

Statistical approach to indices of disease activity in rheumatoid arthritis

With reference to a trial of indomethacin

R. J. MCGUIRE¹ AND V. WRIGHT²

From the Department of Psychiatry and the Rheumatism Research Unit, University of Leeds

When any drug trial is being carried out it is essential that some criterion of improvement be available to assess the efficacy. In the ideal case there is one single measure, reliably and validly assessed, which is acceptable as this criterion. More frequently, however, there are several possible criteria. The usual practice is to deal with each separately and then in the discussion to point out which symptoms responded to treatment. Clinically this has obvious advantages in suggesting the appropriate applications of the treatment, but reliability will be increased if the various scores can be combined in some meaningful way. The unsophisticated way to do this is to add together the various scores, but there may be some clinical justification for weighting major symptoms more than minor ones and this can be done.

In rheumatoid arthritis, Lansbury (1958) has been a keen advocate of indices. His Articular Index weighted the score on each joint by the area of joint (Lansbury and Haut, 1956) and his Systemic Index based on morning stiffness, fatigue, aspirin need, grip strength, and erythrocyte sedimentation rate was believed to measure the inflammatory aspects of the disease (Lansbury, 1956).

A double-blind cross-over trial of indomethacin, previously undertaken (Wright, Walker, and McGuire, 1969) in 24 patients with rheumatoid arthritis to investigate the value of the drug compared with phenylbutazone and placebo, was set up in such a way that the Lansbury Systemic Index could be evaluated and the place of other indices determined. The results of this aspect of the study are reported in this paper.

Plan of trial

The trial comprised 24 patients with definite rheumatoid arthritis (Ropes, Bennett, Cobb, Jacox, and Jessar, 1959)

who were assessed in four stages: before treatment, and at the end of three 4-week periods of treatment with indomethacin, phenylbutazone, or placebo, which was given on a graded dosage schedule (Fig. 1). The full details of the trial are contained in the previous communication (Wright and others, 1969).

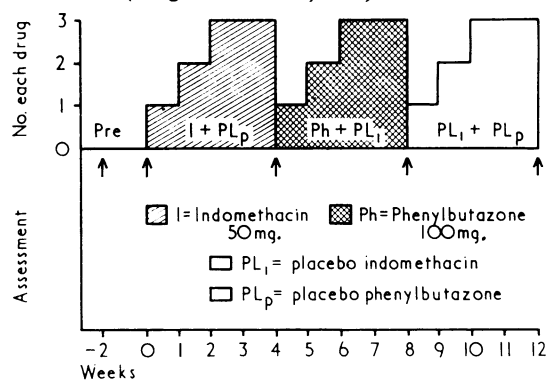


FIG. 1 Plan of trial. The three drug phases were permuted to balance the effects of order of treatment. The stepping in each phase represents increasing dosage, unless side-effects contraindicated it.

Assessments

A full clinical, radiographical, and serological examination of all patients was performed. The functional capacity was classified as normal, adequate, limited, or incapacitated.

Radiographs of the hands and feet were taken, and changes due to rheumatoid arthritis were graded as absent, doubtful, mild, moderate, or severe. The indices used to judge progress at the start of the trial and at the end of each 4-week period were:

(1) Patient's assessment of the degree of pain, recorded as nil (0), mild (1), moderate (2), severe (3), unbearably severe (4).

(2) Patient's assessment of restriction of movement, recorded as nil (0), mild (1), moderate (2), severe (3), completely crippled (4).

Accepted for publication May 12, 1971.

¹ Senior Lecturer in Clinical Psychology.

² Professor of Rheumatology.

This paper was originally presented at a meeting of the Measurement Section of the Royal Society of Medicine, on June 5, 1970.

- (3) Duration in hours of morning stiffness.
- (4) Duration in hours to the onset of fatigue after rising.
- (5) Average number of aspirin tablets taken per day (these were given as aspirin encephal 300 mg.).
- (6) Strength of grip (mm. Hg), the mean of two grips being taken for each hand, the right hand being used first.
- (7) Haemoglobin level (per cent.).
- (8) Erythrocyte sedimentation rate (mm. in first hour).
- (9) White blood count.

The Lansbury Systemic Index was then calculated from morning stiffness, fatigue time, aspirin intake, grip strength, and erythrocyte sedimentation rate (Lansbury, 1958).

Results

Tables I and II summarize the results of the trial, and it will be seen that the Lansbury Index was not significantly altered by either of the active preparations, although it contained measures which, when taken individually, did show a significant difference with both indomethacin and phenylbutazone compared with placebo. This suggested that the weights used by Lansbury in calculating his index were not optimal. More statistically sophisticated approaches to indices were therefore investigated.

MAXIMIZATION OF STATISTICAL SIGNIFICANCE

The first method was to try to maximize the statistical significance of an index. It will be appreciated that statistical significance between two means is normally measured by a t-test, which is the ratio of the difference of the means to the standard error of that

Table II Comparison of mean scores of nine variables, taking the treatments in pairs, order effect ignored

Variable	Placebo v. Indomethacin	Placebo v. Phenylbutazone	Indomethacin v. Phenylbutazone
Pain	**	*	-
Restriction of movement	-	-	-
Morning stiffness	-	*	-
Fatigue	-	-	-
Aspirin	*	*	-
Grip strength	-	*	-
Hb	-	*	-
Erythrocyte sedimentation rate (mm./hr)	-	*	-
Lansbury Index	-	-	-

** P < 0.01; * P < 0.05; - Not statistically significant.

difference. When response to treatment is being considered, the patient concerned acts as his own control to reduce variance due to individual differences. The problem is then one of deciding

Table I Mean scores of variables assessed in trial

Measure	Overall		Order				Drugs			
	Before	After	1st	2nd	3rd	P	Placebo	Indomethacin	Phenylbutazone	P
Pain	2.75	2.04	2.08	1.87	2.17	-	2.37	1.79	1.96	*
Restriction of movement	2.13	1.81	1.88	1.75	1.80	-	1.88	1.88	1.67	-
Morning stiffness (hrs)	3.38	3.64	3.83	2.88	4.21	-	4.25	3.65	3.02	-
Fatigue (hrs)	6.38	6.96	6.83	7.25	6.80	-	6.74	7.24	6.89	-
Aspirin tablets	8.58	7.00	7.08	6.92	7.00	-	7.75	6.62	6.62	-
Grip strength (mm. Hg.)	107.3	111.3	109.3	113.0	111.5	-	106.8	109.9	116.2	-
Hb (per cent.)	85.1	83.7	83.3	83.5	84.3	-	86.1	83.5	81.5	-
Erythrocyte sedimentation rate (mm./hr)	54.7	49.5	55.5	46.4	46.6	-	44.6	49.3	54.6	-
Lansbury index	82.46	75.08	80.29	67.78	77.18	*	78.58	74.68	71.98	-

- Mean differences not statistically significant. * P < 0.05 i.e. probably significant.

whether the mean difference score (Z) is significantly different from zero by means of:

$$t = \frac{\text{Mean } (Z) \times \sqrt{N}}{\text{Standard deviation } (Z)}$$

In the case of an index, Z is not a single measure but is for each individual the weighted sum of the variables involved in the index

$$Z = W_1 X_1 + W_2 X_2 + \dots$$

It should therefore be possible to find weights for the variables which will maximize the value of 't'. Such weights are found by well-known mathematical techniques reducing to a matrix form

$$W = D^{-1} M$$

where W and M are the vectors of weights and means respectively and D^{-1} is the inverse of the variance-covariance matrix. The resulting value of 't' can be found at the same time by means of the formulae

$$t = \sqrt{(M^T W) N} \text{ or } t = \sqrt{(M^T D^{-1} M) N}$$

where M^T is the transpose of the vector of means and N is the total number of cases. To find the comparable value of P , it should be noted that this t has $(N - q)$ degrees of freedom, where q is the number of variables used in the index, but one can have confidence in the statistical significance of the new index only after one has cross-validated the weights by using them on a new series of cases. The appropriate degrees of freedom will then be the usual, *i.e.* one less than the number of cases. A computer can perform all the matrix operations involved in the above equations in a few seconds so that the weights and their likely values can be found for various combinations of variables one might wish to include in the index.

Although at first it might appear that all the variables should be included, it is apparent that only beneficial changes should be maximised. In the

present trial, the technique can be illustrated by application to the changes when indomethacin is compared with the initial 'no-treatment' phase. It is realized that this comparison includes the placebo effects as well as the true effects of indomethacin, but an index aims to include *all* reversible changes in the disease (Lansbury, 1968). In a controlled trial using the index, the placebo effects would be cancelled out.

Table III shows that with indomethacin, six variables improved while three deteriorated, although only two measures were statistically significant. Using these data, it is possible to calculate an improvement index and a deterioration index for the effect of the drug.

Table IV (opposite) shows the maximum value of t obtained when the optimum weights were calculated for increasing numbers of the variables. The reason that the two variables which are most significant on their own do not give the highest t when combined is presumably because they overlap in the relevant aspect of the response. Table III also shows that when the two variables corresponding to analgesic response are omitted the maximum t value equal 2.37, which is significant at the 5 per cent. level, but is much less than the value 5.20, significant at the 0.1 per cent. level, obtained when only the two analgesic measures were used. This confirmed the impression that the main effect of indomethacin is to produce analgesia.

Direct comparison can be made with the Lansbury Systemic Index which uses five of the variables listed. With Lansbury's own weights the value of t obtained in the case of indomethacin is 1.93, a result which is not significant, but with the best weights for these same variables $t = 3.21$, significant at the 1 per cent. level.

The same technique can be applied to the deleterious effects of the treatment. It emerges, however, that the maximum t value ($t = 1.18$) falls below the level of significance, suggesting that these effects were due to chance and might not be found in other trials of the drug.

Table III Mean improvement scores on indomethacin compared with no treatment. Negative sign indicates deterioration on that variable for the average patient

Variable	Improvement score		t	P
	Mean	s.d.		
(1) Pain (0-4)	+0.96	0.95	4.92	< 0.001
(2) Restriction of movement (0-4)	+0.25	0.79	1.54	> 0.1 Not Sig.
(3) Morning Stiffness (hrs)	-0.29	2.66	0.54	> 0.1 Not sig.
(4) Fatigue onset (hrs)	+0.94	2.76	1.66	> 0.1 Not sig.
(5) Aspirin needed (no./day)	+1.96	3.36	2.85	< 0.01
(6) Grip (mm. Hg)	+3.50	28.0	0.61	> 0.1 Not sig.
(7) Haemoglobin (per cent.)	-1.54	9.2	0.82	> 0.1 Not sig.
(8) E.S.R. (mm./1st hr)	+5.33	27.4	0.96	> 0.1 Not sig.
(9) W.B.C. ($\times 100$ /cu. cm.)	-4.13	19.2	1.05	> 0.1 Not sig.

Table IV Values of *t* obtained for various indices when optimum weights are used. The numbers refer to the variables listed in Table III

Variables included in index	Value of <i>t</i> for index	<i>P</i>
(a) Variables which improved		
1	4.92	< 0.001
1, 6	5.47	< 0.001
1, 5, 6	5.86	< 0.001
1, 2, 5, 6	5.88	< 0.001
1, 2, 5, 6, 8	5.89	< 0.001
1, 2, 4, 5, 6, 8	5.89	< 0.001
(b) Variables which deteriorated		
9	1.05	> 0.1 Not sig.
7, 9	1.17	> 0.1 Not sig.
3, 7, 9	1.18	> 0.1 Not sig.
(c) Variables measuring analgesia		
1, 5	5.20	< 0.001
(d) Variables not directly measuring analgesia		
2, 3, 4, 6, 8	2.37	< 0.05
(e) Lansbury Systemic Index		
3, 4, 5, 6, 8		
i) Lansbury weights	1.93	> 0.1 Not sig.
ii) Optimum weights	3.21	< 0.01

It is important to note that these indices give scores for each patient, so that examination of the patients with high amelioration or deterioration

index scores might suggest that certain features are indications or contraindications for the treatment. In this study, however, neither age, sex, duration of illness, x-ray grade, or functional status could on their own be used to predict improvement or relapse on either of the index scores (Table V). A multiple regression equation using these variables might have given accurate prediction, but it was believed better to await cross-validation of the index by a later study before proceeding to this stage, lest the sophistication of the mathematical techniques outrun the reliability of the measures.

Fig. 2 illustrates the distribution of individual patients according to the scores on the six-variable

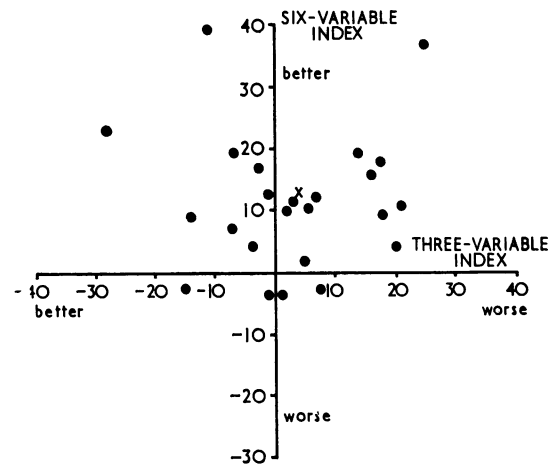


FIG. 2 24 patients plotted according to their scores on the six-variable (improvement) index and the three-variable (deterioration) index. The origin represents no change on either index and the directions better or worse are indicated for each index.

Table V Analysis of variance results for the six-variable (improvement) and three-variable (deterioration) indices, related to the clinical and radiological status of the patients at the beginning of the trial. None of the differences was statistically significant

Classification	Status	Frequency	Six-variable index (improvement)			Three variable index (deterioration)		
			Means	<i>F</i>	<i>P</i>	Means	<i>F</i>	<i>P</i>
(1) Functional status	Adequate	12	8.2	3.00	—	1.3	0.44	—
	Limited/incapacitated	12	15.5			4.8		
(2) X-ray grade	Mild	10	11.1	0.08	—	2.6	0.02	—
	Moderate/severe	14	12.4			3.4		
(3) Duration of arthritis (yrs)	≤ 7	13	11.1	0.12	—	2.7	0.02	—
	> 7	11	12.7			3.5		
(4) Age (yrs)	≤ 55	10	10.1	0.43	—	-2.5	3.58	—
	> 55	14	13.1			7.0		
(5) Sex	Male	4	6.5	1.19	—	2.4	0.01	—
	Female	20	12.9			3.2		

improvement index and the three-variable deterioration index. It is clear that the two indices are comparatively independent of each other, *i.e.* there are two separate effects of the drug, one improving certain variables to a significant level and the other, although not significant, leading to deterioration on certain variables for the average patient. The top right-hand quadrant contains those patients (50 per cent.) who on the whole improved on the six variables but deteriorated on the other three. The top left-hand quadrant contains the eight patients who had greatest benefit from the drug since they improved on both the indices, *i.e.* they had a positive score on the improvement index and a negative score on the deterioration index. Two patients became worse on both and two gave negative scores on both.

FACTOR ANALYSIS

The second statistical approach to indices is that of factor analysis, a method which has been widely used in psychological research. When two variables correlate highly with each other it suggests that there may be an underlying factor or factors which cause the two variables to maintain a certain proportion, and it would be interesting to know what these factors were. Principal component analysis is a mathematical technique which extracts factors from the correlation of many variables, producing simplification of the data from a mathematical point of view.

Fig. 3 shows two factors extracted from the nine variables by this technique. These two factors summarize more than half the information on individual differences contained in the nine variables and illustrate the mathematical simplification produced by factor analysis. The closeness of a variable to the direction of an axis represents the correspondence between the two, so that, for

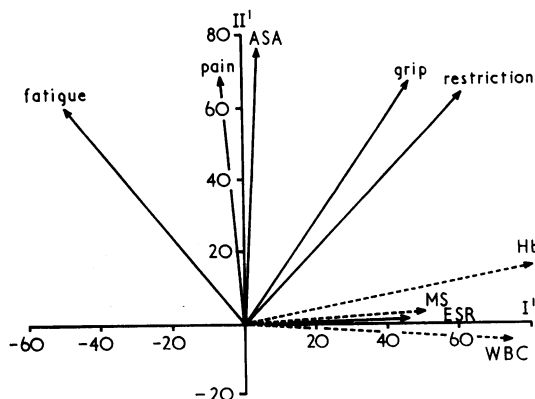


FIG. 3 Factor loadings of changes of the nine variables on two factors which have been extracted by Principal Component Analysis. Variables that showed a deterioration on the drug are indicated by broken lines.

example, fatigue almost completely corresponds to Factor II. Whenever this mathematical approach does not yield clinically meaningful factors it is quite legitimate to rotate the axes to see if the situation can be improved. It is important while doing this to recognize that the variables themselves must not change position. If the axes are rotated to the positions shown in Fig. 4, it will be seen that this is a significant clarification from a clinical point of view, in that the two variables which improved most on treatment line up closely with the new Factor II¹ and the three variables which deteriorated lie in the direction of Factor I¹. The fact that the axes are at right-angles means that they are independent of each other. It therefore follows that the improvement scores of Factor II¹ are unrelated to the deterioration scores of Factor I¹.

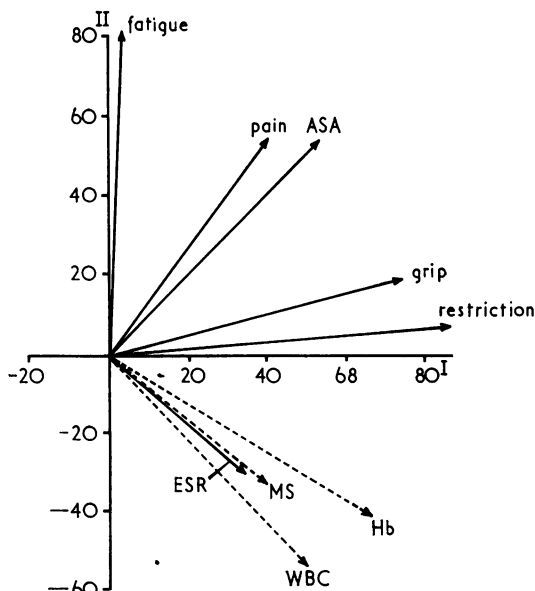


FIG. 4 Axes of Fig. 3 rotated (by approximately 40° clockwise) to line up with the main variables. Variables which improved (except for ESR) line up with one factor, while those which deteriorated line up with the other.

When there are more than two significant factors, it is difficult to rotate the axes by visual inspection, but the technique of varimax rotation carried out by a computer can line up several factors with the variables.

While it is possible to derive weights for each of the variables to obtain factor scores, this is probably not the best way to use the results of the factor analysis, because it is seldom that the factors obtained will produce results which are so clear cut concerning improvement and deterioration.

Probably indomethacin has only an analgesic effect and thus only one improvement axis emerges.

With other drugs and/or diseases there may be two or more aspects of improvement, so that the improvement variables cover each in proportion to the extent to which each factor is involved in the measurement of the variable. In such circumstances, factor analysis will suggest which variables should be used for each improvement index, but the actual weights should then be calculated using the maximum-significance method. Each patient would then have two or more improvement indices relatively independent of each other. Drug comparisons might then suggest different drugs for different patients according to which therapeutic response is most needed.

Table III shows that a simplified five-variable improvement factor would probably be as useful and as reliable as one derived from all six variables which improved. The weights for practical use of this Rheumatoid Index are shown in Table VI. It must be emphasized that the indices used here have been calculated after the outcome of the treatment was known and they must be confirmed by a cross-validation study. It is unlikely that the values of t found in the follow-up will be quite as high as those found originally, but the non-significance of the deterioration index shows that the technique does not automatically produce significant values of t when one is being wise after the event.

Table VI *Optimal practical weights found in present study*

<i>Variable</i>	<i>Measure</i>	<i>Weight</i>
Pain	0-4	12
Restriction	0-4	2
Aspirin need	No./day	1
Grip strength	mm.Hg	0.25
Fatigue onset	hrs	0.1

Discussion

Not everyone is agreed about the value of indices and one eminent rheumatologist has stated:

'I do not think that artificial indices of improvement, obtained by adding together various different criteria, serve any purpose other than to muddle the whole affair' (Bywaters, 1960).

Similarly Mainland (1967) has expressed his reservations:

'Many of us seem to think that if we attach a single-number value to a complex phenomenon, we know more about it than we did before. We forget that all numbers are abstractions and that a qualitative description may be nearer to reality than an index.'

Mainland concluded that even a universally-used composite index would be unlikely to provide satisfactory information concerning the actual condition or changing condition of a patient or groups of patients. This is not necessarily true; the proof of the index is whether it works—which is the position Lansbury (1968) has taken in defence of his Index.

The present study was devised against the background of this controversy. It demonstrated that in this trial the Lansbury Systemic Index did not give significant results although certain individual measures within it did so. Nevertheless, since in rheumatoid arthritis there are subjective phenomena which for the sake of the patient cannot be ignored, steps must be taken to increase their reliability and this is best performed by an appropriate index. Little is to be gained by replacing the subjective impression of the patient by the overall assessment of the clinician, which is equally subjective. Although still largely based on the patient's reports, indices have been calculated by maximization of statistical significance and factor analysis to improve their reliability. For practical purposes a Rheumatoid Index, comprising weighted scores for pain, restriction of activity, aspirin need, grip strength, and onset of fatigue, has been shown to be the most useful in the present study.

Both the statistical approaches to indices described here are flexible procedures; the type and number of variables to be combined can be altered by the experimenters. As here, one can include a variety of measures and conclude from the results that the drug undergoing trial is primarily analgesic. If one had been sure of the clinical effects of the drug, several measures of analgesia could have been included and other measures omitted, thus increasing the reliability of the final index. Similarly, separate aspects of rheumatoid arthritis and other illnesses can produce their own specific indices. Because of the need for cross-validation, the best way to use an index for the first time would be to perform a pilot study on a few cases and then use the weights thus found in the full-scale trial of the drug. If it happens that a drug has two positive effects which are relatively unrelated, two separate indices could be calculated, *e.g.* an analgesic index and an articular index. The drug could then be compared on the appropriate index with other drugs with only a single effect. It is also possible, where two active drugs have been compared, to calculate indices for each of the active drugs and compare the effects on individual patients. In this trial indomethacin and phenylbutazone were, in fact, shown to have mainly analgesic effects and not to be significantly different from each other, but it could well have been that the two drugs had different actions and therefore the

simple question, 'Is A better than B?', is no longer appropriate, since each drug may be suitable for different kinds of patients.

Using this approach it may be possible to predict those patients who will respond to a particular type of therapy. This could in certain cases run across normal clinical classification. Clinicians are traditionally critical of a symptomatic approach to classification, but if difference in response between patients can be clearly demonstrated it raises the question of the homogeneity of the group, particularly where no aetiological agent has been isolated.

These techniques are relatively novel in clinical medicine and are capable of wide application to diseases other than rheumatoid arthritis. The authors will be pleased to give more detailed information to other workers who wish to use these approaches in their own research.

Summary

The value of indices in the assessment of rheumatoid arthritis has been determined in a double-blind cross-over trial of indomethacin compared with phenylbutazone and placebo. The Lansbury Systemic

Index was calculated and shown to have little value in this study. Two other mathematical techniques, factor analysis and a maximization of statistical significance, were used to calculate indices from the data. Clinically significant results were obtained by both methods, demonstrating indomethacin to be an effective analgesic. A Rheumatoid Index, comprising weighted scores for pain, restriction of movement, aspirin need, grip strength, and onset of fatigue, has been shown to be the most useful measure of improvement in this study.

It is suggested that the two mathematical techniques discussed here can be used in separating different aspects of any illness and in deciding which drugs may be effective in dealing with them. Thus in rheumatoid arthritis the analgesic, anti-inflammatory, and other effects of a drug can be ascertained. Advice on the techniques devised by the authors is freely available.

The authors would like to thank Dr. Ardie Lubin, Research Psychologist with the U.S. Navy Medical Neuropsychiatric Research Unit, San Diego, California, for his helpful and stimulating comments while this research was being carried out.

References

- BYWATERS, E. G. L. (1960) 'Treatment and illustrative answers', in 'Controlled Clinical Trials', C.I.O.M.S. Conference, chairman A. Bradford Hill, p. 75. Blackwell, Oxford.
- LANSBURY, J. (1956) *Amer. J. med. Sci.*, **232**, 300 (Quantitation of the activity of rheumatoid arthritis, 5. A method for summation of the systemic indices of rheumatoid activity)
- (1958) *Arthr. and Rheum.*, **1**, 505 (Report of a three-year study on the systemic and articular indices in rheumatoid arthritis)
- (1968) *Ibid.*, **11**, 599 (Clinical appraisal of the activity index as a measure of rheumatoid arthritis).
- AND HAUT, D. D. (1956) *Amer. J. med. Sci.*, **232**, 150 (Quantitation of the manifestations of rheumatoid arthritis. 4. Area of joint surfaces as an index to total joint inflammation and deformity).
- MAINLAND, D. (1967) *Arthr. and Rheum.*, **10**, 71 (The estimation of inflammatory activity in rheumatoid arthritis: role of composite indexes).
- ROPES, M. W., BENNETT, G. A., COBB, S., JACOX, R., AND JESSAR, R. A. (1959) *Ann. rheum. Dis.*, **18**, 49 (1958 Revision of diagnostic criteria for rheumatoid arthritis).
- WRIGHT, V., WALKER, W. C., AND MCGUIRE, R. J. (1969) *Ibid.*, **28**, 157 (Indomethacin in the treatment of rheumatoid arthritis).