

Observer variation and discriminatory value of biopsy features in inflammatory bowel disease

A Theodossi, D J Spiegelhalter, J Jass, J Firth, M Dixon, M Leader, D A Levison, R Lindley, I Filipe, A Price, N A Shepherd, S Thomas, H Thompson

Abstract

If skilled histopathologists disagree over the same biopsy specimen, at least one must have an incorrect interpretation. Thus, disagreement is associated with, although not the cause of, diagnostic error. The present study aimed to determine the magnitude of variation among 10 observers with a special interest in gastrointestinal histopathology. They independently interpreted the same biopsy specimens for morphological features which may discriminate between patients with Crohn's disease and ulcerative colitis and normal subjects. Thirty of 41 features had agreement measures significantly better than expected by chance ($p < 0.05$). The range of agreement in the 45 observer pairs over the final diagnosis was 65-76%. There was good agreement in discriminating between normal slides and those showing confirmed inflammatory bowel disease. For normal slides, however, the term non-specific inflammation was often applied and without any consistency. In addition, true Crohn's disease slides were often and consistently thought to be ulcerative colitis. Having identified 11 important discriminatory morphological features, two multiple regression analyses were then carried out to produce a scoring system for inflammatory bowel disease. These results suggest there is considerable room for improvement in the reliability of colonic biopsy specimen interpretation and that this could probably be achieved using more exact definitions of morphological features and diseases.

(Gut 1994; 35: 961-968)

Clinicians are using colonoscopic mucosal biopsy specimens more frequently to identify patients suspected of having inflammatory bowel disease, to determine disease extent and severity, and to differentiate ulcerative colitis from Crohn's disease.^{1 2} Histological detection and classification of non-infective chronic inflammatory bowel disease depend on accurate interpretation of carefully prepared biopsy material,³ but skilled histopathologists obtaining information from the same biopsy tissue may disagree about the findings.^{4 5}

Giard *et al*⁵ found poor reproducibility of morphological features and final diagnoses in normal subjects and those with inflammatory bowel disease. Cook and Dixon⁴ assessed observer variation in the examination of

macroscopic and biopsy material taken from patients with inflammatory bowel disease. They reported the percentage agreement, but did not take into account the proportion of observed agreement due to chance alone. Surawicz and Belic⁶ and Allison *et al*⁷ considered the value of features in discriminating acute self limited colitis from idiopathic inflammatory bowel disease, and reported both definitions and simple percentage agreement on selected features.

We believe that histopathology provides an important contribution to the diagnosis of inflammatory bowel disease. The aim of the present study therefore was to determine the reliability of the information obtained from colorectal mucosal biopsy specimens by assessing the magnitude of variation among 10 histopathologists with a special interest in gastrointestinal histopathology who independently interpreted the same 76 biopsy specimens. We set out to determine the observer variation associated with the morphological features and the final diagnoses. A scoring system was then developed for identifying patients with inflammatory bowel disease. In addition, at a video recorded meeting we evaluated the question of what leads to disagreement over important discriminatory features which were subsequently redefined.

Methods

Seven consultant histopathologists and three trainees, all with a special interest in gastrointestinal histopathology, were asked to review and independently code the same 76 biopsy specimens. Thirty four specimens came from patients with ulcerative colitis, 24 from patients with Crohn's disease, and 18 from normal subjects. Many of these apparently normal subjects were eventually classified as having the irritable bowel syndrome or other functional bowel disorder. In each instance the final diagnosis had been established by prolonged clinical follow up, radiological and morphological assessments, and, where appropriate, laparotomy or autopsy findings. Two slides taken from different levels of the biopsy material were provided for all the 76 patients. All of the slides were stained with haematoxylin and eosin. The biopsy specimens taken by sigmoidoscopy or colonoscopy were from patients who had been referred to the Westminster Hospital, London, for further evaluation of symptoms suggestive of inflammatory bowel disease.

Variation between the pathologists was

Departments of
Gastroenterology and
Pathology, Mayday
University Hospital,
Croydon
A Theodossi
S Thomas

MRC Biostatistics
Unit, Institute of
Public Health,
Cambridge
D J Spiegelhalter

Departments of
Pathology, St Marks
Hospital, London
J Jass
N A Shepherd

Westminster Hospital,
London
J Firth
M Leader
R Lindley

University of Leeds
M Dixon

St Bartholomew's
Hospital, London
D A Levison

UMDS, Guys Hospital,
London
I Filipe

Northwick Park
Hospital, Harrow
A Price

General Hospital,
Birmingham
H Thompson

Correspondence to:
Dr A Theodossi, Mayday
University Hospital, Mayday
Road, Thornton Heath,
Surrey CR4 7YE.

Accepted for publication
28 October 1993

examined by providing a form on which they recorded their coding of morphological features, as well as their histopathology diagnoses. The coding form comprised 39 different histological features. By design, no attempt was made to provide a strict definition for each of the features which were all ones in regular use. Each pathologist was asked to classify each feature into one of four possible grades: 'absent', 'indefinite', 'little', or 'marked'. Each of the four histological categories of ulcerative colitis, Crohn's disease, normal, and mild non-specific inflammation had to be graded as 'unlikely', 'possibly', or 'likely' to be present.

After the data had been analysed, the important findings were discussed at a meeting in the presence of all 10 histopathologists. At the beginning of the meeting, the histopathologists independently re-examined two of the biopsy specimens. The slides associated with the highest observer variation were then jointly reviewed in order to determine the source of disagreement. The meeting was recorded on a video tape, from which data were subsequently analysed and summarised (see results). In order to improve agreement on 10 important morphological features, three observers (JRJ, NAS, DAL) redefined the features using six slides, three showing high and three low agreements for each feature. At a second meeting, the 10 new definitions, together with the appropriate slides, were reviewed and criticised by the other authors, to improve the definitions further (see Appendix). At this meeting, it was decided that two groups of features, which had been individually graded on the form, would best be defined as composite features. Thus, the results contain a feature 'crypt architectural abnormality', which is calculated as the most severe grade recorded for any one of 'crypt branching', 'crypt distortion', and 'crypt atrophy'. A second composite feature, 'neutrophil infiltration', is the most severe grade given to 'neutrophilic infiltration

of crypt epithelium' and 'neutrophils in lamina propria'.

STATISTICAL METHODS

An overall proportion of agreement was calculated, and then corrected for chance agreement. This kappa coefficient⁸ is provided together with its level of statistical significance in relation to a null value of 0 – just chance agreement. We also reclassified each feature into 'present' ('little' or 'marked') or 'absent' ('absent' or 'indefinite'), and provided proportions of agreement on this scale. These calculations were repeated for the opinion concerning disease.

To determine the value of each feature in discriminating between normal subjects and those with inflammatory bowel disease, and between Crohn's and ulcerative colitis, we began by calculating the frequency of each feature graded 'little' or 'marked' for colitis, Crohn's, and normal subjects. We then calculated the rank correlations between the observers' reported grades of features and the true presence or absence of inflammatory bowel disease. We then repeated the analysis using only the slides of Crohn's and ulcerative colitis tissue and calculated rank correlations for the true presence/absence of ulcerative colitis and Crohn's.

Eleven important discriminatory features were identified, based on their respective frequencies in the three classes of subjects and the extent of agreement between the observers on the reporting of the features. These were then reduced to 10 after carrying out two multiple regression analyses which provided scores firstly, for differentiating 'normals' from patients with inflammatory bowel disease and secondly, for distinguishing Crohn's from ulcerative colitis in patients with confirmed inflammatory bowel disease.

Results

Figure 1 shows that there was considerable disagreement in providing a diagnosis, particularly in the assessment of the likelihood of 'Crohn's disease and 'non-specific inflammation'. True ulcerative colitis is well diagnosed, with 'ulcerative colitis' considered 'possible' or 'likely' in 93% of cases, although 'Crohn's colitis' was also considered at least 'possible' in 63%.

However, for observations on slides that were truly Crohn's disease, only 68% thought 'Crohn's' at least 'possible', whereas 'ulcerative colitis' was thought at least 'possible' in 69%. In 41% of times when an observer saw a true Crohn's slide, 'ulcerative colitis' was thought 'likely'. Thus, Crohn's disease was frequently and consistently mistaken for 'ulcerative colitis', although we note that true Crohn's or ulcerative colitis slides were only very rarely thought 'likely' to be 'normal' or 'non-specific inflammation'.

Similarly, true normal slides were rarely thought 'likely' to be 'ulcerative colitis'; and never thought 'likely' to be 'Crohn's disease'.

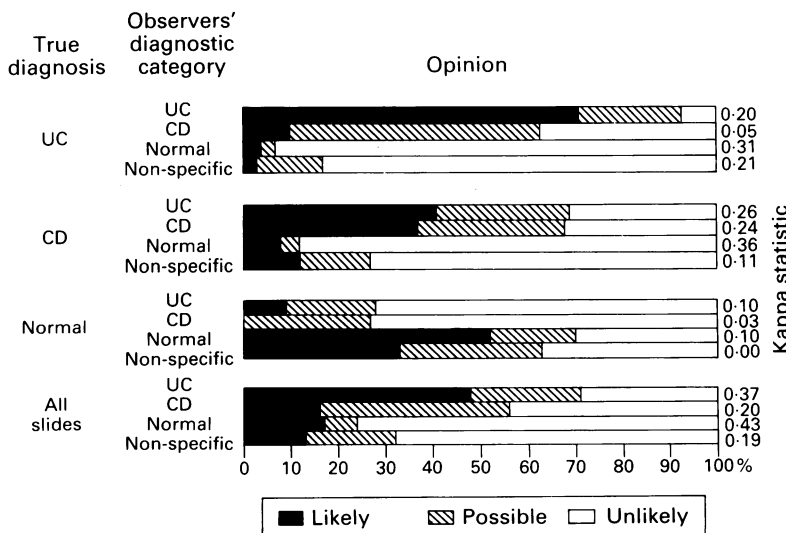


Figure 1: Observer's diagnosis in relation to true diagnosis in biopsy material from patients with Crohn's disease (CD) and ulcerative colitis (UC) and normal subjects.

TABLE I Incidence and interobserver variation on features, ranked by overall kappa statistic and correlation with presence of inflammatory bowel disease v normal, and ulcerative colitis (UC) v Crohn's disease (CD)

Feature	Overall agreement (%)	Kappa*	Significance p value	% Grading reported 'little' or 'marked' when disease is			Rank correlation between grading of feature and	
				UC	CD	Normal	True presence of IBD	True UC
1 Crypt neutrophilia	69 (82)†	0.47 (0.62)‡	<0.001	53	56	2	0.46§	-0.06
2 Neutrophilic (1 or 5) infiltration	60 (82)	0.41 (0.63)	<0.001	66	67	6	0.53§	-0.04
3 Epithelioid granulomas	88 (95)	0.41 (0.60)	<0.001	0	21	0	0.18	-0.41§
4 Langhan's type giant cells	94 (97)	0.40 (0.59)	<0.001	1	11	0	0.13	-0.28§
5 Neutrophils in lamina propria	60 (79)	0.40 (0.57)	<0.001	61	61	6	0.51§	-0.02
6 Crypt distortion	57 (80)	0.38 (0.60)	<0.001	78	51	4	0.53§	0.34§
7 Crypt abscesses distensive	79 (84)	0.37 (0.46)	<0.001	18	32	1	0.27	-0.17
8 Ulceration of surface epithelium	71 (85)	0.36 (0.56)	<0.001	31	24	0	0.33	0.07
9 Crypt architecture (6 or 16 or 21)	54 (79)	0.36 (0.56)	<0.001	86	65	11	0.57§	0.35§
10 Mucin depletion	54 (79)	0.34 (0.57)	<0.001	69	57	17	0.41§	0.15
11 Lymphoid follicles in lamina propria	63 (74)	0.33 (0.44)	<0.001	36	40	32	0.09	-0.04
12 Submucosal inflammation	79 (86)	0.33 (0.44)	<0.001	18	20	1	0.21	-0.01
13 Eosinophils	53 (77)	0.32 (0.51)	<0.001	87	62	16	0.54§	0.31§
14 Crypt abscesses disruptive	81 (86)	0.31 (0.42)	<0.001	14	26	0	0.24	-0.16
15 Submucosal lymphoid aggregates	80 (84)	0.31 (0.39)	<0.001	15	17	11	0.06	-0.05
16 Crypt branching	55 (74)	0.30 (0.47)	<0.001	63	39	5	0.44§	0.30§
17 Paneth cell metaplasia	84 (87)	0.29 (0.36)	<0.001	19	4	6	0.09	0.23§
18 Diffuse inflammation	50 (74)	0.27 (0.34)	<0.001	84	77	41	0.41§	0.05
19 Plasma cells	52 (77)	0.27 (0.25)	<0.001	87	89	57	0.38	0.07
20 Villous mucosal configuration	70 (81)	0.26 (0.37)	<0.001	30	12	6	0.23	0.29§
21 Crypt atrophy	57 (71)	0.24 (0.27)	<0.001	44	23	5	0.34	0.28§
22 Crypt dilatation	55 (67)	0.23 (0.31)	<0.001	57	42	8	0.39§	0.17
23 Inflammatory polyps	90 (96)	0.22 (0.24)	<0.05	3	4	0	0.15	0.01
24 Submucosal fibrosis	93 (95)	0.20 (0.26)	<0.11	6	3	0	0.13	0.04
25 Thickened muscularia mucosa	58 (79)	0.17 (0.30)	<0.001	25	20	6	0.17	0.08
26 Regenerative changes	48 (67)	0.17 (0.30)	0.01	51	44	11	0.32	0.09
27 Crypt hyperplasia	52 (68)	0.15 (0.24)	<0.001	41	29	11	0.26	0.15
28 Oedema of lamina propria	50 (61)	0.15 (0.18)	<0.001	44	41	30	0.11	0.03
29 Neuronal hyperplasia	96 (97)	0.14 (0.22)	0.27	4	4	0	0.07	0.07
30 Submucosal oedema	85 (90)	0.10 (0.13)	0.19	5	9	5	0	-0.11
31 Superficial aphthoid ulcer	92 (94)	0.10 (0.10)	0.27	5	4	0	0.13	0.05
32 Vascularity	49 (59)	0.10 (0.14)	<0.001	46	43	17	0.24	0.02
33 Lymphocytes	42 (67)	0.08 (0.15)	<0.001	79	85	49	0.36	-0.04
34 Fissuring	97 (99)	0.08 (0.02)	0.39	1	1	0	0.07	0.02
35 Foreign body type giant cells	94 (96)	0.05 (0.02)	0.39	3	3	0	0.07	-0.08
36 Focal inflammation	67 (78)	0.05 (0.09)	0.20	10	26	7	0.06	-0.18§
37 Histocytes	34 (54)	0.05 (0.07)	<0.001	45	55	27	0.25	-0.08
38 Fat in lamina propria	97 (98)	0.01 (0)	0.50	2	0	0	0.07	0.06
39 Submucosal lymphangiectasia	96 (98)	0 (-0.02)	0.50	1	2	0	0.08	-0.02
40 Vasculitis	99 (99)	-0.01 (-0.01)	0.50	0.3	0.4	0	0.04	-0.04
41 Muciphages in lamina propria	82 (90)	-0.02 (-0.02)	0.50	5	3	7	0	0.02

*Kappa=1-(100-overall agreement %)/(100-expected agreement by chance %).

†‡The overall agreement and kappa values in brackets are the results obtained when the gradings were reclassified into a present-absent scale.

§Top 10 features.

They were frequently thought 'possible' or 'likely' 'non-specific inflammation', although there was no agreement on this, and the grade was as if given at random.

Table I displays 41 features ranked in order of the observer agreement as measured by a kappa statistic. The overall proportion of each grading is based on $10 \times 76 = 760$ observations. The figures for overall agreement vary with the rarity of the feature: 'fat in lamina propria' (#38) has overall agreement of 97% but this is only because it was so often recorded as absent - on the rare occasions it was seen by an individual, it was never recorded by one of his colleagues. The kappa statistic corrects for this chance agreement, although for more frequent findings the standard error of the kappa statistic decreases and hence increases the statistical significance of the improvement over agreement by chance alone. This can be seen by comparing features 'superficial aphthoid ulcer' (#31) and 'vascularity' (#32): both have poor agreement (kappa=0.10), but the latter feature is more common and hence its kappa value is significantly different from zero. Separate 'pairwise' analysis comparing each observer with every other showed no individual standing out as being distinct in their opinions, with overall agreement varying between 65% and 75%.

Features which are rarely observed, usually (but not inevitably) have low kappa values. For example, muciphages (#41) were present in 6% of the observations leading to a kappa of -0.02. Epithelioid granulomas (#3) were also rarely observed (in 7% of observations) but the associated kappa value was considerably higher at 0.41, indicating that when seen there was much closer agreement on their presence. By contrast, common features such as lymphocytes (#33) may have poor agreements.

When the features are reclassified into just 'present' or 'absent' (table I - note †‡), it is inevitable that agreement will improve. However, it can be seen there is little change in overall ranking. Thus, a feature such as 'lymphocytes (#33) does not have poor agreement simply because of minor disputes over whether the abnormality should be called 'little' or 'marked'; there is serious disagreement with one third of paired comparisons of pathologists giving conflicting opinion over the presence or absence of 'lymphocytes'.

As a formal method of investigating the relationship between features and the presence of inflammatory bowel disease, we show the rank correlation coefficient (Spearman's rho) between the grade given to the feature and the true presence/absence of disease. The 10 features with the highest correlation with true

TABLE II Details of two true-Crohn's disease (CD) slides in which substantial dispute existed concerning the likelihood of 'CD'

	Grades	Unlikely	Possible	Likely
(A) Slide no 11 (CD). Eight histopathologists' diagnostic opinion concerning 'CD':				
Opinion concerning 'mucin depletion' (#10)	Absent Indefinite Little Marked		1 1 2	2 2
(Spearman rank correlation=0.76, p<0.05)				
(B) Slide no 56 (true CD). Nine histopathologists' diagnostic opinion concerning 'CD':				
Opinion concerning 'Epithelioid granulomas' (#3)	Absent Indefinite Little Marked	2	1 1 1	1 2 1
(Spearman rank correlation=-0.76, p<0.05)				

Entries in the tables show the number of observers grading each combination of opinion about 'Crohn's' and a selected feature for each slide, eg, two observers thought 'mucin depletion' was 'marked', and also stated an opinion that Crohn's disease was 'unlikely' to be the true diagnosis. The observed relationship suggests that improvement in agreement on these and other features may improve diagnostic agreement.

inflammatory bowel disease are indicated by asterisks. There is a strong relationship between good agreement and discriminatory power - no feature ranking below 23 in agreement is in the top 10 for discrimination. The most useful features for distinguishing inflammatory bowel disease from normal were thus neutrophils (#1, #5, and their composite #2), crypt architectural abnormality (#6, #16, and the composite measure #9), mucin depletion (#10), eosinophils (#13), diffuse pattern of inflammation (#18), and crypt dilatation (#22).

We now consider features that are useful for distinguishing Crohn's from ulcerative colitis, and hence the final column of Table I is based solely on the 58 inflammatory bowel disease slides. The rank correlations for features and true ulcerative colitis are calculated and the 10 most powerful features identified: epithelioid granulomas (#3), Langhan's type giant cells (#4), crypt architecture (#6, #16, #21, and #9), eosinophils (#13), Paneth cell metaplasia (#17), villous mucosal configuration (#20), and focal inflammation (#36). Again better agreement is strongly related to discriminatory value, although the importance of 'focal inflammation' persists in spite of poor agreement. This suggests that a firmer definition may lead to improved diagnostic power.

To illustrate further the apparent relationship between disagreements on specific features and disagreements between claimed

TABLE IV Suggested strategy showing the conclusion to be drawn for any combination of the inflammatory bowel disease (IBD)/normal and ulcerative colitis (UC)/Crohn's disease (CD) scores

	UC/CD score			
	≤-3	-2 to 2	'Normal'	≥3
IBD/normal score : ≤1	-	-	-	-
: 2 to 4	'CD'	'Possible IBD'	'UC'	'UC'
: ≥5	'CD'	'IBD'	'UC'	'UC'

Note: the top right combination cannot be obtained from the scores given in Table III. The top left combination can only be obtained if 'epithelioid granulomas' is the only abnormal feature recorded among the seven contributing to the IBD/normal score. This has not been observed in our series.

diagnoses, we consider two slides (#11 and #56) which were actually of Crohn's disease but which gave rise to substantial dispute. Table II shows the relationship between opinion concerning 'Crohn's disease' and the opinion given to two selected features. The clear relationship, while not proving that viewing a particular feature causes the claimed diagnosis to be made, strongly supports the idea that reduced disagreement on specific influential features could give rise to better agreement on claimed diagnosis, with accompanying increase in accuracy.

The scoring systems based on all 76 biopsy specimens and derived from the discriminant analyses are shown in Table III. Thus, a biopsy in which an observer reported diffuse inflammation, eosinophils, crypt architectural abnormalities, and Paneth cell metaplasia would receive an inflammatory bowel disease/normal score of 6 and a ulcerative colitis/Crohn's disease score of 4. In Table IV, the two scoring systems are seen together forming a 'strategy' in which the inflammatory bowel disease/normal and ulcerative colitis/Crohn's scores classify each slide into one of the cells.

Table V shows the consequences had that strategy been used by the observers in reviewing the slides. We can now compare the accuracy of the observers' original diagnosis shown in Figure 1 with that of the scoring system strategy. For normal slides the scoring system makes more positive diagnosis of 'normal' (66% v 52%) for a similar number of confident, but wrong, diagnoses of inflammatory bowel diseases (9% v 7%). For Crohn's disease, the scoring method makes fewer correct positive diagnoses (28% v 37%), makes the same number of false negative diagnoses of 'normal' (10% v 8%), but makes substantially fewer false confident diagnoses of 'ulcerative colitis' (11% v 41%), tending to be more cautious and classifying these slides as indeterminate 'inflammatory bowel disease'. For ulcerative colitis slides, the scoring system is definitely more diffident in asserting 'ulcerative colitis' (37% v 71%), makes similar false negative 'normal' diagnoses (3% v 4%), but makes substantially fewer false diagnoses of 'Crohn's' (6% v 10%).

Altogether 363 (52%) of slides would have been given a confident diagnoses, with 81% accuracy, while a further 213 (31%) would have been classified as indeterminate inflammatory bowel disease, with 99% accuracy.

TABLE III Ten selected important features and their contribution to indices for discriminating normal from inflammatory bowel disease (IBD) slides, and ulcerative colitis (UC) from Crohn's disease (CD). (Ulceration of the surface epithelium (#8) was included in the discriminant analyses but its rounded coefficient was zero in both indices.) Numbering of features corresponds to that in Table I

Feature	IBD/normal score	UC/CD
#2 Neutrophilic infiltration	1	-1
#3 Epithelioid granulomas	1	-6
#9 Crypt architectural abnormalities	3	1
#10 Mucin depletion		1
#13 Eosinophils	2	3
#17 Paneth cell metaplasia		2
#18 Diffuse inflammation	1	-2
#20 Villous mucosal configuration		1
#22 Crypt dilatation	1	
#36 Focal inflammation	3	-2
Range of score	0 to 12 (high favours IBD)	-11 to 8 (negative favours CD, positive favours UC)

TABLE V Conclusions that would have been drawn had the observers used the strategy in Table IV on 689 readings of slides, divided by disease category

Conclusion	True disease (%)		
	Normal	CD	UC
'Normal'	108 (66)	22 (10)	10 (3)
'CD'	4 (2)	59 (28)	2 (6)
'UC'	9 (6)	23 (11)	126 (37)
'IBD'	1 (1)	74 (35)	138 (44)
'Unsure'	41 (25)	32 (15)	40 (13)
	163 (100)	210 (100)	316 (100)

CD=Crohn's disease; UC=ulcerative colitis.

PANEL MEETING (INCLUDING REVIEW OF SLIDES SHOWING MARKED DISAGREEMENT)

Two slides were selected for re-reading before the meeting. Table VI shows the considerable disagreement that was observed. Different histopathologists use the term 'non-specific inflammation' in different ways. Some pathologists apply 'non-specific inflammation' to biopsy specimens where they are confident that inflammatory bowel disease or another condition such as irradiation colitis is present, but they cannot discriminate between these conditions. Others restrict the term for biopsies where they cannot discriminate between Crohn's and ulcerative colitis. Still others like to use the term when they believe Crohn's is likely to be present, but where the specimen shows insufficient evidence to make a definite diagnosis of Crohn's. Some histopathologists use the term to indicate that mild non-specific inflammation is present but there is no histological evidence of inflammatory bowel disease. The results from the present study indicated, however, that there was very little agreement on 'non-specific inflammation'.

Focal inflammation was the only feature with a low agreement measure that the pathologists considered to be important. Other features with low kappa values were thought to be rare and unusual and therefore unimportant. The terms 'absent', 'indefinite', 'little', and 'marked' were criticised. For example, how many granulomas should be present on the slide before changing one's opinion from 'little' to 'marked'? And if there were only micro-granulomas present, should they be described as 'little' or 'marked'. Different pathologists had different definitions for focal inflammation. One observer would describe an isolated crypt abscess as focal inflammation.

Others used the term when they saw patchy chronic inflammation on the slide. But how many of these 'patches' must one see before changing one's opinion from 'little' to 'marked'? A patch of chronic inflammatory cells that consisted only if lymphocytes could be the edge or part of a lymphoid follicle. Thus, some observers would call such a patch focal inflammation only if they could also see polymorphs amongst the lymphocytes. Similarly, patchy areas of inflammation which are regularly spaced on the slide should raise one's suspicions that these might all be arising from lymphoid follicles.

One observer commented that a well known feature such as mucin depletion should be easy to detect by everybody. Yet slides were shown where the observers were divided on the presence or absence of mucin depletion. One of the pathologists therefore concluded that there was probably general agreement on mucin depletion but differences in opinion occurred because of the different ways in which different histopathologists interpreted any one feature in the presence of all the others. On reviewing the slide where on the first reading three observers had seen and four had not seen granulomas, and where on the second reading no one saw granulomas, it took a long time to find a granuloma. Once one was seen, all 10 pathologists agreed that granulomas, or rather micro-granulomas, were present. The relevance of micro-granulomas in the diagnosis of Crohn's disease has been emphasised by Rotterdam *et al.*¹⁰ This slide highlighted the problem of observation of features. All 10 pathologists concluded that important features can be missed because of the problem of observation even among experienced observers. In addition, this video recorded meeting showed clearly that the observers may have different working definitions for morphological features considered to be well known and in routine use.

Discussion

This study, conducted by experienced histopathologists on colorectal biopsies from patients with inflammatory bowel disease and patients with functional diarrhoea, has shown serious disagreements both in the interpretation of morphological features and the final histopathological diagnoses. Experienced

TABLE VI Opinions of 10 observers reviewing two slides, before a video taped meeting; all observations taken on same two microscopes within a one hour period

	Slide #11 (true CD)				Slide #26 (true normal)			
	Absent	Indefinite	Little	Marked	Absent	Indefinite	Little	Marked
21 Crypt atrophy	6	1	2	1	9		1	
36 Focal inflammation	8		2		3		3	1
3 Epithelioid granulomas	10				9	1		
4 Langhan's type giant cells	9		1		10			
10 Mucin depletion	1		7	2	10			
8 Ulceration of surface epithelium	3	4	2	1	10			
	Unlikely	Possible	Likely		Unlikely	Possible	Likely	
UC	3	2	5		9		1	
CD	3	6	1		7	2	1	
Normal	10				5	4	1	
Non-specific inflammation	10				2	4	4	

histopathologists can accurately discriminate between biopsy specimens from normal subjects and those from inflammatory bowel disease, but the differentiation of Crohn's from ulcerative colitis is less reliable. In addition, patients with functional diarrhoea (our 'normal' group) were often found to show features of non-specific inflammation. Giard *et al*⁵ also found better agreement on ulcerative colitis than Crohn's disease and that agreement was poor for normal mucosa. Myren¹¹ noted that pathologists had difficulty distinguishing ulcerative colitis from non-specific inflammation and normal biopsy tissues from those with non-specific inflammation. However, 'ordinary non-specific acute and chronic inflammation is the major form taken by colitis'.¹² Although the term cannot therefore be discarded, it should be used cautiously to avoid mistaking normal individuals for those with inflammatory bowel disease.

Over the past two decades several papers¹³⁻¹⁵ have addressed the problem of differentiating Crohn's from ulcerative colitis, and idiopathic inflammatory bowel disease from acute self limited colitis.¹⁶ However, both Crohn's and ulcerative colitis typically have a long and variable course due to unpredictable remissions and relapses.¹⁷ In addition, many of the diagnostic features occur with similar frequency in the two conditions, and neither disease has a pathognomonic finding, 'present in every case of the one and absent in every case of the other'.¹⁸ Thus, sometimes only colitis indeterminate¹⁹ can be established.

Previously, the accuracy of biopsy diagnosis of ulcerative colitis and Crohn's disease was found to be 70% and 40% respectively in 126 patients.¹⁷ In contrast to previous reports,^{12 13} the present study suggests that mucin depletion does not distinguish Crohn's from ulcerative colitis. Its main value is to distinguish normal biopsy tissues from specimens showing inflammatory bowel disease. Crypt abscesses have been reported either with similar frequency in ulcerative colitis and Crohn's,^{12 13 20} or, more often, in ulcerative colitis.³ Interestingly, we found crypt abscesses more often in Crohn's disease.^{4 12} Colonoscopic biopsy, however, only rarely yields submucosal tissue, and, surprisingly, we found submucosal fibrosis in 6% of ulcerative colitis and in only 3% of Crohn's disease tissues.

Although, in our material, submucosal oedema occurred more often in Crohn's colitis, the difference was not significant (5% *v* 9%) and as it is associated with a low agreement measure, we do not recommend using it in distinguishing Crohn's from ulcerative colitis. It is widely accepted that Paneth cell metaplasia does not discriminate between Crohn's disease and ulcerative colitis.^{3 20-22} Our study, however, suggests this is an important marker for ulcerative colitis. We and others²⁰ have not confirmed either that a diffuse pattern of inflammation discriminates between Crohn's and ulcerative colitis,³ or that pseudopolyps are found more often in ulcerative colitis,¹² but we can agree with Price and Talbot³ that a villous mucosal configura-

tion occurs more often with ulcerative colitis.

Since crypt distortion (4%), crypt branching (5%), crypt atrophy (5%), and crypt architectural changes (11%) were seen in our normal group, we agree that minor degrees of fibrosis and crypt distortion probably do occur in the healing phase of acute self limited colitis.¹² It is likely therefore that minor inflammatory changes, together with minor crypt changes, are found in patients who after prolonged clinical follow up do not have inflammatory bowel disease. Clearly, patients with these minor morphological changes will need careful follow up with repeat biopsies.

The lamina propria in a normal colonoscopic biopsy specimen contains the occasional eosinophil.^{3 20 21 23} Others have suggested that mucosal eosinophilia is a marker of the severity^{3 12 23} or chronicity^{2 4} of ulcerative colitis, but we have found that mucosal eosinophilia is useful in that it distinguishes normal subjects from those with inflammatory bowel disease and its presence favours the diagnosis of ulcerative colitis.²⁵

Lymphoid aggregates have been found to be good discriminators in idiopathic inflammatory bowel disease.^{6 7 19} On average one lympho-glandular complex occurs approximately every 2 cm in normal subjects, and one every 0.7 cm in patients with colitis.²⁶ We found lymphoid follicles in similar frequency in Crohn's colitis, ulcerative colitis, and normal subjects. Clearly the value of lymphoid tissue is controversial and we recommend lymphoid aggregates should not be used as a major discriminator for idiopathic inflammatory bowel disease.

Features such as granulomas and focal inflammation can be regarded as helpful positive discriminators of Crohn's disease. Our study suggests the finding of chronic inflammation in the absence of a villous mucosal configuration, crypt branching, Paneth cell metaplasia, and crypt atrophy are strong pointers to possible Crohn's disease, and thus the latter features can be described as important negative features which tend to favour Crohn's.

The combination of 'little' and 'marked' as 'present' may lead to loss of some information and thus, for example, may prevent a histopathologist from using a 'marked' thickening of the muscularis mucosae in support of a diagnosis of ulcerative colitis. But there is no gold standard whether the feature is 'little' or 'marked' in a particular biopsy specimen. The present study has shown poor agreement on this feature, although when 'little' and 'marked' were lumped together, agreement improved (kappa value from 0.17 to 0.3). Better agreement may lead to less error, since if two observers agree on the presence or absence of a finding they are both either correct or incorrect. By contrast, if they disagree then at least one of them is incorrect. Thus, only agreement may be associated with full accuracy.

The classification of inflammatory bowel disease should be based on all available evidence including clinical, radiological, and

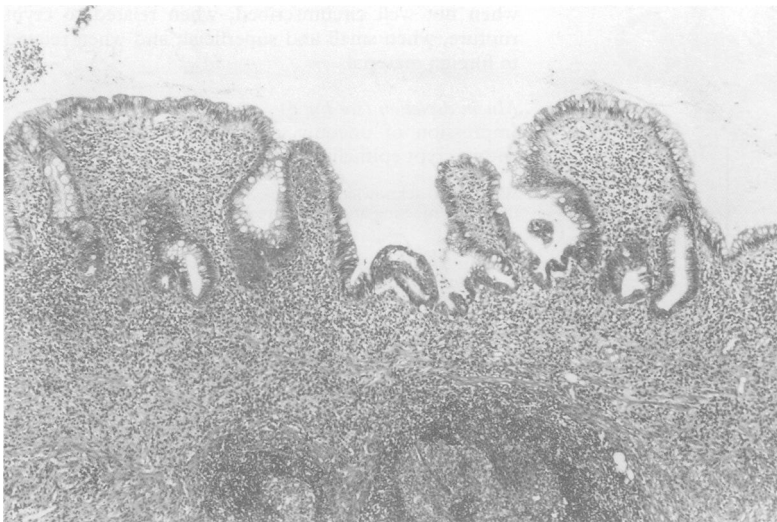


Figure 2: Biopsy specimen showing crypt atrophy, crypt banding, crypt distortion, and a villous surface (haematoxylin and eosin, original magnification $\times 40$).



Figure 3: Biopsy specimen showing neutrophil infiltration (haematoxylin and eosin, original magnification $\times 100$).

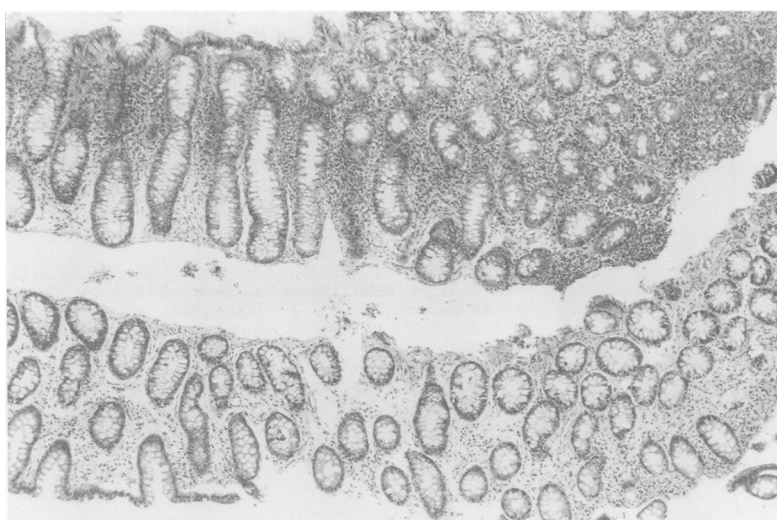


Figure 4: Biopsy specimen showing discontinuous inflammation (haematoxylin and eosin, original magnification $\times 40$).

histopathological data. In addition, sequential information may be required over a period of time.^{3 19} To ask for the final diagnosis on the basis of two slides, taken out of the usual context of histopathology decision making, may be considered unrealistic. We have found considerable disagreement, however, between experienced observers reviewing the same slides. In addition, this disagreement exists whether the biopsy specimens are taken from the rectum or colon, either through the sigmoidoscope or the colonoscope, and whether they are active or quiescent, and disagreement is intimately related to inaccuracy since some of the reported observations will be correct and others incorrect. Furthermore, deeper sections or additional biopsy specimens will not necessarily guarantee better agreement.

It is therefore our unshaken belief that our study has shown considerable error in using colorectal biopsy specimens for classifying patients. In addition, we have identified features with such poor reproducibility that we recommend caution when using them for decision making. The strategy in Table IV should be used as an aid, in conjunction with clinical judgment and the observer's intuition. Furthermore, in clinical decision making, the recommendation from the strategy should be carefully integrated with other objective data obtained from the clinical, laboratory and radiological evaluation of the patient.

Appendix

DEFINITIONS

Crypt branching (see Fig 2)

The presence of two or more branched crypts in an adequate sized biopsy specimens. Crypt branching needs to be distinguished from the normal branching seen between mucosal hillocks.

Crypt distortion (see Fig 2)

Irregular shaped, non-parallel, variable diameter, cystic as opposed to the normal parallel, test tube shaped crypts.

Crypt atrophy (see Fig 2)

Refers to the increased gap (space) between the base of the crypt and muscularis mucosa, but not necessarily crypt themselves. There are fewer crypts and there is an increased ratio of lamina propria to crypt epithelium. Crypts may be reduced, normal or increased in height.

It is worth mentioning that all of these definitions can only really be applied if there is an adequate sized biopsy that has been appropriately orientated. In addition, if there is crypt atrophy and crypt distortion, it is difficult to identify crypt branching.

We recommend that the three definitions, crypt branching, crypt distortion, and crypt atrophy, be grouped as 'architectural abnormalities' and that this feature is of major importance in discriminating between normal biopsies and those taken from chronic inflammatory bowel disease.

Neutrophil infiltration (see Fig 3)

The unequivocal presence of neutrophils infiltrating the lamina propria or the crypt epithelium, or both, as opposed to neutrophils found inside small blood vessels. The presence of even a few neutrophils is a relevant finding. They may be present either in the epithelium, the lamina propria, or they may be found in relation to a ruptured crypt.

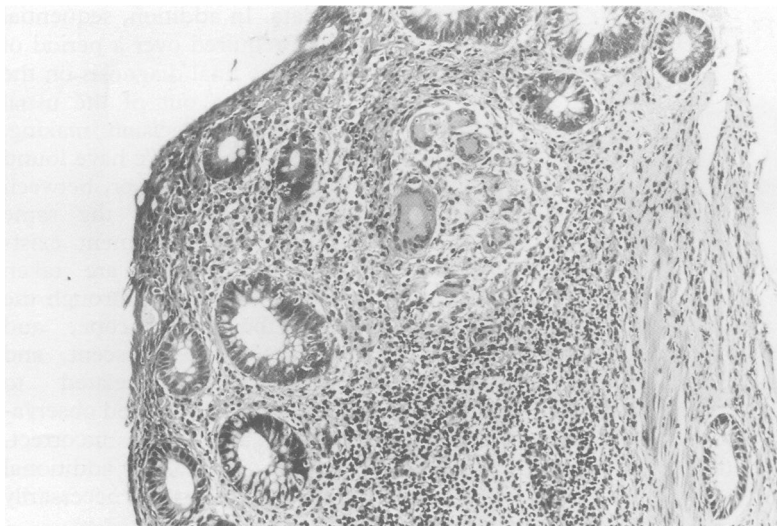


Figure 5: Biopsy specimen showing mucin depletion (haematoxylin and eosin, original magnification $\times 100$).

Eosinophils

These are present in small numbers in the normal lamina propria. There is no definition of the baseline of normality, but the subjective impression of increased eosinophilia was reasonably consistent between observers. Such factors as thickness of sections, staining and degree of accompanying inflammation, will influence subjective impression.

Focal inflammation has been re-named as discontinuous inflammation (Fig 4)

This is defined as the presence of oedematous or normal mucosa, as defined by the presence of a normal architecture and the absence of inflammation, together with the presence of active inflammation in another part of the same biopsy.

Scattered lymphoid aggregates do not constitute reliable evidence of active inflammation. Demonstration of skip areas within a colonoscopic series would be indicative of discontinuous inflammation.

Epithelioid granuloma (see Fig 5)

An aggregate of epithelioid histiocytes with or without multi-nuclear giant cells. Caution should be exercised

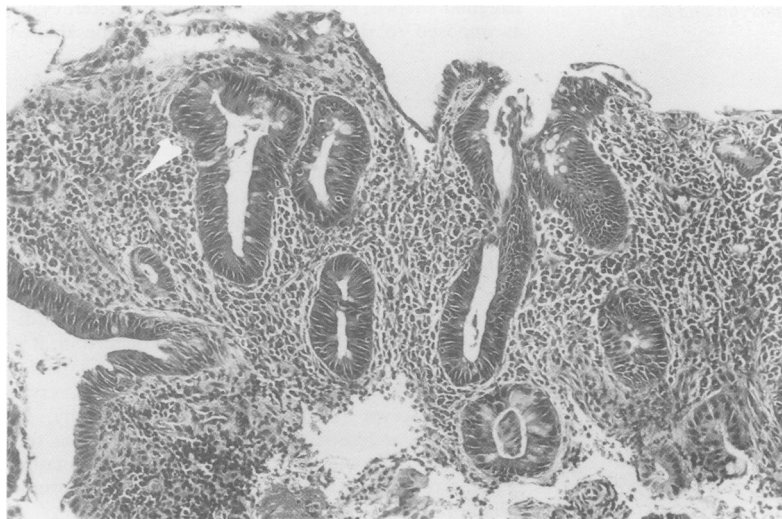


Figure 6: Biopsy specimen showing granuloma (haematoxylin and eosin, original magnification $\times 100$).

when not well circumscribed, when related to crypt rupture, when small and superficial, and when related to foreign material.

Mucin depletion (see Fig 6)

Impression of unequivocal reduction of goblet cells within crypt epithelium.

We wish to acknowledge Merck Sharp and Dohme; support for the panel meeting and video preparation.

- Durley P, Weston PMT, Williams NS. Colonoscopy or barium enema as initial investigation of colonic disease. *Lancet* 1987; ii: 549-51.
- Lindsey DC, Freeman JG, Cobden I, Record CU. Should colonoscopy be the first investigation for colonic disease? *BMJ* 1988; 296: 167-9.
- Talbot IC, Price AB. *Biopsy pathology in colorectal disease*. London: Chapman and Hall, 1987.
- Cook MG, Dixon MF. Analysis of the reliability of detection and diagnostic value of various pathological features in Crohn's disease and ulcerative colitis. *Gut* 1973; 14: 255-62.
- Giard RWM, Hermans J, Ruitter DJ, Hoedemaeker PJ. Variations in histopathological evaluation of non-neoplastic colonic mucosal abnormalities, assessment and clinical significance. *Histopathology* 1985; 9: 535-41.
- Surawics CN, Belic L. Rectal biopsy helps to distinguish self-limited colitis from idiopathic inflammatory bowel disease. *Gastroenterology* 1987; 92: 318-28.
- Allison MC, Hamilton-Dutoit SJ, Dhillon AP, Pounder RE. The value of rectal biopsy in distinguishing self-limited colitis from early inflammatory bowel disease. *Q J Med* 1987; 65: 985-95.
- Fleiss JL. *Statistical methods for rates and proportions*. 2nd Ed. New York: Wiley, 1981.
- Armitage P, Berry G. *Statistical methods in medical research*. Blackwell, Oxford.
- Rotterdam H, Korelitz BI, Sommers SC. Microgranulomas in grossly normal rectal mucosa in Crohn's disease. *Am J Clin Pathol* 1977; 67: 550-4.
- Myren J, Serck-Hanssen A, Solberg L. Routine and blind histological diagnosis on colonoscopic biopsies compared to clinical - colonoscopic observations in patients with and without colitis. *Scand J Gastroenterol* 1976; 11: 135-40.
- Yardley JH, Donowitz M. Colorectal biopsy in inflammatory bowel disease. In: Yardley JH, Morson BC, Abell MR, eds. *The gastrointestinal tract*. Baltimore: Williams and Wilkins Co, 1977: 50-94.
- Glotzer DJ, Gardner RC, Goldman H, Hinrichs HR, Rosen H, Zetzel L. Comparative features and course of ulcerative colitis and granulomatous colitis. *New Engl J Med* 1970; 282: 582-7.
- Schachter H, Goldstein MJ, Rappaport H, Fennesy JJ, Kirsner JB. Ulcerative and granulomatous colitis - validity of differential diagnostic criteria. *Ann Intern Med* 1970; 72: 841-51.
- Clamp SE, Myren J, Bouchier RAD, Watkinson G, De Dombal FT. Diagnosis of inflammatory bowel disease: an international multi-centre scoring system. *BMJ* 1982; 284: 91-5.
- Anonymous. Which type of colitis? [Editorial]. *Lancet* 1988; i: 336.
- Frei JV, Morson BC. Medical audit of rectal biopsy diagnosis of inflammatory bowel disease. *J Clin Pathol* 1982; 35: 341-4.
- Hywell Jones J, Lennard-Jones JE, Morson BC, Chapman M, Sackin MJ, Sneath PHA, Spicer CC, Card WI. Numerical taxonomy and discriminant analysis applied to non-specific colitis. *Q J Med* 1973; 168: 715-32.
- Price AB. Overlap in the spectrum of non-specific inflammatory bowel disease 'colitis indeterminate'. *J Clin Pathol* 1978; 31: 567-77.
- Goldman H, Antonioli DA. Mucosal biopsy of the rectum, colon and distal ileum. *Hum Pathol* 1982; 13: 981-1012.
- Whitehead R. Ulcerative colitis, Crohn's colitis, ischaemic colitis, pseudomembranous and antibiotic-associated colitis. In: *Mucosal biopsy of the gastrointestinal tract*. 3rd Ed. Whitehead R. London: WB Saunders Company, 1985: 213-57.
- Price AB, Morson EC. Surgical pathology of Crohn's disease and ulcerative colitis. *Hum Pathol* 1975; 6: 7-29.
- Heatley RV, James PD. Eosinophils in the rectal mucosa. *Gut* 1978; 20: 787-91.
- Willoughby CP, Piris J, Truelove SC. Tissue eosinophils in ulcerative colitis. *Scand J Gastroenterol* 1979; 14: 395-9.
- Anthonisen P, Riis T. Eosinophilic granulocytes in the rectal mucosa in patients with ulcerative colitis and Crohn's disease of the ileum and colon. *Scand J Gastroenterol* 1971; 6: 731-4.
- O'Leary AD, Sweeney EC. Lymphoglandular complexes of the colon; structure and distribution. *Histopathology* 1986; 10: 267-84.