Osteoarthritis in Human Skeletal Remains

TONY WALDRON¹ AND JULIET ROGERS²

Palaeopathology Study Group: ¹Institute of Archaeology, University College, London, UK: and ²Department of Rheumatology, Bristol Royal Infirmary, Bristol, UK

ABSTRACT

Thirty-eight participants at the VIIIth European Meeting of the Paleopathology Association took part in a study of inter-observer variation in scoring osteoarthritis in human skeletal remains. Ten specimens representing different joints were used and five criteria of osteoarthritis were scored. Eleven of the 38 participants ranked themselves as beginners, 13 as experienced and six as very experienced; the data were subsequently examined using the results from these 30, comparing beginners with experts. Agreement as to whether or not changes were present on the specimens and on the degree of change was seldom complete but was greater when scoring eburnation and the presence of new bone on the joint surface than for the three other criteria. There was little difference between beginners and experts.

Although all the specimens were chosen to meet our published criteria for osteoarthritis, the experts were unanimous in agreeing the diagnosis in only three cases and the beginners in only one.

These results suggest that more work needs to be done to develop operational definitions for the classification of disease in palaeopathology and that great care must be taken when comparing disease frequencies between studies.

Keywords: Palaeopathology, Osteoarthritis, Epidemiology, Inter-observer variation.

Introduction

In epidemiological studies of joint diseases, where results from different observers are to be compared, it is common to estimate the degree of variation between the observers;^{1,2} if the variation is small, then the conclusions drawn from such studies will have more credibility than if it is great. Now that epidemiological techniques are increasingly being applied to palaeopathology it seemed important to make some assessment of the extent of inter-observer variation amongst its practitioners when scoring diseased joints, since to the best of our knowledge this has not been attempted before.

Materials and methods

Each of the specimens used in the study (and illustrated in Figure 1) was considered to have

the features of osteoarthritis by the criteria that we have published before.³ Briefly, these criteria state that osteoarthritis is present when there is eburnation on the joint surface or, in the absence of eburnation, when two of the following are present: marginal osteophytes, new bone on the joint surface, pitting on the joint surface or deformation of the joint contour.

Ten specimens were chosen to illustrate some or all of these changes at as many different joints as possible. The specimens (and joints) were as follows: (1) left tibia (tibio-femoral joint); (2) right distal humerus (elbow joint); (3) right trapezium (first carpo-metacarpal joint); (4) right distal clavicle (acromio-clavicular joint); (5) left tibia (tibio-femoral joint); (6) right humeral head (gleno-humeral joint); (7) lumbar vertebra (inferior facet joint); (8) left scapula (gleno-humeral joint); (9) right femoral head (hip joint); and (10) right first metatarsal (metatarso-phalangeal joint).

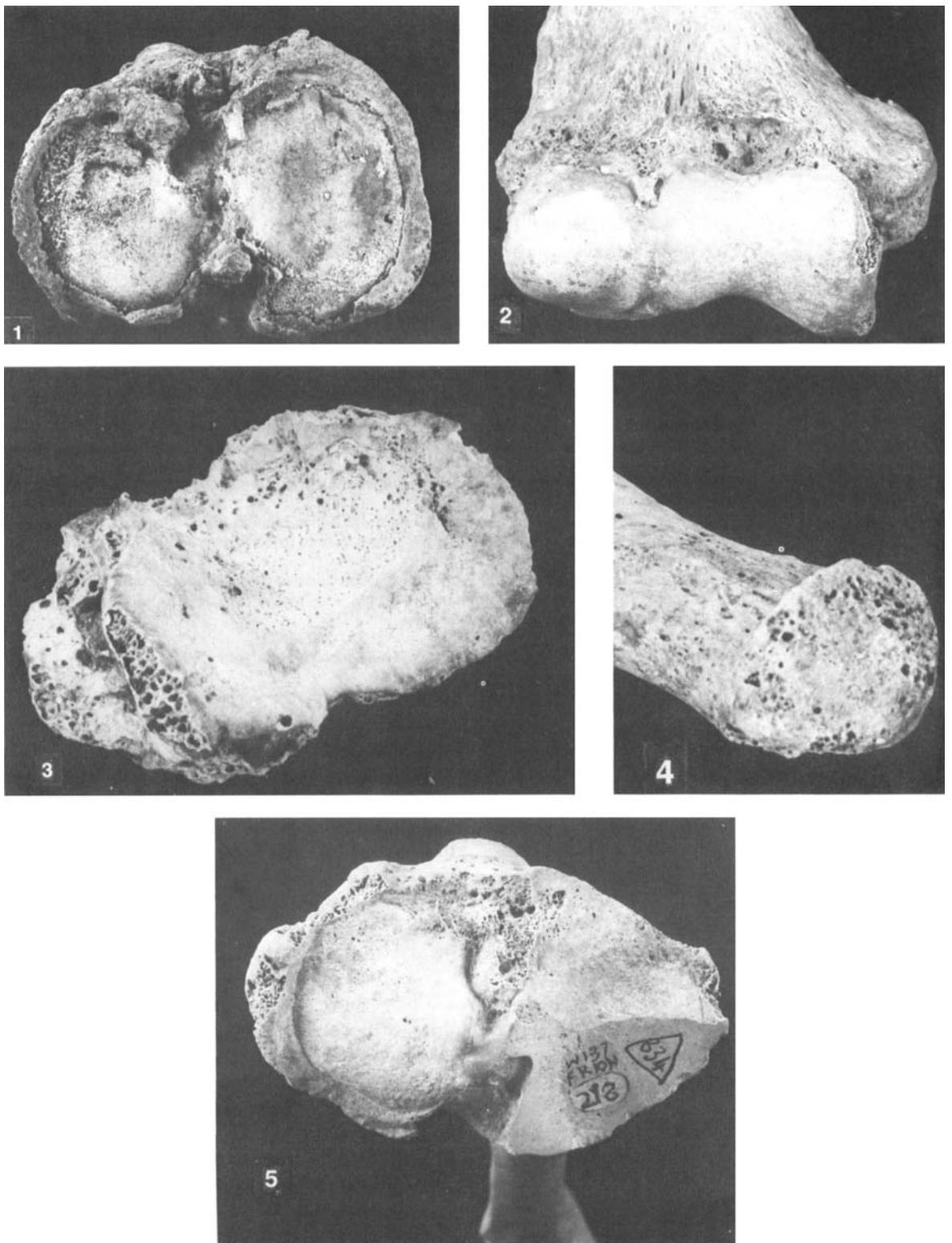
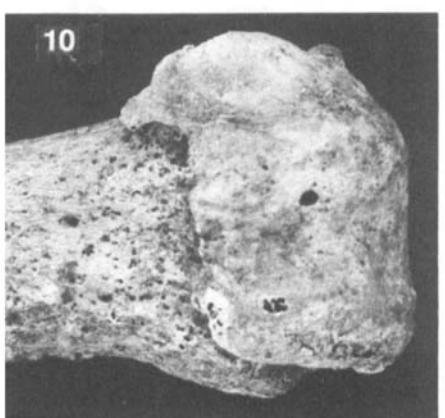
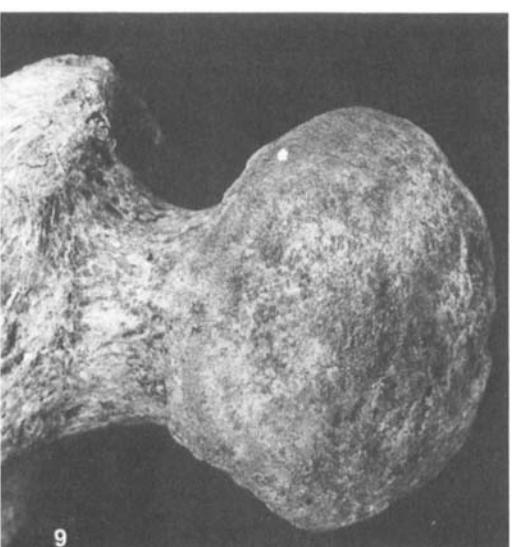
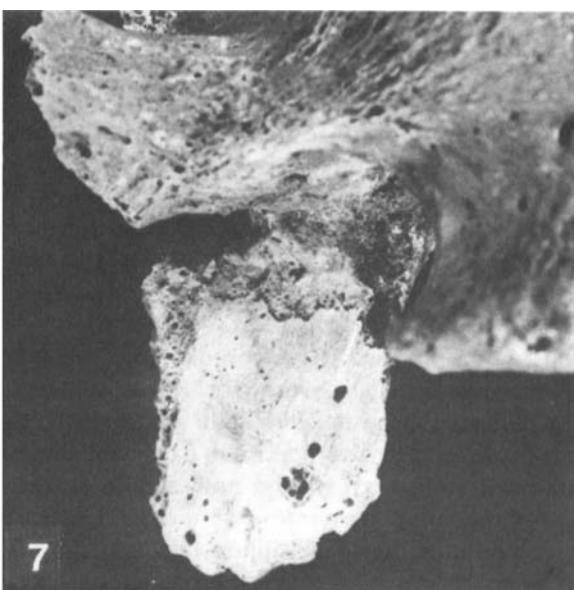


Figure 1. Specimens used for inter-observer variation. 1, left tibia; 2, right distal humerus; 3, right trapezium; 4, right distal clavicle; 5, left tibia; 6, right humeral head; 7, lumbar vertebra; 8, left scapula; 9, right femoral head; 10, right first metatarsal.



The participants in the study were all delegates at the VIIIth European Meeting of the Paleopathology Association, which was held in Cambridge from 19 to 22 September 1990. At the start of the conference, the purpose of the study was explained and delegates were invited to take part. They were asked to grade each of the five diagnostic criteria noted above for each specimen as being present or absent; if they considered that changes were present in the joint, then each was to be scored on a three point scale from 1 (least severe) to 3 (most severe). Finally, they were asked to state whether or not they considered the specimen to have osteoarthritis. Although we did not ask the participants to give their names, we did ask them to give their principal discipline (anthropologist, palaeopathologist or other) and to state their level of expertise (beginner, experienced or very experienced).

Results

Of the 83 delegates (excluding the authors of this paper), 38 (45.8%) took part in the trial. Of

these, nine gave their principal discipline as anthropologist, and 13 as palaeopathologist; seven were in different disciplines and nine gave no response.

Eleven of the 38 participants ranked themselves as beginners, 13 as experienced and six as very experienced; eight were too unsure of themselves to respond. Since one of the aims of the exercise was to measure differences between beginners and those who considered themselves experienced, we analysed the data only from the 11 beginners and the 19 experts, grouping together those who were 'experienced' and 'very experienced'. We did not analyse the data according to discipline since the numbers within each group were too small.

For each of the 10 specimens there were five criteria to be scored, giving an overall matrix of 50 scores. For each of the 50 cells we calculated what we termed a concordance score, which was simply taken as the largest number agreeing a particular grade expressed as the total number scoring (which in a very few cases was less than the absolute number in the group). The mean concordance scores were calculated for each diagnostic criterion and also for the five criteria

Table 1. Concordance scores in inter-observer error study.

| Specimen | Beginner (B) or expert (E) | Eburnation | Marginal osteophyte | Changes in joint contour | Pitting on joint surface | New bone on joint surface | Mean (standard error (SE)) |
|----------|----------------------------|-------------|---------------------|--------------------------|--------------------------|---------------------------|----------------------------|
| 1 | B | 63.6 | 63.6 | 45.5 | 72.7 | 54.5 | 59.9 (4.6) |
| | E | 73.7 | 57.9 | 36.8 | 47.4 | 42.1 | 51.6 (6.5) |
| 2 | B | 36.4 | 36.4 | 36.4 | 63.6 | 54.5 | 45.5 (5.7) |
| | E | 52.6 | 42.1 | 52.6 | 63.2 | 57.9 | 53.7 (3.5) |
| 3 | B | 54.5 | 36.4 | 54.5 | 45.5 | 63.6 | 50.9 (4.6) |
| | E | 63.2 | 52.6 | 42.1 | 52.6 | 52.6 | 52.6 (3.3) |
| 4 | B | 100 | 45.5 | 63.6 | 45.5 | 72.7 | 65.5 (10.1) |
| | E | 94.7 | 52.6 | 36.8 | 68.4 | 68.4 | 64.2 (9.6) |
| 5 | B | 54.5 | 45.5 | 54.5 | 54.5 | 45.5 | 50.9 (2.2) |
| | E | 68.4 | 52.6 | 57.9 | 63.2 | 52.6 | 58.9 (3.1)* |
| 6 | B | 72.7 | 45.5 | 72.7 | 90.9 | 72.7 | 70.9 (8.4) |
| | E | 100 | 63.2 | 47.4 | 89.5 | 68.4 | 73.7 (8.4) |
| 7 | B | 45.5 | 45.5 | 36.4 | 63.6 | 72.7 | 52.7 (6.1) |
| | E | 68.4 | 52.6 | 36.8 | 42.1 | 47.4 | 49.5 (6.1) |
| 8 | B | 54.5 | 63.6 | 45.5 | 72.7 | 45.5 | 56.4 (4.3) |
| | E | 63.2 | 57.9 | 47.4 | 63.2 | 52.6 | 56.9 (4.3) |
| 9 | B | 81.8 | 45.5 | 54.5 | 54.5 | 45.5 | 56.4 (7.9) |
| | E | 100 | 68.4 | 59.9 | 47.4 | 57.9 | 66.7 (7.9) |
| 10 | B | 81.8 | 45.5 | 63.6 | 72.7 | 54.5 | 63.6 (6.4) |
| | E | 68.4 | 52.6 | 52.6 | 68.4 | 52.6 | 58.9 (3.9) |
| Mean SE | B | 64.5 (6.1) | 55.3 (2.8) | 52.7 (3.8) | 63.6 (4.5) | 58.2 (3.6) | |
| | E | 75.3 (5.3)* | 55.3 (2.3) | 47.0 (2.3) | 60.5 (4.4) | 55.3 (2.6) | |

* $p < 0.05$

for each specimen and the results are shown in Table 1. It should be noted that the beginners and experts did not always agree on which grade should be allocated to each criterion, but this was not taken into account when analysing the results; our interest was only in the degree to which the two groups agreed on their grading, not in whether it was the same grading in both cases.

Agreement was clearly easier to achieve for some changes than for others and concordance scores tended to be higher when scoring eburnation and new bone on the joint surface than for the other changes. It was interesting to note that the experts had *lower* mean concordance scores than the beginners for three of the five criteria and a higher mean score only for eburnation, but this was the only significant difference ($t = 2.75, p < 0.05$). Higher mean scores were achieved on six of the specimens by the experts but in only one case (specimen 5) was the difference significant ($t = 4.72, p < 0.01$).

As a second stage in the analysis, we decided to compare the numbers in each group registering change or no change for each of the criteria. In this instance we were not interested in how many agreed on the grade but merely whether change was present or absent. The number scoring change as *present* in each cell was subtracted from the total number responding and then expressed as a percentage. If all the group scored change as being present, then the percentage was zero; if they all scored it as being absent, the percentage was 100. These data are shown for each criterion and for each specimen in Figure 2. It should be noted when looking at these figures that the measure of agreement increases at both extremes and is least at 50% since then the group is equally split as to whether change is present or absent.

In general, agreement is reasonable although there are some notable exceptions. For example,

there was a good deal of disagreement about whether or not specimens 4 and 7 had marginal osteophyte; whether specimens 2 and 5 had pitting on the joint surface and whether specimens 3, 5, 7 and 8 had new bone on the joint surface. In only very few cases was there unanimity that change was present or absent, and the experts were no more likely to agree amongst themselves than the beginners (see Table 2).

Finally, we looked to see the proportions who considered that the joints had osteoarthritis. The experts were unanimous for three of the specimens, 1, 3 and 6, whereas the beginners were unanimous only about specimen 6 (see Figure 3). There was only one dissenting voice amongst the experts for specimens 5, 7 and 8 and amongst the beginners for 1, 2, 5 and 8. The greatest degree of dissent amongst the experts was shown over specimen 9 and for the beginners over specimen 7. In only a single case (specimen 7), however, was the difference between experts and beginners statistically significant ($p < 0.05$).

Discussion

We think that is the first study of inter-observer variation that has been carried out in palaeopathology and the results are not entirely comforting. We chose to carry out the study using bones that showed some or all the features of osteoarthritis on the grounds that osteoarthritis is the most common disease found in human skeletal remains and so most experienced palaeopathologists would have had plenty of practice at observing it, and also because we thought that it was comparatively easy to tell whether or not the characteristic changes were present, whatever the cause of the changes may have been.

Table 2. Number of cases with complete agreement on presence or absence of criteria of osteoarthritis.

| Beginners (B) or experts (E) | Eburnation | Marginal osteophyte | Changes in joint contour | Pitting on joint surface | New bone on joint surface |
|------------------------------|------------|---------------------|--------------------------|--------------------------|---------------------------|
| B | 5 | 4 | 4 | 2 | 0 |
| E | 5 | 2 | 4 | 2 | 1 |

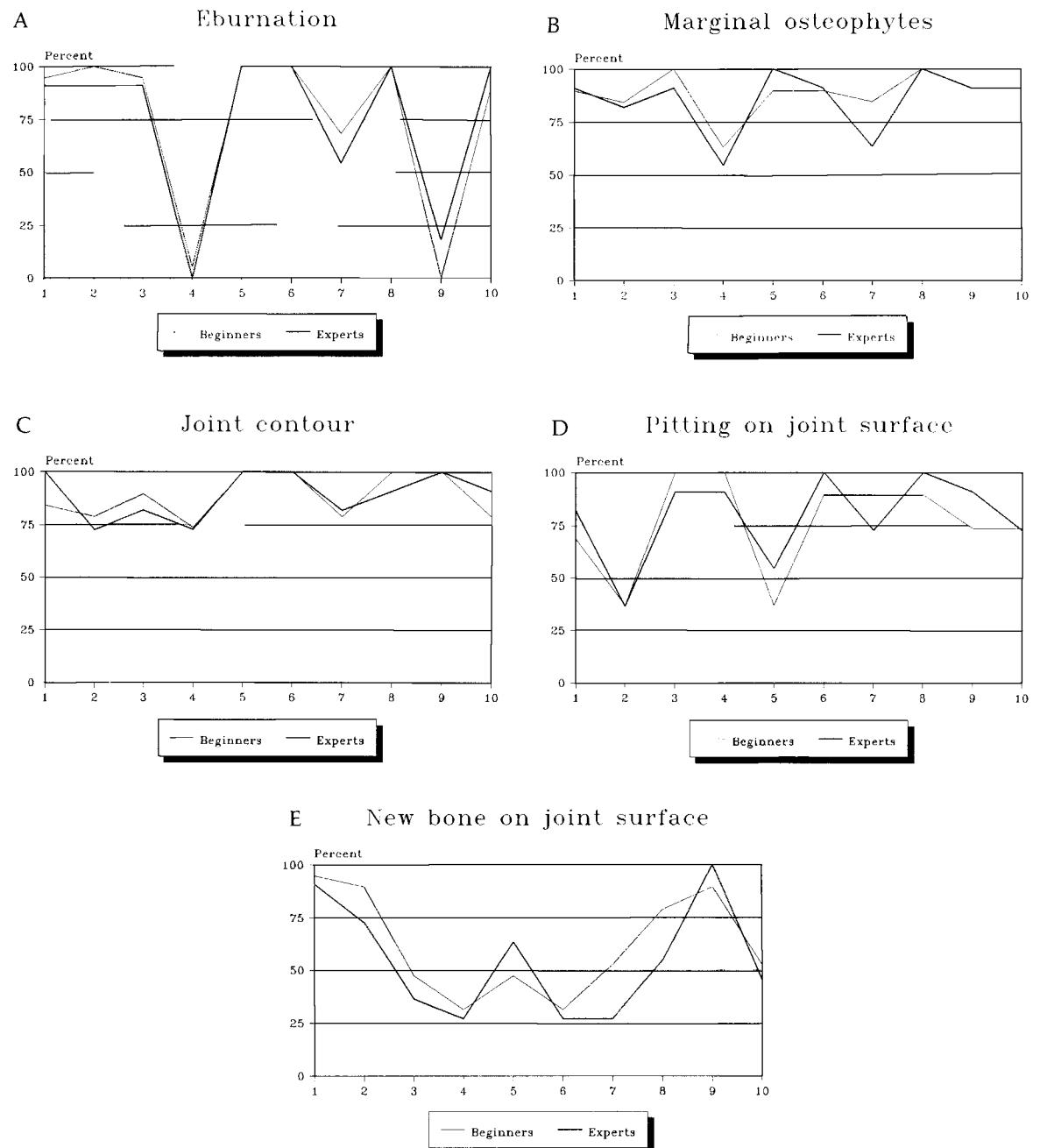


Figure 2. Proportion of beginners and experts agreeing on (a) eburnation, (b) marginal osteophytes, (c) alteration in joint contour, (d) pitting on joint surface and (e) new bone on joint surface being present or absent. For explanation, see text.

What has emerged from the study is that there is seldom complete agreement as to whether pathological changes are present and that agreement on the severity of the changes that are seen is seldom likely to be achieved by

more than half the observers. What is even more disturbing, however, is that experts seldom achieve more consistent results than beginners. Of course we have to enter the caveat here that the self-rating may not have been entirely

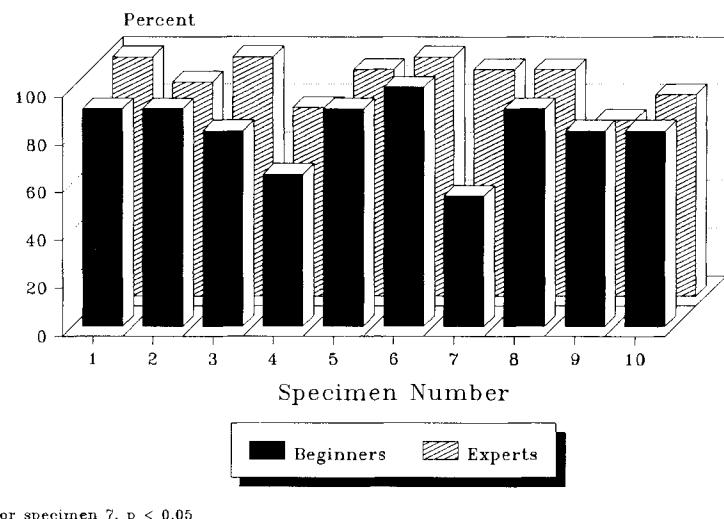


Figure 3. Proportion of beginners and experts agreeing on the diagnosis of osteoarthritis for each specimen.

accurate and that some bias is very likely in the direction of experienced workers under-rating themselves, but we have no means of checking this since the study was carried out anonymously.

There was also considerable disagreement when it came to attributing the diagnosis of osteoarthritis to the specimens. We chose all ten as meeting our published criteria for osteoarthritis in palaeopathological material and so, to some extent, this was a test of the acceptance of our criteria. As it was, the experts were unanimous on three specimens only (1, 3 and 6) and the beginners were unanimous only about specimen 6. There was a significant difference in the number of experts and beginners only over the diagnosis of specimen 7; 94.7% of the experts considered it to have osteoarthritis but only 54.5% of the beginners. Considering this result, and some of the other results pertaining to specimen 7, it seems likely that some of the participants were scoring the wrong joint. The inferior facet joint (shown in Figure 1) had very obvious eburnation and scoring on the joint surface and could not possibly have been mistaken for anything other than osteoarthritis; the fact that so many people did misdiagnose it suggests that they were actually scoring the normal, superior facet joint; this, at least, is our most charitable explanation.

In general, there was better agreement about the diagnosis when eburnation was present,

although three of the experts and two of the beginners did not diagnose specimen 10 as having osteoarthritis even though there was an obvious area of eburnation on the inferior surface of the head of this metatarsal.

One practical outcome of this study must be that greater efforts will need to be taken to agree on the criteria for the classification of pathological changes in palaeopathology. What is required for palaeopathological work is the development of operational definitions, since it is not valid entirely to rely on either clinical or radiological diagnoses. For example, we have shown that radiology is not reliable in detecting changes that are evident to the palaeopathologist.⁴

Finally, these results demonstrate that great care must be taken when comparing disease frequencies between studies unless the authors cite the criteria used in arriving at their classifications or use a common, agreed source of reference.

References

1. Cooper, C., Cushnaghan, J., Kirwan, J., Rogers, J., McAlindon, T., McCrae, F. and Dieppe, P. A. Radiographic assessment of the knee joint in osteoarthritis. *British Journal of Rheumatology*, 1990; **29**: Supplement 1: 19.
2. Cushnaghan, J., Cooper, C., Dieppe, P., Kirwan, J., McAlindon, T. and McCrae, F. Clinical assess

- ment of osteoarthritis of the knee. *Annals of the Rheumatic Diseases*, 1990; **49**: 768-770.
3. Rogers, J., Waldron, T., Dieppe, P. and Watt, I. Arthropathies in palaeopathology: the basis of classification according to most probable cause. *Journal of Archaeological Science*, 1987; **14**: 179-193.
4. Rogers, J., Watt, I. and Dieppe, P. Comparison of visual and radiographic detection of bony changes at the knee joint. *British Medical Journal*, 1990; **300**: 367-368.