Ann. rheum. Dis. (1973), **32**, 354

New York clinical criteria for ankylosing spondylitis

A statistical evaluation

J. M. H. MOLL* AND V. WRIGHT†
From the Rheumatism Research Unit, University Department of Medicine, the General Infirmary at Leeds, and Royal Bath Hospital, Harrogate

Criteria for use in population surveys to study the epidemiology of ankylosing spondylitis were proposed at the CIOMS Symposium in Rome in 1961 (Kellgren, Jeffrey, and Ball, 1963), and a revised set of criteria were introduced at the next symposium in New York in 1966 (Bennett and Wood, 1968). The Rome and New York criteria are shown in Tables I and II respectively. The Rome criteria specify that the diagnosis of ankylosing spondylitis should be made when bilateral sacroiliitis and one of five clinical criteria are present, or when four clinical criteria are present. Using these criteria, therefore, it is possible to make a diagnosis of spondylitis without radiological evidence of disease. However, bilateral sacroiliitis is recognized as the most important criterion and has a weight three times greater than any of the clinical criteria. The Rome and New York criteria have been evaluated previously (Bennett and Burch, 1968; Chalmers, Danchot, Kellgren, King, Pikhlak, Sievers, Strevens, Tait, and Wood, 1969), but evaluation was assessed from largely subjective data—apart from the objective measurement of chest expansion. A further drawback relating to this earlier work was that no correction was made for the effects of age and sex on chest and spinal mobility. Recently, we have developed simple *objective* clinical methods to measure spinal mobility in each of the stipulated (New York, 1966) directions of movement—anterior flexion (Macrae and Wright, 1969), lateral flexion (Moll, Lyanage, and Wright, 1972a), and extension (Moll, Lyanage, and Wright, 1972b). Furthermore, we have particularly emphasized the importance of allowing for the effects of age and sex not only on spinal mobility (Moll and Wright, 1971) but also on chest expansion (Moll and Wright, 1972).

Employing these new objective techniques, the New York diagnostic criteria have been applied to a

Table I *Criteria for diagnosing ankylosing spondylitis (Rome, 1961)*

Clinical criteria
1. Low back pain and stiffness for more than 3 months which is not relieved by rest
2. Pain and stiffness in the thoracic region
3. LIMITED MOTION IN THE LUMBAR SPINE
4. LIMITED CHEST EXPANSION
5. History or evidence of iritis or its sequelae
Radiological criterion
6. *X* ray showing bilateral sacroiliac changes characteristic of ankylosing spondylitis (this would exclude bilateral osteoarthrosis of the sacroiliac joints)

Table II *Clinical criteria for ankylosing spondylitis (New York, 1966)*

(A) Diagnosis
1. LIMITATION OF MOTION OF THE LUMBAR SPINE in all THREE PLANES—anterior flexion, lateral flexion, and extension
2. History or the presence of PAIN at the dorso-lumbar junction or in the lumbar spine
3. LIMITATION OF CHEST EXPANSION to 1 in. (2·5 cm.) or less, measured at the level of the fourth intercostal space
(B) Grading
DEFINITE AS:
1. *Grade 3–4 bilateral* sacroiliitis with *at least one clinical criterion.*
2. *Grade 3–4 unilateral* or *Grade 2 bilateral* sacroiliitis with *Clinical criterion 1* (limitation of back movement in all three planes) or with both *Clinical criteria 2 and 3* (back pain and limitation of chest expansion)
PROBABLE AS:
Grade 3–4 bilateral sacroiliitis *with no clinical criteria*

Accepted for publication November 9, 1972.
* Senior Registrar in Rheumatology and Physical Medicine, Oxford Regional Hospital Board (formerly Research Assistant, Department of Rheumatology, University of Leeds).
† Professor of Rheumatology, University Department of Medicine, Leeds.

large group of normal and spondylitic subjects examined during a family study of psoriatic arthritis (Moll, 1971), and the data have subsequently been used to evaluate these criteria. In other words, objective measurement has enabled accurate evaluation of the clinical parameters laid down for the diagnosis of ankylosing spondylitis, and has thus helped to resolve the current controversy as to which criteria are valuable and which are not.

In addition, each patient's chest and spinal mobility has been assessed by conventional objective techniques and the results compared with data obtained by purely objective procedures. In this way, the importance of exact measurements in making a clinical diagnosis of spondylitis has been estimated.

The evaluation reported in this paper therefore concerns: (1) a statistical assessment of the relative value of the various New York clinical criteria for ankylosing spondylitis, and (2) a study of the relative importance of employing objective and subjective methods in making such a diagnosis.

Clinical material

In view of the known association between ankylosing spondylitis and psoriatic arthritis, it was anticipated that a relatively large number of spondylitis cases would be encountered in a family study of psoriatic arthritis. It was considered, therefore, that the primary genetic study of the disease (Moll, 1971) might provide a reasonable opportunity, as a secondary investigation, to test the validity of the New York criteria for ankylosing spondylitis.

A total of 412 subjects was examined during the survey. This population comprised 104 probands, 201 first-degree relatives, 31 second-degree relatives, and 76 spouse controls. The sex distribution and prevalence of sacroiliitis in the study population are shown in Table III. The age distribution of both male and female subjects followed a Gaussian (normal) distribution. The completion rate, calculated from the overall response to participate in the survey, was more than 80 per cent.

Methods

All subjects were examined *subjectively* without making any correction for the effect of age or sex, and *objectively* using the new methods evolved in the department. All objective measurements were corrected for age and sex. To avoid bias, subjective assessment was always made before objective assessment. In this way, subjective and objective examination of each individual enabled diagnosis to be made according to two sets of criteria:

(1) *The original New York criteria* (a largely subjective diagnosis);

(2) *A set of modified New York criteria* (a largely objective diagnosis).

In the case of the original criteria, back mobility was assessed in all three directions, and the results recorded as simply 'normal' or 'abnormal' on a purely subjective basis. The instructions given to the patient before each movement were identical to those given during the objective measurement. Chest expansion was estimated conventionally with a tape-measure and was regarded as abnormal, regardless of age and sex, if the value was 2·5 cm. or less.

In the case of the modified criteria, back mobility was ascribed a numerical value in each plane of movement and this was regarded as abnormal if the value was less than 2 S.D. from the mean for that particular sex and decade. Likewise, correction for age and sex was applied to measurements of chest expansion.

The methods used in the present study may be divided into:

(1) Objective clinical method to measure chest expansion;

(2) Objective clinical method to measure spinal mobility in three directions (anterior flexion, lateral flexion, and extension);

(3) Radiological technique for the study of pelvic *x* rays for evidence of sacroiliitis.

(1) Measurement of chest expansion

The details of this method and charts enabling a correction to be made for the effect of age and sex on chest expansion have been previously reported (Moll and Wright, 1972).

Table III *Details of study population showing sex distribution and sex prevalence of sacroiliitis (Grade 2–4)*

Relationship of subject	*Total examined*	*Sex*		*Prevalence of sacroiliitis*†			
		Male	*Female*	*Male*	*Female*	*Total*	
						No.	*Per cent.*
Probands*	104	33	71	10	13	23	22·1
First-degree relatives	201	122	79	5	8	13	6·4
Second-degree relatives	31	10	21	1	0	1	3·2
Spouse controls	76	52	24	1	0	1	1·3
Total	412	217	195	17	21	38	9·2

* 88 had true psoriatic arthritis, 16 had psoriasis and other arthritis.
† Three relatives (1 male; 2 females) had no clinical features of ankylosing spondylitis; there were no probands or spouse controls in this category.

All observations were made on subjects unclad to the waist and standing with hands on head and with arms flexed in the frontal plane. Circumferential chest mobility was assessed by means of a conventional centimetre tape-measure at the level of the 4th intercostal space as recommended at the New York Symposium (Bennett and Wood, 1968). Measurements were obtained at the height of maximal inspiration and expiration, and considerable care was taken not to pull the tape too tightly while making the measurements. The observations were always preceded by detailed instructions and, more importantly, a personal demonstration of what was required.

(2) Measurement of spinal mobility

ANTERIOR SPINAL FLEXION

This method is a modification of a technique originally described by Schober (1937) and recently reported by us elsewhere (Macrae and Wright, 1969).

Three marks were inked on the skin overlying the lumbo-sacral spine with the subject standing erect (Figure A). The first mark was placed at the lumbo-sacral junction which is represented by the spinal intersection of a line joining the dimples of Venus. Further marks were inked 5 cm. below and 10 cm. above the lumbo-sacral junction. The subject was then asked to bend forward and touch his toes, and the new distance between the upper and lower marks was measured. The distraction between these two marks has been found in a separate study to correlate very closely ($r = +0{\cdot}97$; $P < 0{\cdot}001$) with anterior flexion measured radiologically.

LATERAL SPINAL FLEXION

This new method (Moll and Wright, 1972a) involved inking two marks on the skin of the lateral trunk with the subject standing erect (Figure B). The upper mark represents the point where a horizontal line through the highest point on the iliac crest crosses the coronal line. The subject was then instructed to bend sideways as far as possible by sliding the hand down the homolateral thigh. On completing this movement, the new distance between these marks was measured. A separate study revealed that the distance between the first and second measurements correlated reasonably closely ($r = +0{\cdot}79$; $P < 0{\cdot}001$) with lateral thoraco-lumbar flexion measured radiologically.

SPINAL EXTENSION

The landmarks used in this technique (Moll and others, 1972b) were identical to those employed in the method to measure lateral flexion. A simple plumb-line was constructed; this consisted of a pointed weight suspended by a thread approximately 20 cm. long. The point of the plumb-line was positioned to coincide with the lower mark and the thread held at the upper mark (Figure C). To facilitate the manipulation, the subject was asked to stand erect with hands on head. Without support, he was instructed to bend over backwards as far as possible without flexing the knees and with groins approximated to the edge of the examination couch. This was specified in order to minimize pelvic tilting and secondary hip flexion. At the point of maximal extension the distance traversed by the plumb-line pointer was marked on the skin of the flank and measured. A reasonable correlation ($r = +0{\cdot}75$; $P < 0{\cdot}01$) was found between clinical and radiological measurements of extension.

(3) Radiological procedure

The basic radiological requirements for arthritis surveys laid down by the New York Symposium (Bennett and Wood, 1968) were followed and included an antero-posterior view of the pelvis taken with the subject supine.

In view of the notorious difficulty in diagnosing sacro-iliac joint abnormality because of intra and inter-observer

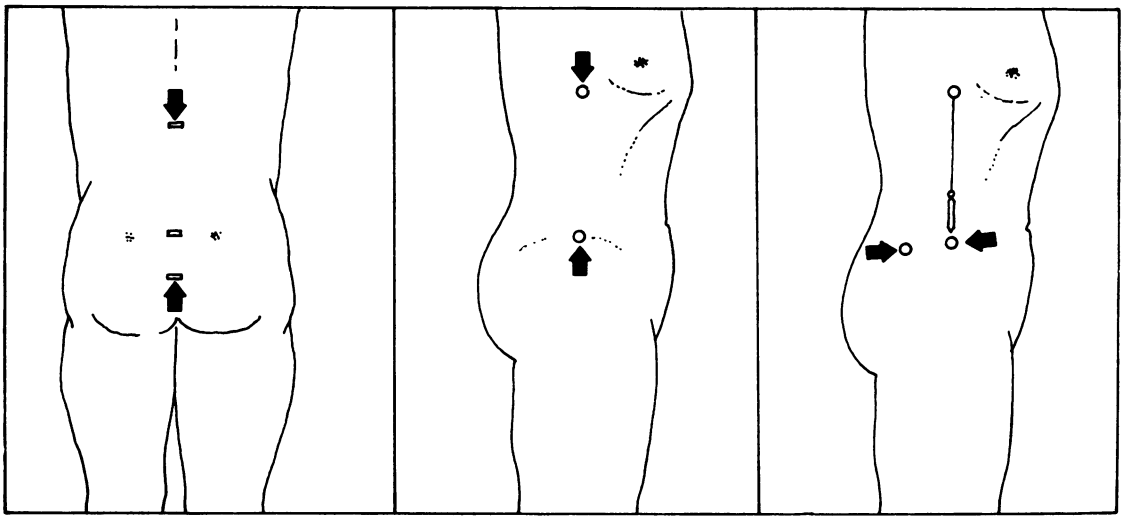

FIGURE **(A)** *Position of skin marks for measurement of anterior spinal flexion*
(B) *Position of skin marks for measurement of lateral spinal flexion*
(C) *Position of skin marks and plumb-line for measurement of spinal extension*

variation, certain precautions were taken to minimize this source of error. In order to reduce bias which might be introduced by a knowledge of the patient's clinical details, all films were read 'blind', and to obviate personal prejudice, the entire collection of pelvic films was read by two observers. The inter-observer concordance, based on the percentage of films in which agreement was achieved at the first reading, was 85 per cent.

Results

Analysis of the data for both the original and the modified criteria has been evaluated statistically by the following procedures:

(1) Distribution of positive tests;

(2) Distribution of false positive and false negative tests;

(3) Specificity and sensitivity;

(4) Overall evaluation in terms of the Youden index;

(*1*) *Distribution of* POSITIVE TESTS

The distribution of positive tests is shown in Table IV. Four points deserve particular emphasis:

(*i*) Approximately 50 per cent. of subjects in each of the relationship groups (probands, first and second-degree relatives, and spouses) were found to have thoraco-lumbar pain. It was particularly noted that in the spouse control group pain at this site was present in as many as 60 per cent. of subjects.

(*ii*) Despite the relative frequency of limitation of individual spinal movements, particularly anterior flexion and extension, only a few subjects were found to have limitation of spinal movement in *all three* directions.

(*iii*) Also noted was the significant difference ($0{\cdot}05 > P > 0{\cdot}02$; $\chi^2 = 5{\cdot}07$; d.f. = 1) between the number of subjects found to have limited chest expansion according to the original New York criteria (31 subjects) and the number diagnosed according to the modified criteria (16 subjects).

(*iv*) When spinal extension was assessed subjectively and without allowing for age or sex, 26 subjects were found to have abnormal mobility in this direction. However, assessing this movement objectively and allowing for age and sex, only twelve subjects were found to have limited extension. This difference was statistically significant ($0{\cdot}05 > P > 0{\cdot}02$; $\chi^2 = 5{\cdot}41$; d.f. = 1).

(*2*) *Distribution of* FALSE POSITIVE *and* FALSE NEGATIVE TESTS

As in previous studies (Bennett and Burch, 1968), the presence of radiological sacroiliitis (at least bilateral Grade 2 involvement) was taken to represent the true diagnostic status, and it was against this parameter that each clinical criterion was compared.

False positive and false negative results were defined as follows:

$$\text{FALSE POSITIVE (normals wrongly included)} = \frac{\text{Subjects without sacroiliitis with positive test}}{\text{Total subjects without sacroiliitis (374)}} \times 100$$

$$\text{FALSE NEGATIVE (real cases missed)} = \frac{\text{Subjects with sacroiliitis with negative test}}{\text{Total subjects with sacroiliitis (38)}} \times 100$$

Table V (overleaf) shows that, apart from the significant difference ($0{\cdot}05 > P > 0{\cdot}02$; $\chi^2 = 4{\cdot}80$; d.f. = 1) between the percentage of false positive limited chest expansion measurements, no significant differences were observed between the original and modified New York criteria evaluations. Thoraco-lumbar pain yielded the highest population of false positive results (51 per cent.) and limitation of back

Table IV *Distribution of* POSITIVE TESTS *according to the* original (*largely subjective*) *and* modified (*largely objective*) *New York criteria. Statistically significant differences between original and modified criteria are denoted* (*) *and* (†)

Relationship of subject	*Total examined*	*Original New York criteria*						*Modified New York criteria*					
		Thoraco-lumbar pain	*Limited chest expansion*	*Limited back movement*				*Thoraco-lumbar pain*	*Limited chest expansion*	*Limited back movement*			
				AF	*LF*	*Ext*	*All*			*AF*	*LF*	*Ext*	*All*
Proband	104	66	12	7	2	11	3	66	8	6	2	5	2
First-degree relatives	201	95	11	6	3	12	0	95	3	5	2	6	0
Second-degree relatives	31	11	1	0	0	2	0	11	0	0	0	1	0
Spouse controls	76	46	7	3	0	1	0	46	5	2	0	0	0
Total No.	412	218	31*	16	5	26†	3	218	16	13	4	12	2
Per cent.		52	7	4	1	6	0·7	52	4	3	1	2	0·5

* $0{\cdot}05 > P > 0{\cdot}02$; $\chi^2 = 5{\cdot}07$; d.f. = 1
† $0{\cdot}05 > P > 0{\cdot}02$; $\chi^2 = 5{\cdot}41$; d.f = 1

AF = Anterior spinal flexion
LF = Lateral spinal flexion
Ext. = Spinal extension
All = Spine limited in all three planes

Table V FALSE POSITIVE *and* FALSE NEGATIVE TESTS *results for each clinical parameter according to both the original and modified New York criteria*

(Statistically significant differences between original and modified criteria denoted (*))

Clinical criterion		*False positive test (normals wrongly included)*				*False negative test (real cases missed)*			
		Original		*Modified*		*Original*		*Modified*	
		No.	*per cent.*	*No.*	*per cent.*	*No.*	*per cent.*	*No.*	*per cent.*
Thoraco-lumbar pain		192	51	192	51	12	32	12	32
Limited chest expansion		25	7*	12	3	32	85	34	90
Limited back movement	AF	4	1	4	1	26	69	29	77
	LF	1	<1	0	0	34	90	34	90
	Ext.	15	4	7	2	27	72	33	87
	All planes	0	0	0	0	35	93	36	95

* $0{\cdot}05 > P > 0{\cdot}02$; $\chi^2 = 4{\cdot}80$; d.f = 1.

movement (all three directions) the lowest (0 per cent.). Conversely, main criteria giving rise to the highest frequency of false negative results were limitation of back movement in all three planes (93–95 per cent.) and limitation of chest expansion (85–90 per cent.). In contrast, thoraco-lumbar pain yielded only 32 per cent. false negative results.

A separate analysis revealed that 8 per cent. of subjects (1 male and 2 female relatives) with unequivocal Grade 2–4 sacroiliitis were clinically entirely normal.

(3) SPECIFICITY *and* SENSITIVITY *of each criterion*

The following definitions form the basis of this evaluation (Sartwell, 1968):

SENSITIVITY (capacity of test to give a negative response if disease present) =

$$\frac{\text{True positive}}{\text{True positive} + \text{False negative}} \times 100$$

SPECIFICITY (capacity of test to give a negative response if disease absent) =

$$\frac{\text{True negative}}{\text{True negative} + \text{False positive}} \times 100$$

Expressed differently, sensitivity is inversely proportional to false negativity and specificity is inversely proportional to false positivity. Table VI shows the sensitivity and specificity of each clinical parameter according to both the original and the modified New York criteria. No statistically significant difference was seen between the results obtained from these two sets of criteria. The highest sensitivity was shown by thoraco-lumbar pain (68 per cent.) and the lowest by limited back movement measured in all three directions (5–7 per cent.). The sensitivity of limited chest expansion as a clinical parameter was also low (10–15 per cent.). Conversely, the highest specificity was

Table VI SENSITIVITY *and* SPECIFICITY *of each clinical parameter according to the original and modified New York criteria*

Clinical criterion		*Sensitivity*				*Specificity*			
		Original		*Modified*		*Original*		*Modified*	
		No.	*per cent.*	*No.*	*per cent.*	*No.*	*per cent.*	*No.*	*per cent.*
Thoraco-lumbar pain		26	68	26	68	182	49	182	49
Limited chest expansion		6	15	4	10	349	93	362	97
Limited back movement	AF	12	31	9	23	370	99	370	99
	LF	4	10	4	10	373	>99	374	100
	Ext.	11	28	5	13	359	96	367	98
	All planes	3	7	2	5	374	100	374	100

Table VII YOUDEN INDEX *for each clinical parameter according to the original and modified New York criteria, and respective order of relative value of each clinical parameter as diagnostic indices of ankylosing spondylitis* (Statistically significant difference between original and modified criteria denoted (*))

Clinical criterion		*Youden index*		*Order of relative value*	
		Original	*Modified*	*Original*	*Modified*
Thoraco-lumbar pain		16	16	3	2
Limited chest expansion		8	7	5	5
Limited back movement	AF	30	22	1	1
	LF	10	10	4	4
	Ext.	24*	11	2	3
	All planes	7	5	6	6

* $0{\cdot}02 > P > 0{\cdot}01$; $\chi^2 = 5{\cdot}92$; d.f. = 1.

shown by limitation of back movement in all three directions (100 per cent.). Limitation of chest expansion also proved highly specific (93–97 per cent.), The specificity of thoraco-lumbar pain, however, was only 49 per cent.

(4) Overall evaluation in terms of the YOUDEN INDEX

The Youden index (Youden, 1950), previously used by Bennett and Burch (1968) to evaluate the Rome criteria for ankylosing spondylitis, was applied to the present data. The index is defined as:

$$\text{Sensitivity} + \text{Specificity} - 100$$

The value of each criterion according to the Youden index is shown in Table VII. No statistical difference in Youden index was observed between the main features (thoraco-lumbar pain, limited chest expansion, and limited back movement in all three directions) of the original and modified criteria, but a significant difference was observed between the Youden indices of spinal extension. In the adjacent columns of the same Table the respective order of relative value of each criterion is indicated. The order of value was found to be similar whether assessed by the original or modified criteria. In both evaluations limited anterior spinal flexion gained the highest score and back movement limited in all three directions the lowest.

Overall analysis of data obtained by different methods of evaluation

Table VIII summarizes the results obtained by the different statistical methods of evaluation. The relative order of value of each clinical criterion, assessed by both the original (O) and modified (M) New York criteria, is indicated numerically (*i.e.* first = 1, second = 2, third = 3, etc.). Two features illustrated by this Master Summary Table deserve particular emphasis:

Table VIII *Summary Table showing order of relative value of criteria according to different methods of evaluation*

(1 = most valuable criterion; 6 = least valuable criterion)

Clinical criterion		*Positive tests*		*False positive tests*		*False negative tests*		*Sensitivity*		*Specificity*		*Youden index*	
		O	*M*	*O*	*M*	*O*	*M*	*O*	*M*	*O*	*M*	*O*	*M*
Thoraco-lumbar pain		1	1	6	6	1	1	1	1	6	6	3	2
Limited chest expansion		2	2	5	5	4	4=	4	4=	5	5	5	5
Limited back movement	AF	4	3	3	3	2	2	2	2	3	3	1	1
	LF	5	5	2	1=	5	4=	5	4=	1=	1=	4	4
	Ext.	3	4	4	4	3	3	3	3	4	4	2	3
	All planes	6	6	1	1=	6	6	6	6	1=	1=	6	6

O = Original New York criteria. M = Modified New York criteria.

(1) With the occasional exception, the results obtained by both the original (largely subjective) and modified (largely objective) New York criteria show no significant difference. The comparability of the results obtained by these two sets of criteria is striking, regardless of the evaluation method used. It was concluded from the data that the introduction of objective methodology on an epidemiological scale resulted in surprisingly little difference in the overall results. Furthermore, where a significant difference between the objective and subjective assessment could be demonstrated, this did not affect the criterion's order of relative value as an index of ankylosing spondylitis.

(2) The order of relative value of criteria followed two general patterns according to the evaluation technique used.

On the one hand, evaluations based on *positive tests, false negative tests, sensitivity*, and the *Youden index* resulted in the order:

(*i*) Thoraco-lumbar pain,

(*ii*) Limitation of chest expansion,

(*iii*) Limitation of spinal mobility in all three directions.

On the other hand, basing the order of value on *false positive tests* and *specificity*, the sequence was reversed:

(*i*) Limitation of spinal mobility in all three directions,

(*ii*) Limitation of chest expansion,

(*iii*) Thoraco-lumbar pain.

The order of relative value of these criteria therefore depends partly on the type of evaluation method used.

Discussion

As with any criteria in general use, it is not sufficient that they be universally agreed, but also, and more importantly, that their *precise value as diagnostic indices be critically and objectively tested.* In order to achieve the latter, evaluation of the clinical criteria of ankylosing spondylitis has been carried out using both objective and subjective methods. In view of the absence of any convincing qualitative difference between psoriatic spondylitis and idiopathic ankylosing spondylitis (Fletcher and Rose, 1955; Graber-Duvernay, 1957; Wright, 1957; Dixon and Lience, 1961; Reed, 1961), it was assumed that sacroiliitis reflected the true diagnostic status, and it was against this parameter that each clinical criterion was compared.

The assumption of truth in sacroiliitis has also been adopted in previous studies (Bennett and Burch, 1968). As with most so-called true indices of disease, sacroiliitis is by no means perfect as an early and easily identifiable pointer to the diagnosis of spondylitis, as it is likely that in a proportion of subjects, albeit a minority, clinical features may precede radiological abnormality. However, in our present state of knowledge of ankylosing spondylitis, there is no more suitable non-clinical indicator of the disease than radiological sacroiliitis—especially if observations are properly controlled and acceptable radiographic techniques employed. It is important, when using this feature as the reflection of true disease status, that certain sources of potential error be considered. A separate study in our department (Macrae, Haslock, and Wright, 1971) has drawn attention to difficulties in sacroiliac joint interpretation arising from age variations, overlying bowel shadow, exaggerated lumbar lordosis, anatomical abnormalities at the lumbo-sacral junction, Paget's disease, and secondary malignant deposits. However, satisfactory grading of sacroiliac joints in population surveys can be achieved if pelvic *x*-ray examination is made by two observers independently. (This precaution was taken in the present study.) Diagnostic accuracy is further enhanced by a knowledge of the subject's age and sex. Inter-observer error was found to be particularly low in the case of sclerosis, erosion, and ankylosis, but unacceptably high for joint width.

The surprising fact that virtually the same evaluation resulted from both objective and subjective methodology suggests that objectivity is not of such central importance in contributing value to these criteria as was previously thought. Moreover, the comparability of the results obtained by objective and subjective means is consistent with the general observation that an experienced clinician often appears to arrive almost intuitively at the correct diagnosis, rather in the manner that a familiar face or place is recognized. It is curious, however, that the actual subjective processes involved in making such a diagnosis are still unknown (McGirr, 1969). Despite our inability to demonstrate any clear advantage of objectivity over subjectivity in this study, the considerable value of making measurements in clinical rheumatology remains undisputed. The value of objectivity depends not only on the obvious direct advantage based on increased accuracy of assessment but also on the indirect advantage of enabling individual correction to be made for age and sex. It was on account of the latter property that we were able to demonstrate the importance of allowing for the effect of age and sex. This was particularly evident in the case of chest expansion. In the study of false positive tests shown by limited chest expansion, it was found that seven subjects were falsely diagnosed to have limited chest expansion when no correction was made for age and sex, compared with false positive diagnosis in only three ($0{\cdot}05 > P > 0{\cdot}02$; $\chi^2 = 4{\cdot}80$; d.f. = 1) when such correction was made.

The importance of variations due to sex was examined in another context. In previous family studies of ankylosing spondylitis, apparent differences in the

distribution of asymptomatic sacroiliitis between male and female relatives have been reported. For example, in the study by Bremner, Emery, Kellgren, Lawrence, and Roth (1968), sacroiliitis unaccompanied by clinical features was found in 15·6 per cent. (20 of 128) of male relatives compared with 8·3 per cent. (5 of 60) of female relatives. In a population study by the same unit, 3·2 per cent. (4 of 122) of males but no females (0 of 66) had asymptomatic sacroiliitis. However, in our analysis of these data, neither of these sex differences was statistically significant at the 5 per cent. level. These results are therefore consistent with our own observations that asymptomatic sacroiliitis tends to be equally distributed between the sexes. There was a slight excess of asymptomatic female relatives in our series, but the sex difference was not statistically significant. In view of this, no special sex-specificity correction was considered necessary in the final evaluation of clinical criteria.

A further point concerns the inconsistency between evaluations of criteria of ankylosing spondylitis by different authors. Table IX shows the order of relative value of criteria judged by the present investigation (New York criteria) compared with the results obtained by two other groups using similar (Rome) criteria (Bennett and Burch, 1968; Chalmers and others, 1969). The striking discordance between these evaluations is self-evident and probably reflects inadequacy and incomparability of criteria rather than inaccuracy of their means of measurement. Evidence for this conclusion is based on our observation that clinical objectivity makes little difference to the results of the overall evaluation.

Abramson (1967) has drawn attention to similar differences between evaluations of clinical criteria in rheumatoid arthritis. This author also believed that lack of comparability between sets of criteria was one explanation for the inter-observer differences between evaluation studies. Differences in clinical methods and standards were also thought to be relevant. As was the case in our inter-survey comparisons, different nationalities were involved in Abramson's study, so it is possible that cultural and linguistic differences also contributed to the discordance between evaluations. However, it is likely that differences in the composition of case material provided the most potent cause of inter-evaluation differences, both in our study and in that of Abramson.

The practical value of the Youden index (sensitivity + specificity − 100), currently a popular measure of the relative value of criteria (Blumberg, 1957; Bennett and Burch, 1968; Chalmers and others, 1969), is seriously doubted. Sartwell (1968) and Acheson (1968) have also questioned the value of the Youden index. Taking thoraco-lumbar pain as an example, this was found to be a highly sensitive though poorly specific criterion. In other words, the criterion was characterized by two widely differing features. By amalgamating these two 'minor' indices (sensitivity and specificity) to form the 'major' index (the Youden index), it is clear that the individual significance of the minor indices becomes entirely obscured. It is suggested that it might be more realistic to abandon the Youden index and, instead, to report sensitivity and specificity separately. It might be argued, on the other hand, that the Youden index offers a compromise between sensitivity and specificity, and that this is what is needed in epidemiological studies. It is difficult to say what should constitute the lowest acceptable level for the Youden value in epidemiological studies of ankylosing spondylitis, but it is probably true, as Wood has stated (Wood, 1972), that in this field any value less than 20 is hopeless and any value less than 10 useless. Moreover, it should be noted that it is possible for the Youden index to have a negative value, in which case the value of criteria having negative indices would be even worse than useless. On the other hand, it is possible to have Youden values of up to 100, but this value or values approaching this figure are achieved only by ideal or optimal criteria—in this case the presence of sacroiliitis. Using this level of arbitration (*i.e.* Youden index of 20 or more), reference to Table VII suggests that anterior flexion measured objectively and subjectively (Youden indices 22 and 30 respectively) and extension measured subjectively (Youden index 24) are the only criteria likely to be of much value.

Table IX *Inconsistency between evaluations of clinical criteria for ankylosing spondylitis by various authors*

Clinical criterion	*Present study New York criteria*		*Manchester study (Chalmers and others, 1969) Rome criteria*	*Pima study (Bennett and Burch, 1968) Rome criteria*
	Original	*Modified*		
Thoraco-lumbar pain or low back pain and stiffness	1	1	1	3
Limited chest expansion	2	2	3	1
Limited back movement	3	3	2	2

As noted previously, the order of relative value of criteria depends largely on the method of evaluation used. The important implication arising from this observation is that the type of evaluation to be used must relate to the yield of positive diagnoses required. A further point is that certain relationships exist between some of the evaluation methods used. We have already drawn attention to the inverse relationship between false negativity and sensitivity, and also to a similar relationship between false positivity and specificity. Furthermore, it is often not appreciated that the Youden index can behave like sensitivity or specificity, or even Bayesian probability, depending on the amount of weight sensitivity and specificity are contributing to the index. When calculated in terms of the highest percentage of positive tests, false negative tests, sensitivity, or Youden index, thoraco-lumbar pain scored the highest value, and restriction of back movement in all three directions the lowest. However, assessed in terms of the highest proportion of false positive tests and specificity, the converse was obtained. It is clear from this dichotomy that to obtain proper insight into the value of criteria a number of statistical tests, including particularly those which reveal degrees of sensitivity and specificity, are likely to be more revealing than simply the results of a single test.

In order to examine in further depth the meaning of the evaluation results in terms of specificity and sensitivity, individual criteria will be discussed. For example, in the case of limited back movement in all three directions as an individual index of ankylosing spondylitis this was found to be excessively specific and apt to engender diagnostic loss of actual cases. This is sometimes referred to as 'an error of the first kind' (Cliffe, 1968). Thoraco-lumbar pain, on the other hand, suffers from 'an error of the second kind' in that it is too sensitive. In other words, application of this criterion will mean that relatively few real cases are lost but that a high proportion of individuals who do not have the disease are included. The place of limited chest expansion is intermediate, though nearer in criterion characteristics to limited back movement (low sensitivity; high specificity) than to thoraco-lumbar pain (high sensitivity; low specificity).

A further observation concerned the comparable pattern of behaviour of spinal mobility in each direction of movement, and it was concluded that for usual purposes, and more particularly in epidemiological work, measurement in one direction only should be adequate. In view of the fact that lateral flexion tends to be normal in cases of prolapse of the lumbar disc and abnormal in cases of ankylosing spondylitis (compared with sagittal mobility which is affected in both conditions), perhaps this plane of movement should be chosen to represent the status of spinal mobility.

As a result of the present investigation, it was concluded that in their present form the New York criteria for spondylitis require urgent re-appraisal. It is felt that criterion weighting may considerably improve the value of these criteria. It is recognized, however, that a form of criterion weighting already exists in the present system. This is based on increasing the number of clinical criteria required to fulfil the diagnosis of ankylosing spondylitis the less impressive the radiological evidence for the disease. Such a scheme is obviously relatively crude and unrealistic. We suggest that a preferable approach would be *numerical* weighting of criteria, such as has been done previously in thyrotoxicosis (Crooks, Murray, and Wayne, 1959). This might be done by finding the best compromise between sensitivity and specificity for each criterion by means of another application of probability theory (the likelihood ratio) which has been recently outlined by Cliffe (1968).

However, the definitive question concerning the level of the absolute value necessary to constitute a positive diagnosis remains arbitrary. This would naturally be related to the intrinsic characteristics of the disease under study and concerns the inevitable problem of deciding the relative proportion of real cases one wishes to diagnose at the expense of including entirely normal individuals. In the case of a treatable lethal disease, the sensitivity level would clearly have to be set at 99 per cent. or more. However, in the case of a non-lethal condition such as ankylosing spondylitis, a lower level of sensitivity would probably be more acceptable. Furthermore, this would obviate the undesirable probability of subjecting falsely diagnosed normal individuals to potentially hazardous therapy such as irradiation. On the other hand, considering the therapeutic possibilities in spondylitis, the sensitivity level would not have to be set too low; otherwise real cases, and thus potential candidates who would benefit from early treatment, might be missed.

Summary

The New York clinical criteria for ankylosing spondylitis have been evaluated. In a study of 412 subjects, clinical examination was carried out subjectively and objectively. Criterion evaluation was applied to both subjective and objective data.

The following points emerged from the investigation:

(*a*) No overall difference was found between subjective and objective results;

(*b*) Calculated in terms of frequency of *positive* and *false negative tests*, *sensitivity*, and the *Youden index* (sensitivity + specificity − 100), criteria were listed in the following order of value:

(*i*) Thoraco-lumbar pain,

(*ii*) Limited chest expansion,

(*iii*) Limited back movement.

However, in terms of *false positive tests* and *specificity*, the reverse order was obtained.

It was calculated that individual criteria in their present form fail to provide a satisfactory diagnostic index of ankylosing spondylitis. Thoraco-lumbar pain is too sensitive and too non-specific, and limited chest and back mobility are too insensitive and too specific. In order to obviate this problem it is suggested that criteria should be numerically weighted.

References

ABRAMSON, J. H. (1967) *J. chron. Dis.*, **20**, 275 (On the diagnostic criteria of active rheumatoid arthritis)

ACHESON, R. M. (1968) *In* 'Population Studies of the Rheumatic Diseases', ed. P. H. Bennett and P. H. N. Wood, p. 312. Excerpta Medica Foundation, Amsterdam (Int. Congr. Ser. No. 148)

BENNETT, P. H., AND BURCH, T. A. (1968) *Idem*, pp. 305–313

—— AND WOOD, P. H. N. (1968) *Idem*, p. 456

BLUMBERG, M. S. (1957) *Operations Research (Baltimore)*, **5**, 351 (Evaluating health screening procedures)

BREMNER, J. M., EMERY, A. E. H., KELLGREN, J. H., LAWRENCE, J. S., AND ROTH, H. (1968) *In* 'Population Studies of the Rheumatic Diseases', ed. P. H. Bennett and P. H. N. Wood, pp. 299–304. Excerpta Medica Foundation, Amsterdam

CHALMERS, T. M., DANCHOT, J., KELLGREN, J. H., KING, D., PIKHLAK, E., SIEVERS, K., STREVENS, E., TAIT, B., AND WOOD, P. H. N. (1970) *Ann. rheum. Dis.*, **29**, 200 (A test of diagnostic criteria—experience in England and Wales Heberden Society, September, 1969. Abstract)

CLIFFE, P. (1968) 'Computers in medicine' *in* 'Recent Advances in Medicine', ed. D. N. Baron, N. Compston, and A. M. Dawson, 15th ed., pp. 13–14. Churchill, London

CROOKS, J., MURRAY, I. P. C., AND WAYNE, E. J. (1959) *Quart. J. Med.*, **28**, 211 (Statistical methods applied to the clinical diagnosis of thyrotoxicosis)

DIXON, A. ST. J., AND LIENCE, E. (1961) *Ann. rheum. Dis.*, **20**, 247 (Sacroiliac joint in adult rheumatoid arthritis and psoriatic arthropathy)

FLETCHER, E., AND ROSE, F. C. (1955) *Lancet*, **1**, 695 (Psoriasis spondylitica)

GRABER-DUVERNAY, J. (1957) *Rev. Rhum.*, **24**, 288 (À propos de la spondylarthrite psoriasique)

KELLGREN, J. H., JEFFREY, M. R., AND BALL, J. (eds) (1963) 'The Epidemiology of Chronic Rheumatism', vol. 1, p. 326. Blackwell, Oxford

MACRAE, I. F., HASLOCK, D. I., AND WRIGHT, V. (1971) *Ann. rheum. Dis.*, **30**, 58 (Grading of films for sacro-iliitis in population studies)

—— AND WRIGHT, V. (1969) *Ibid.*, **28**, 584 (Measurement of back movement)

MCGIRR, E. M. (1969) *In* 'Computers in Medicine', Proc. Symp. Computers in Medicine, Blackburn College of Technology and Design, 1968, ed. J. Rose, pp. 19–29. Churchill, London

MOLL, J. M. H. (1971) 'A Family Study of Psoriatic Arthritis'. D.M. thesis, University of Oxford

——, LIYANAGE, S. P., AND WRIGHT, V. (1972a) *Rheum. phys. Med.*, **11**, 225 (An objective clinical method to measure lateral spinal flexion)

——, ——, —— (1972b) *Ibid.*, **11**, 293 (An objective clinical method to measure spinal extension)

—— AND WRIGHT, V. (1971) *Ann. rheum. Dis.*, **30**, 381 (Normal range of spinal mobility: an objective clinical study)

——, —— (1972) *Ibid.*, **31**, 1 (An objective clinical study of chest expansion)

REED, W. B. (1961) *Acta derm.-venereol. (Stockh.)*, **41**, 396 (Psoriatic arthritis. A complete clinical study of 86 patients)

SARTWELL, P. E. (1968) *In* 'Population Studies of the Rheumatic Diseases', ed. P. H. Bennett and P. H. N. Wood, p. 312. Excerpta Medica Foundation, Amsterdam

SCHOBER, P. (1937) *Münch. med. Wschr.*, **84**, 336 (Lendenwirbelsäule und Kreuzschmerzen)

WOOD, P. H. N. (1972) Personal communication

WRIGHT, V. (1957) *Brit. J. Radiol.*, **30**, 113 (Psoriasis and arthritis: a study of the radiographic appearances)

YOUDEN, W. J. (1950) *Cancer (N.Y.)*, **3**, 32 (Index for rating diagnostic tests)