Gut, 1987, **28**, 474–481

Special report

IOIBD* report no 1: observer variation in calculating indices of severity and activity in Crohn's disease

F T DE DOMBAL AND A SOFTLEY

SUMMARY Observer variation in calculating various indices for estimating the severity and activity of Crohn's disease is reported. Seven prospective users of Crohn's disease activity indices were presented with 10 'cases histories' compiled from relevant patient data and asked to calculate (independently) various indices of severity and activity from this information. The results showed a disquieting degree of observer variation in all indices studied. Similar results were obtained when 15 members of the International Organisation for the Study of Inflammatory Bowel Disease (IOIBD) reviewed one case history and also when members independently reviewed a series of real life cases. It is suggested that each index so far proposed is open to considerable observer variation, which casts some doubt upon the validity of studies so far carried out. Clarification of indices and the use of ranking methods can, however, do much to overcome this discrepancy.

Rapid, reliable, and reproducible estimation of the severity and activity of Crohn's disease at a given time ranks second only in importance to accuracy of diagnosis in its implications – both for the management of an individual patient and for the conduct of clinical therapeutic trials.

In the 1970's a major step forward was taken with the development of the Crohn's disease Activity Index (CDAI) for use in the National Cooperative Crohn's Disease Study[1] (NCCD), to allow uniform decentralised clinical evaluation and decision making throughout the period of the study.[2]

The original CDAI has since undergone modification by Harvey and Bradshaw,[3] by the World Organisation of Gastroenterology,[4] and by the NCCD authors themselves.[5] Other indices (or lists of factors affecting severity) have been published,[6–8] or widely communicated (Table 1). Yet there is a lack of study relating to the reproducibility of each index; and the practical value of an index which is poorly reproducible must be open to considerable question. An index may be poorly reproducible for two main reasons. First, there may be variation in eliciting data, particularly in respect of subjective clinical data. In addition problems may arise in calculating indices (even from objective data) if the criteria for calculation and/or the method of doing so are not clear to the user.

In this report from the International Organisation for the Study of Inflammatory Bowel Disease (IOIBD) we present the result of studies designed to explore both points.

Methods

MATERIAL

There were four 'parts' to the present study – the first three deal with calculation of indices, the fourth with elicitation of patient data.

Part 1

The first part of this study involved a group of seven prospective users of Crohn's Disease Activity Indices (five consultants in surgery or gastroenterology, and two research assistants). Each was provided with 10 'case histories' (Fig. 1) and asked to calculate (for each 'patient'), eight indices of severity.

Part 2

Next, 15 clinicians attending the 1984 (Copenhagen) meeting of the IOIBD were presented with case data from a single 'patient' (Fig. 1), and asked to calculate each of eight indices for this 'patient'.

Part 3

During this process, considerable discussion took place and some obvious causes of discrepancy were

Address for correspondence: Clinical Information Science Group, University Department of Surgery, St James University Hospital, Leeds LS9 7TF.

Received for publication 12 August 1986.

*IOIBD=International Organisation for the Study of Inflammatory Bowel Disease. (For membership see Appendix 1.)

Table 1 *Showing Attributes estimated in, and overlap between, various 'indices of severity' in Crohn's Disease*

Attribute assessed	*IOIBD assessment*	*Harvey & Bradshaw CDAI*	*S African assessment*	*Dutch AI*	*CDAI*	*European SAI*
Pain	X	X	X		X	X
Bowel habit	X	X	X	X	X	X
Perianal complications	X		X			X
Fistula	X					
Other complications	X	X	X	X	X	X
Mass	X	X	X	X	X	X
Body weight/build	X		X		X	X
Temperature	X		X	X		X
Tenderness	X		X			
Haemoglobin/haematocrit	X		X		X	X
Albumin	X			X		X
General wellbeing		X	X		X	
ESR				X		
'Quetelet' score				X		
Sex				X		
Prev. resection				X		
Drugs					X	

Fig. 1 *Sample of case history used in present study (together with diary card) for calculating CDAI (not shown here)*

PATIENT NAME: A ASSESSOR:

REG NO: 12345 STATUS:

BASIC DATA: Age: 35 Male Height: 5' 10" (1 m 78) Weight: 140 lbs (64kg)

HISTORY: Patient known to have Crohn's Disease for five years. Previous resection right colon and 10 cms terminal ileum four years ago

Patient produced 7 day card immediately on presentation, filled out as shown overleaf:

On interrogation patient claims to be having moderate left-sided abdominal pain usually at defaecation, no pain to-day. Bowels are normally open about 4–5 times per day, some motions are looser than others. Some blood in stool but no slime.

Patient records indicate that patient has suffered from pain in both knees, erythema nodosum and iritis. Drugs being taken for diarrhoea include Salazopyrine (1 gm tds) and codeine phosphate.

EXAMINATION: On examination looks reasonably well. Patient is afebrile. On examination of abdomen there is tenderness in lower half abdomen with suspicion of mass in right lower quadrant. Patient looks well nourished and no fistula is seen. On rectal examination however, patient has fistula in ano. No other abdominal findings.

INVESTIGATIONS: Recent investigations are as follows:-

Haemoglobin 10·5 grms% Albumin 3·2 grms/litre ESR 25 mm (1 hr)

Endoscopy reveals friable mucosa with small ulcers in rectum up to 12 cms at which point mucosa becomes normal.

Biopsy in past has been reported as 'typical Crohn's Disease' with granulomata and giant cells.

Radiology in last few months has revealed apparent recurrence of Crohn's Disease proximal to line of resection extending 10 cms with narrowing of bowel, and further changes suspicious of Crohn's Disease in rectum and rectosigmoid.

FOR PATIENTS PLEASE CALCULATE:

1. CDAI ..... 2. DUTCH AI .....
3. 'HARVEY BRADSHAW' SCORE ..... 4. BEST-BECKTEL .....
5. LEEDS CDAI SCORE ..... 6. OXFORD SCORE .....
7. EUROPEAN SCORE ..... 8. SOUTH AFRICAN SCORE .....

Finally, please add your own assessment of this patient's 'severity' of disease

Very mild |——————————————————————| Very severe

identified and clarified. In the third part of this study, seven further prospective index users (once again five consultants and two research assistants) repeated the study outlined in Part 1, to determine whether such clarification had any effect.

Part 4

Finally, at the 1985 IOIBD meeting (in Jönkoping, Sweden) a series of six patients known to have Crohn's disease were interviewed and examined by a panel of six senior and experienced gastroenterologists familiar with Crohn's disease and its assessment. Patient case histories were also available to the panel. (All participants, were thoroughly fluent in the English language; but in order to minimise any language difficulties, a bilingual local doctor familiar both with terminology and the patient's case details was present at each interview).

INDICES

A substantial number of indices form the basis of the analysis in this presentation (Table 1). Participants were furnished with identical data, including reprints of relevant articles, for hand calculation of the CDAI,[1] Harvey and Bradshaw's Index,[3] the Dutch AI,[6] (where because of its complexity a specific 'worked example' was included), and the OMGE Index.[4] Where indices were unpublished, a full description was furnished to participants on the basis of information supplied by the creators of each index.

STATISTICAL ANALYSIS

This problem is discussed in a separate footnote.

Results

PART 1 ANALYSIS OF 'CASE HISTORIES' BY PROSPECTIVE INDEX USERS

Ten 'case histories' were analysed independently by seven prospective index users. The resultant calculations of the CDAI are illustrated (Fig. 2 and Table 2). Considerable discrepancy is apparent; the range between the lowest and the highest estimate for each patient was often a matter of several hundred points.

Such discrepancy could be caused by one or two individual observers unfamiliar with the index. Table 2 therefore illustrates the 'scatter' between the 'middle 5' observers – excluding (for each case history) the lowest and highest estimates. Such an analysis clearly biases the study in favour of the index concerned. Nevertheless, the residual variation is still very high.

INDIVIDUAL DATA ITEMS

Because calculation of the CDAI involves calculating several individual items, a further attempt was made to determine which items were responsible for the interobserver variation (Table 3).

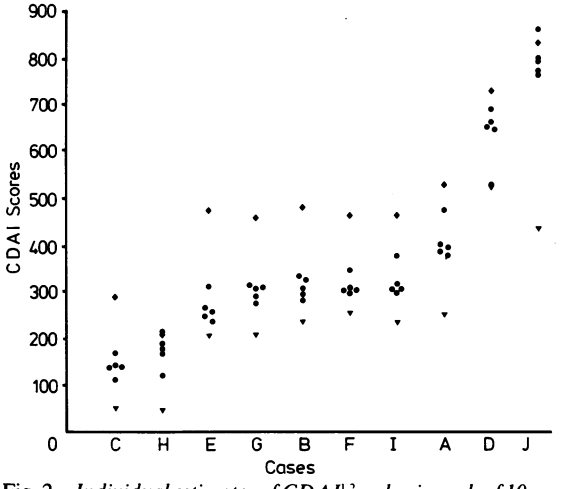

Fig. 2 *Individual estimates of CDAI[1,2], value in each of 10 case histories by seven prospective users of CDAI in Part 1 of study. Notice (a) considerable range of estimation in each instance and (b) two observers appear to have completely misunderstood the CDAI (as compared with the other five observers).*

The items given rise to most variation were number of liquid stools per day (due to dissention as to what constitutes a liquid stool), complications (some observers felt the given list to be incomplete, and added their own complications), the haematocrit ('rules of thumb' for calculating this differ from country to country) and the weight deficit (the method of calculating ideal weight varies from country to country).

The biggest single problem, however, was inherent in the CDAI itself. For in the CDAI scheme indi-

Table 2 *Estimates of CDAI by seven observers for 10 case histories in Part 1 of study; showing for each patient median value of seven CDAI estimations, range of estimation and 'scatter' – that is, range of residual estimation after highest and lowest excluded*

Case history	*Median of 7 CDAI estimations*	*Range low–high*	*Scatter (excl. low/high)*
C	132	44–281	108–162
H	172	41–207	119–202
E	261	201–473	235–310
G	304	203–457	272–309
B	305	235–478	280–330
F	305	254–462	298–345
I	302	229–464	292–375
A	391	249–529	373–472
D	651	521–730	525–690
J	794	434–864	763–834

Table 3 *Source of observer variation in eliciting CDAI, showing variation between observers for separate items. Note all variations in observation multiplied by 'factor score'; consequent effect on CDAI considerably enhanced*

CDAI item	*Observed* median variation*	*CDAI factor score*	*Effect of observed variation on CDAI*
Stools	6	2	12
Pain	2	5	10
Wellbeing	2	7	14
Complications	1	20	20
Haematocrit	3	6	18
Weight	5	1	5

*Median variation between each pair of observers for each of 10 cases.

vidual estimations are multiplied by 'discriminant factors' – so that – for example, disagreement about the presence or absence of a single complication adds no less than ±20 to the final CDAI total.

MODIFIED CDAI INDICES

The same group of observers also calculated (for each of the 10 case histories) additional values for the CDAI as modified by Harvey and Bradshaw[3] and by the World Organisation of Gastroenterology Research Committee[4] (Fig. 3). The results at first sight show less interobserver variation, yet this is misleading, for the possible range of values is far smaller than the CDAI, and the variation is just as high (expressed as a percentage of the median level).

The effect of this variation is, in practice, considerable. If an arbitrary cutoff between 'active' and 'inactive' disease is set at a score of 8 (and a cutoff between 'moderate' and 'severe' disease at 16), around 10% of patients in any given series will be 'misclassified' (because of interobserver variation).

OTHER INDICES

Calculation of the Activity Index proposed from

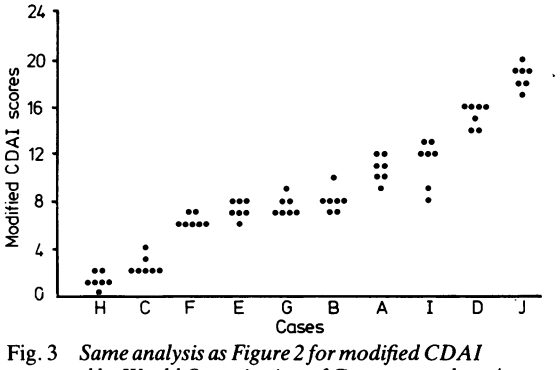

Fig. 3 *Same analysis as Figure 2 for modified CDAI proposed by World Organisation of Gastroenterology.*[4]

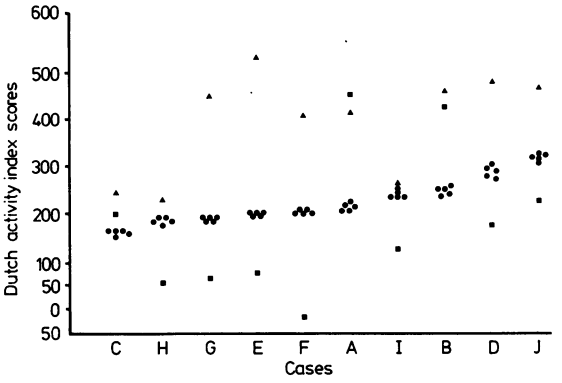

Fig. 4 *Same analysis as Figure 2, estimates being of Dutch Activity Index.*[6]

Holland[6] showed a totally different pattern of observer variation (Fig. 4). Most observers were able to calculate to within one or two points, the precise value of this Index. Nevertheless two out of seven subjects apparently failed to understand the method of calculation and produced consistent (but totally erroneous) estimations.

Other indices studied were the 'Best-Becktel-B' Index[5] and the Index proposed by the Cape Town group (personal communication). Observer variation for these two indices was also considerable.

The final 'Index' tested was that proposed by the Study Group which later formed the IOIBD. This merely consisted of a list of 10 factors thought to be important by members of the Study Group. As might be expected there was good observer agreement in recording this, 95% of all estimates lying within one of the consensus mean score for each patient.

PART 2 SINGLE PATIENT ASSESSMENT

It could be argued that the variation noted in Part 1

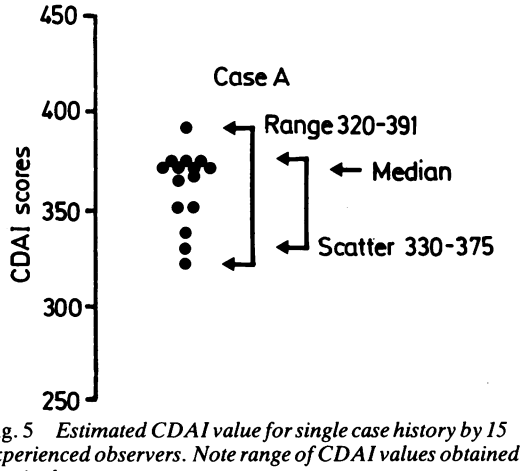

Fig. 5 *Estimated CDAI value for single case history by 15 experienced observers. Note range of CDAI values obtained for single case.*

Table 4 *Illustration of overall results in Part 2 of study (15 estimates of single patient data by experienced observers). Note considerable range of estimation in each of current indices*

Index	Part 1	Part 2		
	Median of 7 estimations	Median of 15 estimations	Range of estimations	Inter-quartile range*
CDAI	391	371	320–391	339–371
Harvey Bradshaw	11	10	6–12	9–11
OMGE	11	9	6–11	8–11
Dutch AI	208	208·6	208–417	208–260
S African	17	14	10–21	12–19
SAI	—†	289	115–524	199–300
Oxford	5	5	3–14	5–6

*9 estimations of 208·6 (correct value); †not estimated in Part 1.

represented inexperience on the part of the observers. In the second part of the study, a single 'case history' (Fig. 1) was presented to 15 members and colleagues attending the IOIBD Copenhagen meeting in July 1984 – resulting in a series of 15 independent estimations about a single case.

Figure 5 indicates the 15 individual estimations of the CDAI value. The range of estimations was 71 (320 to 391) and the scatter (ignoring the highest and lowest values) 45. As regards the source of this variation between observers, similar considerations apply to those already discussed.

Table 4 summarises the data from the 15 individual calculations for each of the other indices calculated (including the SAI used by the European Study Group.[7] Once again the same patterns were noted –

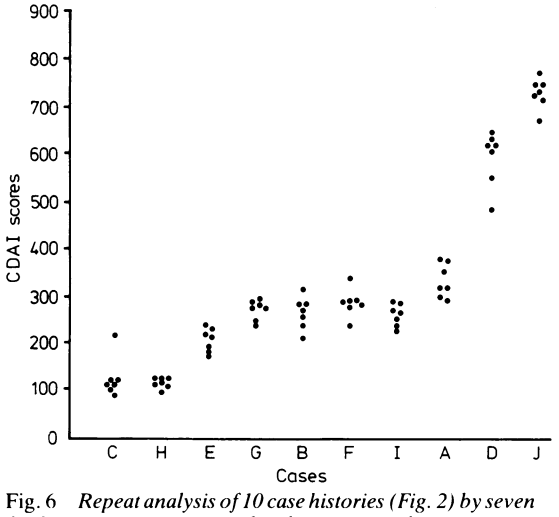

Fig. 6 *Repeat analysis of 10 case histories (Fig. 2) by seven further prospective users, after discussion and attempt to define terminology. Note considerable reduction in range of scatter of estimates.*

particularly in respect of the Dutch Activity Index, where nine of the 15 subjects calculated the index to be 208·6 (the figure arrived at by the instigators of the index themselves, who participated in the study), whereas other estimations ranged up to 417 (twice this value).

PART 3 FURTHER ANALYSIS BY PROSPECTIVE USERS AFTER DISCUSSION

As a result, considerable discussion about definition of terminology took place. Subsequently a further seven prospective users repeated the study described in Part 1. Figure 6 illustrates the CDAI calculations by seven observers participating in the repeat study. There was considerable reduction in interobserver variation; the scatter was considerably less in almost all patients than that observed in Part 1 (see Table 2).

The results as regards other indices of activity were very similar. In particular, this applied to the Dutch Activity Index (where six of the seven observers produced virtually identical scores) and the South African and 'Best-Becktel-B' indices (which still showed considerable interobserver variation).

The repeat study also compared three individuals who had participated in both Parts 1 and 3 (Fig. 7). These data also suggested that the effects of discussion and previous definitions of terminology may be quite considerable.

IMPORTANCE OF RANKING

Previous studies have explored the use of ranking methods of analysis;[1] and when this is done, the findings become more favourable. Table 5 shows the consensus ranking of patients (from least severe to most severe) and the individual ranking of each observer. The correlation values of these rankings are extremely high.

PART 4 ELICITATION OF PATIENT DATA

Subsequently, six experienced gastroenterologists interviewed and examined six patients –

Table 5 *Case histories in Part 3 ranked on the basis of CDAI scores*

Observer	Ranking of 10 case histories										Rho*
	Lowest									Highest	
1	C	H	E	I	B	G	F	A	D	J	1·00
2	C	H	E	I	B	F	G	A	D	J	0·998
3	H	C	E	B	I	G	A	F	D	J	0·995
4	H	C	E	B	F	G	I	A	D	J	0·992
5	C	H	E	I	G	F	B	A	D	J	0·987
6	C	H	E	B	F	G	I	A	D	J	0·992
7	C	H	E	B	I	G	F	A	D	J	1·00
Consensus	C	H	E	B	I	G	F	A	D	J	

*Spearman Rho rank correlation coefficient, comparing each observers individual ranking with consensus ranking.

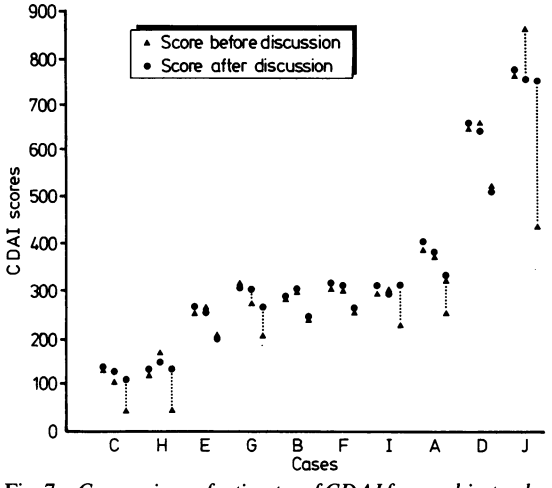

Fig. 7 *Comparison of estimates of CDAI from subjects who participated twice in present study. Note less interobserver variation in Part 3 (values ringed) than in Part 1 estimations.*

independently – on the same day (Fig. 8). There was wide variation in individual assessments of the CDAI – particularly in respect of patient 5 (where estimates ranged from 50 to 500). In only two patients was there 'good agreement', and even here, there was in each case a variation of over 50 points between the highest and lowest estimations. As regards the Dutch AI, far closer estimates were obtained, marred only by occasional wide discrepancies by individual single observers. Other indices calculated showed similar variation to the CDAI.

Discussion

No evaluation can ever determine that an index is 'valid'; only that an index may be useful in clinical work and/or therapeutic trials. If an index cannot be reproducibly calculated by those for whom it is designed then its subsequent use in clinical or research work will be less than optimal.

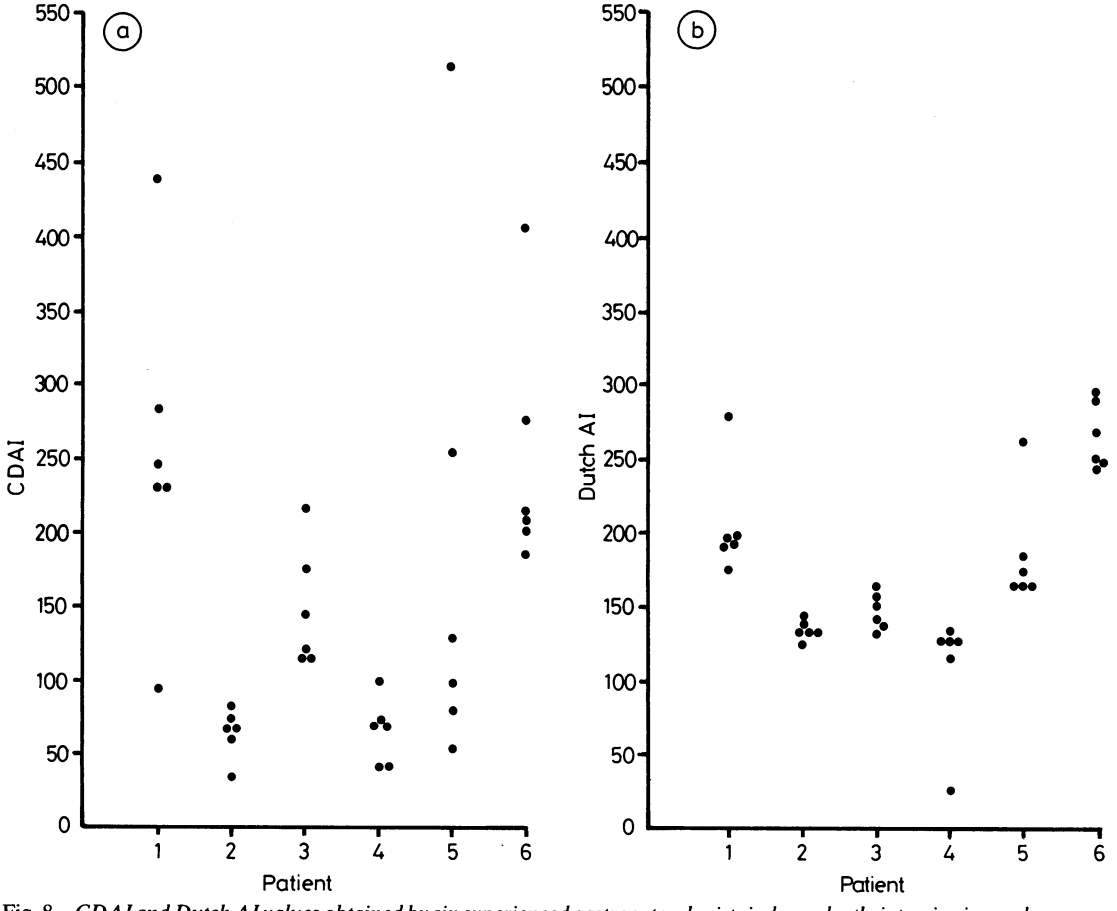

Fig. 8 *CDAI and Dutch AI values obtained by six experienced gastroenterologists independently interviewing and examining six individual patients.*

As with any other evaluation, it is possible to level criticisms at the present study. Thus, it is possible to argue that the observers were unrepresentative, and lacked competence or lacked adequate guidance. Yet the observers were carefully chosen, being either typical prospective users of indices, or those with considerable experience and interest in the field, and all were provided with the authors' original descriptions of each index. All appeared to understand these.

It is also possible to argue that the statistical analysis of the resultant data is imprecise (see Appendix 2). But no statistical consideration can disguise the very considerable degree of observer variation noted in the present trial in calculating the various indices from identical information bases. This ranges from practical matters of detail (such as what constitutes a 'complication', how one calculates haematocrit levels, and how one works out ideal weights), to complete failure (in some instances) to understand the index in question by the subject under study.

In the present study a considerable effort was made to involve those who would be actual users of the various indices in clinical trials. The participants (clinical consultant surgeons and physicians and their research assistants) are exactly the people who should be able to score systems reproducibly if the system is to be of practical use. The fact that manifestly they cannot do so must therefore cast considerable doubt on the validity of results obtained from existing studies which have used the various indices concerned.

On a more constructive note, the improvement in Part 3 indicates that much can be done by careful discussion; and many of the causes of variation are open to simple remedial methods. A second constructive suggestion which emerges from this study concerns the use of ranking methods. It is possible to invest numbers with an importance quite beyond what is appropriate; and Table 5 indicates that (after detailed discussion) consultants and research assistance on both sides of the Atlantic can rank a test series of patients in almost identical order using the CDAI (or other indices).

It thus appears from the present study that, whilst all of the indices so far proposed are open to considerable observer variation in calculating them, clarification of the indices and the use of ranking methods can do much to overcome this deficiency – and thereby provide in future a better basis for the scientific evaluation of Crohn's disease and its treatment.

Appendix 1

MEMBERSHIP OF IOIBD

S C Truelove*, Chairman, UK, R G Farmer, Vice-Chairman, USA, F T de Dombal*, Scientific Secretary, UK, S Baker* (Canada), V Binder*, P Riis (Denmark), C Andre*, L Descos*, A P Heckets-Weiler*, R Modigliani* (France), H Goebell, H Malchow (W Germany), Ph van Elteren*, P A M van Hees, S A S Pena*, J H M van Tongeren, I T Weterman* (Netherlands), I N Marks, J P Wright (South Africa), G Hellers, L Hulten, G Jarnerot (Sweden), R Allan, H J F Hodgson*, D Jewell*, E G C Lee†, G Watkinson* (UK), W R Best, B I Korelitz, A I Mendeloff, G B Rankin*, D B Sachar, J Singleton (USA).

*Members taking part in one or other of the studies described.

†Mr Emanoel Lee sadly died between taking part in these studies and their presentation in this paper.

Appendix 2

STATISTICS

Statistical analysis of observer variation in clinical medicine has been handicapped over the years by poor agreement amongst statisticians as to the methodology of carrying out this form of assessment. There is unfortunately still no 'gold standard' by which observer variation in clinical medicine can be measured, and the difficulties are compounded in the present instance because the 'indices' estimated here represent numerical information about what is manifestly non-numerical data. For these reasons, some brief statistical comments may be appropriate. The most widely used coefficient of agreement in clinical studies is the Kappa statistic of Cohen.[9] In fact, Cohen suggested not one but several versions of the Kappa statistic; Goodman and Kruskal[10] produced an identical statistic (which they called lambda), whilst Guttman[11] had also suggested a 'comparable' (lambda) statistic (which was quite differently defined, as was the 'comparable' (pi) statistic of Scott).[12] In this context both Saiger[13] and Schleff[14] have pointed out that it is not the mere presence of observer variation, but its magnitude and its effect in practical terms, which are the critical features to be measured. This view is supported by many authorities, such as Norbert Weiner.[15]

For these reasons (as will be apparent from the Tables and Figures) we have chosen (in lieu of complex and potentially fallible statistical analysis) to set out our results rather fully and to use simple non-parametric analyses wherever possible (such as median and range of estimations made). Also, as observer variation in this study is due to (a) poor calculation and (b) complete failure to understand the indices concerned, there are two types of discrepancy, slight random variation between estimates, and consistently large errors (Figs 2 and 4). In order to take this into account, we have (in addition to the range of estimations) measured (for each assessment) the 'scatter' of estimation – that is once the highest and lowest values for each estimation have been excluded. This 'scatter' of estimation (roughly corresponding to an interquartile range in Parts 1 and 3) probably represents a more realistic assessment of what informed observers should be able to record in clinical practice in the given circumstances.

Naturally, this type of assessment biases the analysis quite strongly in favour of each index. Yet the observer variation

is still relatively large, despite this bias – and thus this form of analysis in no way invalidates the major conclusions in this study.

The 'numerical' data from this study are not parametric data; and hence there is (as stated) a need for non-parametric methods of expression (such as median and range). This feature of the data, however, also underlines the importance of the use of ranking methods (such as those suggested by the NCCD authors themselves.[1]

Finally, the data from this study undoubtedly reveal a major deficiency in the use of discriminants for calculating indices which relate to situations in clinical medicine where there is likely to be significant observer variation. For example, Table 3, even where there is only slight observer variation this may lead to considerable discrepancy between the final CDAI estimates produced – because each observer variation is magnified by a factor of up to 20. This problem must be taken into account in the construction of any future indices.

The study was carried out under the auspices of the International Organisation for the Study of Inflammatory Bowel Disease, whose membership is listed separately. In addition, a number of colleagues participated in the study as prospective users, in one or other of the various parts of the trial. These were Mr W A F MacAdam, Dr A G Morgan, Mrs S E Clamp (UK), Drs J R Clarke (USA), S Ogren (Sweden), J Rasmusson and B Davidson (Denmark) and their help is warmly acknowledged. Finally, the International Organisation for the Study of Inflammatory Bowel Disease is grateful to Pharmacia AB, Uppsala, Sweden for financial support in respect of its research programme.

References

1 Best WR, Becktel JM, Singleton JW, Kern FJr. Development of a Crohn's disease activity index. National Co-operative Crohn's Disease Study. *Gastroenterology* 1976; **70:** 439–44.
2 Winship DH, Summers RW, Singleton JW, *et al.* Study design and conduct of the study. National Co-operative Crohn's Disease Study. *Gastroenterology* 1979; **77:** 829–42.
3 Harvey RF, Bradshaw JM. A simple index of Crohn's Disease Activity. *Lancet* 1980; **1:** 514.
4 Myren J, Bouchier IAD, Watkinson G, *et al.* The OMGE Multinational inflammatory bowel disease survey 1976–1982. A further report on 2657 cases. *Scand J Gastroenterol* 1982; **19:** suppl. 95: 1–27.
5 Best WR, Becktel JM. The Crohn's Disease activity index as a clinical instrument. In: Booth CC, Weterman I, Pena S, Haex CC, eds. *2nd International Workshop on Crohn's Disease.* Noordwijk/zee; Holland: de Baak, 1980.
6 van Hees PAM, van Elteren Ph, van Lier HJJ, van Tongeren JHM. An index of inflammatory activity in patients with Crohn's Disease. *Gut* 1980; **21:** 279–86.
7 Geobell H, Jeskinsky JH, Weinbeck M, Schomerus H. *European Co-operative Crohn's Disease Study (ECCDS): An index of severity and activity in Crohn's Disease (SAI).* In press.
8 Data from 1981 Oxford meeting of study group forming the IOIBD. (Cited by Myren, *et al.*[4]).
9 Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Measurement* 1960; **20:** 37–46.
10 Goodman LA, Kruskall WH. Measures of association for cross classifications. *J Am Statist Assoc* 1954; **99:** 732–64.
11 Guttman L. An outline of the statistical theory of prediction. In: Horst P, ed. *The prediction of personal adjustment.* New York: Social Science Research Council, 1941.
12 Scott WA. Reliability of content analysis. The case of nominal scale coding. *Public Opinion Quarterly* 1955; **19:** 321–5.
13 Saiger GL. Observations on the probability of error in medical diagnosis. *Am J Intern Med* 1982; **56:** 860–4.
14 Schleff TJ. Decision rules, types of error and their consequences in medical diagnosis. *Behav Sci* 1963; **8:** 97–107.
15 Weiner N. Nonlinear prediction and dynamics. *Proceedings 3rd Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley: University of California Press, 1956; **3:** 247.