J Clin Pathol 1983;**36**:385–391

Observer variation in the histological grading of rectal carcinoma

GDH THOMAS,* MF DIXON,* NC SMEETON,† NS WILLIAMS‡

*From the Departments of *Pathology, †Statistics and ‡Surgery, University of Leeds*

SUMMARY The variation between two observers in grading 100 biopsies and the corresponding main specimens of rectal carcinomas has been examined. Using kappa statistics, which take account of chance agreement, we found a highly significant level of agreement. As expected, higher levels were obtained for intraobserver agreement. However, disagreements between observers were in many instances "haphazard" and there were differences in bias between them. Fifty paired biopsies and main tumours were graded by five observers and the results analysed for bias and by kappa statistics for overall and conditional agreement. These methods revealed significant overall agreement but the levels for some observer pairs did not differ significantly from chance. Examination for observer bias indicated differing standards of grading, and haphazard disagreements reached high levels for some observer pairs. The intraobserver agreement between the grade of the biopsy and the corresponding main tumour varied from 56–69% but only 52% of the poorly differentiated tumours were diagnosed as such in the preoperative biopsy by the "specialist" observer. The poor predictive value was not improved by taking multiple biopsies. We conclude that the grade of a rectal carcinoma cannot be accurately assessed on a preoperative biopsy and that this has serious implications for the management of low rectal cancers. Furthermore the wide discrepancies in diagnostic standards between some pathologists mean that studies on the treatment and prognosis of rectal cancer which utilise histological grade for comparison purposes must be viewed with considerable scepticism.

Histological grading is a routine practice in the pathological reporting of large-bowl cancer. Tumour grade has been widely employed in comparisons of treatment and as a prognostic index, although its value in this regard is much less than that of a careful assessment of the extent of spread. Increasingly, the degree of differentiation found in biopsy samples is being used in making decisions on surgical management. In particular, tumour grade has an important bearing on the choice of patients for transanal local excision[1,2] and anterior resection with low pelvic anastomosis, especially when using a stapling gun. The major consideration with the latter procedure is the adequacy of the distal margin and we have previously shown that only poorly differentiated carcinomas are likely to show distal intramural spread greater than 1 cm from the gross tumour margin.[3] Thus, in the surgical treatment of low rectal carcinomas the histological grade established by preoperative biopsy assumes major importance. Surprisingly, the relation between the grade allocated to a biopsy sample and that found in the main tumour has not been adequately assessed. Furthermore it is likely that there will be wide discrepancies between pathologists in making these highly subjective decisions. With these problems in mind we determined to answer three questions:

1 What is the intra- and interobserver variation in grading rectal carcinoma?

2 What is the correlation between the grade of a biopsy and the grade of the subsequently removed carcinoma?

3 Can this correlation be improved by taking multiple biopsies prior to the operation?

Material and methods

One hundred resection specimens of rectal adenocarcinoma which had been preceded by a biopsy deemed adequate for grading purposes, were obtained from the Department of Pathology files.

At least two paraffin-processed haematoxylin and

Accepted for publication 17 November 1982

eosin stained sections of the tumour in the major resection specimen were available and where more than two blocks were taken, the two sections showing the largest area of carcinoma were selected. The majority of biopsy specimens had been sectioned at three levels; in these cases the haematoxylin and eosin section at level 2 was selected for use in this study.

Having noted which were the corresponding sections from the biopsy and the resection specimen ("main tumour"), they were relabelled with random numbers. The two principal observers, MFD (a Consultant with a special interest in gastrointestinal pathology) and GDHT (a Senior Registrar) independently graded each biopsy and main tumour into one of three grades representing well, moderate or poor differentiation. Both observers had familiarised themselves with the criteria for grading rectal cancer proposed by Dukes[4-6] and Grinnell[7] and were in keeping with those recently illustrated by Blenkinsopp *et al.*[8] Each observer repeated the grading after an interval of several weeks. Fifty biopsies and their corresponding main tumour sections were selected at random from the original group and were graded by three consultant colleagues. There was no attempt to define criteria with these observers; they were merely asked to grade them as they would a routine specimen.

In order to determine whether or not examination of multiple biopsies increases the accuracy of preoperative grading, several biopsies (3–12, mean 6·5) were taken from various parts of 32 rectal cancers whilst the patients were under anaesthesia immediately prior to surgical removal. The sections of the biopsies and the resected main tumours were coded and assessed by a single observer (MFD). Two cases (5 and 7 biopsies) had to be rejected as none contained adenocarcinoma.

STATISTICAL ANALYSIS

Previous experience of grading rectal carcinomas indicated that approximately two-thirds of the sample would be graded as moderately differentiated. A substantial level of observer agreement may therefore be accounted for by chance alone. A statistic which measures the agreement between two observers whilst taking chance into account is the kappa-statistic, introduced by Cohen,[9] given by,

$$\kappa = \frac{p_o - p_e}{1 - p_e} \qquad \ldots\ldots 1$$

where p_o and p_e are observed and expected proportions of agreement. This takes the value +1 for perfect agreement, zero for chance agreement and negative values for less than chance agreement. Where kappa lies between zero and one, significance tests exist to determine whether the value is significantly different from zero (see Cohen[9]). We assume in our study that the different types of disagreement carry equal weighting.

If the value of kappa is very close to one, then the agreement can be considered satisfactory. If it is not significantly different from zero, then we conclude that either agreement is no better than that expected by chance or that a real difference may exist but further observations would be needed to demonstrate this. If kappa is significantly positive but not close to one, further investigation may highlight reasons for disagreement.

Given unequal margins in an agreement matrix (for example, the 3 × 3 Tables 1 and 2) there is a maximum possible proportion of agreements, p_{max}, less than one. This is calculated by taking for each category, the smaller of the row or column totals, summing these values and dividing by the number of items categorised. When p_{max} is say, less than 0·8, differences in the marginals, caused by systematic bias between observers, explain a considerable amount of disagreement. Denote by p_s the proportion of disagreements explained by systematic bias or *systematic* disagreement $(1-p_{max})$. The disagreement which cannot be explained by differing marginals is the difference between p_o and p_{max} and is *haphazard* disagreement, p_h. This shows up inconsistencies such as if observer A labels a case *poor* and B labels it *moderate* and then the next case is labelled oppositely. Retraining of the observers to agree on the proportions in each category would minimise systematic disagreement. This agreement on proportions is essential since complete agreement is otherwise impossible. Further training is then needed to define more precisely the characteristics of each category and hence reduce haphazard disagreement.

If p_{max} is not close to one because of marginal differences, then it is possible to quantify the nature of these differences with two coefficients. These indicate the direction of disagreements between two observers. The "optimism" (that is, the tendency to allocate better differentiation) of B relative to A is defined by

OPT = (proportion *well* by B − proportion *well* by A) − (proportion *poor* by B − proportion *poor* by A). ……2

If OPT is positive then B is more optimistic than A, whilst a negative value indicates relative pessimism. A difference between OPT and zero of more than 0·2 indicates considerable differences in opinion between A and B. When looking at intraobserver agreement, changes in OPT between two sets of grades are a measure of "observer drift".

The second coefficient is "centrality". The cen-

trality of B relative to A is defined by

CEN = proportion *moderate* by B − proportion *moderate* by A. 3

A positive value of CEN indicates that B is less ready to give extreme allocations than is A. Again, cases where CEN differs from zero by more than 0·2 are particularly indicative of differences in boundaries between respective classes. In the tables of results, values of OPT and CEN of 0·2 or more are indicated with an asterisk. The two coefficients are illustrated diagrammatically in Figs. 1 and 2.

It is possible, when the total of disagreements is large, for p_{max} to be fairly close to one and yet there is considerable imbalance in optimism or centrality between two observers. This is probably because haphazard disagreement is still the primary source of disagreement.

Recapping,

$$p_o + p_s + p_h = 1 \quad .....4$$

From the point of view of clinical management, the recognition of poorly differentiated carcinomas in preoperative biopsies is of major importance. Therefore the conditional agreement between observers was examined by calculating the probability that one observer reports poor differentiation given that the other did so.

Results

The allocation of grades and the intraobserver agreement between the two "runs" for the two principal observers are shown in Table 1—biopsies and

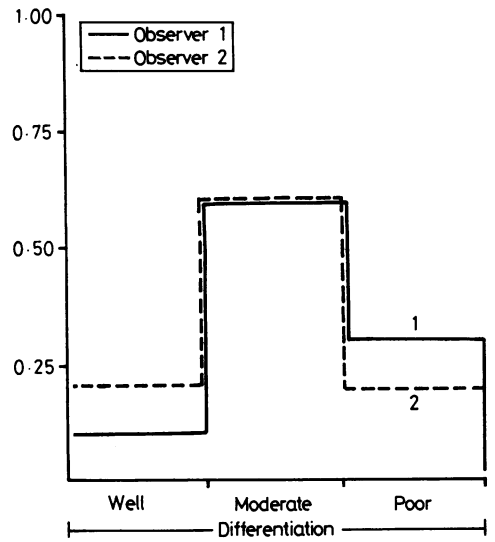

Fig. 1 *Histogram showing proportions of grades allocated by two observers 1 and 2. Observer 2 demonstrates relative optimism but there is no difference in centrality.*

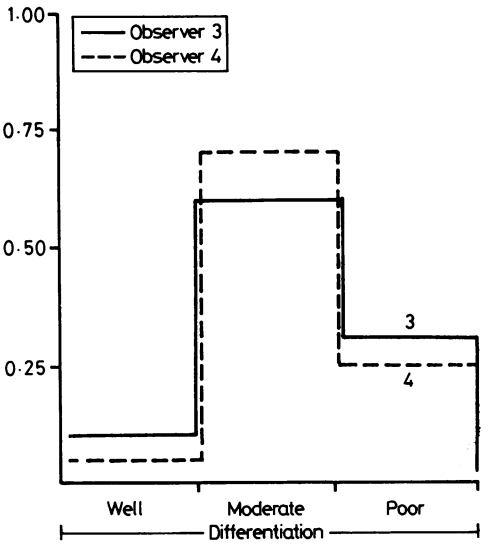

Fig. 2 *In this example the two observers (3 and 4) demonstrate differing centrality. Observer 4 is less prepared than observer 3 to give grades other than moderate.*

Table 1 *Intraobserver agreement in grading 100 biopsies*

		MFD Run 1			*% Agreement*	*Kappa*	*95% confidence limits*
		Well	*Mod*	*Poor*			
MFD Run 2	Well	5	6	0			
	Mod	2	54	1	80%	0·613	0·461–0·765
	Poor	0	11	21			
		GDHT Run 1					
		Well	*Mod*	*Poor*			
GDHT Run 2	Well	4	4	0			
	Mod	11	44	3	74%	0·539	0·387–0·691
	Poor	0	8	26			

Table 2—main tumours. High percentage agreements and kappa values which are significantly greater than zero ($p < 0.001$) were obtained for both observers. Nevertheless the kappa values are not close to 1 and the nature of the disagreements can be explored further.

Taking the biopsy allocations of MFD, the sum of the lowest row and column marginals (7 + 57 + 22) is 86. Thus p_{max} is 0·86 and the systematic disagreement (p_s) between the two runs is 0·14. Haphazard disagreement (p_h) given by $p_{max} - p_o$ is 0·06.

$$OPT = (0.11 - 0.07) - (0.32 - 0.22) = -0.06 \quad \text{.....(equation 2).}$$

$$CEN = 0.57 - 0.71 = -0.14 \quad \text{.....(equation 3).}$$

From these results it would appear that the kappa value is depressed largely as a consequence of a change in centrality between the two runs. The biopsy allocations of GDHT give $p_s = 0.07$, $p_h = 0.19$, OPT = −0·12 and CEN = +0·02. Thus there is relative pessimism comparing run 2 to run 1 and haphazard disagreement in 19% of cases.

For main tumour grades allocated by MFD (Table 2) the two runs give $p_s = 0.06$, $p_h = 0.14$, OPT = 0 and CEN = −0·06. There is thus more haphazard disagreement than with the biopsies but no change in optimism and less difference in centrality. The latter factors explain the slightly higher kappa value.

Main tumour grades for GDHT give $p_s = 0.09$, $p_h = 0.14$, OPT = −0·13 and CEN = −0·05. Compared with the biopsy results there is less haphazard disagreement and this explains the higher kappa value.

The agreement between the two principal observers in grading biopsies and main tumours (using run 1) is shown in Table 3. MFD has much more centrality and is less optimistic than GDHT, hence the values of kappa. Haphazard disagreement accounts for 10% of biopsy and 21% of main tumour cases.

The levels of agreement between five observers

Table 2 *Intraobserver agreement in grading 100 main tumours*

		MFD Run 1			*% Agreement*	*Kappa*	*95% confidence limits*
		Well	*Mod*	*Poor*			
MFD Run 2	Well	9	4	0			
	Mod	1	51	6	80%	0·630	0·485–0·775
	Poor	0	9	20			
		GDHT Run 1					
		Well	*Mod*	*Poor*			
GDHT Run 2	Well	7	2	0			
	Mod	6	54	3	77%	0·546	0·383–0·709
	Poor	0	12	16			

Table 3 *Interobserver agreement (MFD v GDHT) for grades assigned to biopsies and main tumours (n = 100)*

	po	*pe*	*pmax*	*Kappa*	*ps*	*ph*
Biopsies	0·77	0·4817	0·87	0·556 (0·397, 0·715)	0·13	0·10
Main tumour	0·72	0·4976	0·93	0·443 (0·268, 0·618)	0·07	0·21

Table 4 *Agreement between five observers grading 50 biopsies*

	po	*Kappa*	*95% limits*	*ps*	*ph*	*OPT*	*CEN*
AB	0·74	0·500	(0·266, 0·734)	0·14	0·12	0·10	−0·16
AC	0·60	0·266	(0·017, 0·515)	0·22	0·18	0·28*	−0·16
AD	0·72	0·410	(0·147, 0·672)	0·10	0·18	−0·12	−0·08
AE	0·54	0·210	(−0·027, 0·447)	0·26	0·20	−0·06	−0·26*
BC	0·48	0·115	(−0·121, 0·351)	0·12	0·40	0·22*	−0·02
BD	0·58	0·231	(−0·020, 0·481)	0·12	0·30	−0·18	0·06
BE	0·64	0·414	(0·198, 0·631)	0·12	0·24	−0·12	−0·12
CD	0·60	0·312	(0·079, 0·546)	0·24	0·16	−0·40*	0·08
CE	0·54	0·279	(0·062, 0·496)	0·22	0·24	−0·34*	−0·10
DE	0·58	0·284	(0·051, 0·517)	0·18	0·24	0·06	−0·18

Table 5 *Agreement between five observers grading 50 main tumours*

	p_o	Kappa	95% limits	p_s	p_h	OPT	CEN
AB	0·74	0·447	(0·188, 0·706)	0·10	0·16	−0·06	0·06
AC	0·68	0·391	(0·145, 0·637)	0·06	0·26	0·06	−0·06
AD	0·78	0·532	(0·288, 0·776)	0·08	0·14	0·10	0·06
AE	0·56	0·258	(0·026, 0·490)	0·22	0·22	−0·06	−0·22*
BC	0·66	0·333	(0·075, 0·590)	0·12	0·22	−0·08	−0·12
BD	0·76	0·456	(0·187, 0·724)	0·02	0·22	−0·04	0
BE	0·70	0·500	(0·288, 0·712)	0·28	0·02	−0·20*	−0·28*
CD	0·64	0·290	(0·028, 0·553)	0·12	0·24	0·04	0·12
CE	0·50	0·180	(−0·048, 0·407)	0·16	0·34	−0·12	−0·16
DE	0·52	0·194	(−0·041, 0·428)	0·28	0·20	−0·16	−0·28*

Table 6 *Intraobserver agreement between biopsy and corresponding main tumour grade for five observers*

Observer	n	p_o	Kappa	95% limits	p_s	p_h
MFD	100	69	0·358	(0·170, 0·548)	0·09	0·22
GHT	100	69	0·420	(0·250, 0·590)	0·10	0·21
C	50	60	0·317	(0·085, 0·549)	0·16	0·24
D	50	64	0·249	(−0·029, 0·526)	0·12	0·24
E	50	56	0·296	(0·076, 0·516)	0·04	0·40

Table 7 *Agreement of second observer given that first allocates as poor (%)*

(a) Biopsy

Observer		Second				
		A	B	C	D	E
First	A	—	82	60	90	80
	B	67	—	42	75	83
	C	86	71	—	100	86
	D	60	60	47	—	80
	E	44	56	33	67	—

(b) Main tumour

Observer		Second				
		A	B	C	D	E
First	A	—	54	69	62	69
	B	74	—	88	88	100
	C	69	54	—	54	77
	D	89	78	78	—	89
	E	45	40	50	40	—

Table 8 *Intraobserver agreement when (a) biopsies and (b) main tumours are designated poorly differentiated (%)*

Observer	Biopsy → main tumour	Main tumour → biopsy
A	67	52
B	48	68
C	86	46
D	40	67
E	72	65

grading the same 50 biopsies and main tumours are shown in Tables 4 and 5. In general, kappa values obtained between observer pairs including C, D or E are slightly lower than those between the principal observers, A (MFD) and B (GDHT), especially with the biopsies. All values are much less than one but most are significantly more than zero. A and B do not have great differences in optimism and centrality between each other, compared with other observer pairs. For biopsies, C is very much more optimistic with respect to the others and seems to be most out of line. E is much less centralised than A.

For the main tumours, E is much more pessimistic and much less centralised than the rest and is most out of line. Overall agreement with A is significant since all other observers individually agree significantly with A. For biopsy grades haphazard disagreement is lowest between A and B (12%) but is very high for some other pairs (up to 40%). With the main tumour grades, however, two other pairings achieved lower haphazard disagreement and higher kappa values.

For each observer, the agreement between biopsy and main tumour grades are given in Table 6. A and B have the highest kappa values, followed by C and E and lastly D, whose value is not significant. Haphazard disagreements are large in all cases (> 20%), so this is a cause for concern. Most observers have differences in centrality or optimism between biopsy and main tumour grades.

Looking at the conditional agreements of one observer given that another has allocated poor

(Table 7) we see that even between A and B the chance of agreement can be as low as 54%. This illustrates the point that since gradings of poor differentiation are relatively scarce, large percentage-wise errors can be made with respect to poor gradings without affecting the value of kappa unduly. This is a pitfall in using kappa and it is always advisable to look at conditional probabilities. Between other observers, agreement drops to as little as 33%.

The intraobserver agreement between the grade given to the biopsy and the main tumour with respect to those labelled poor is given in Table 8. It can be seen that for biopsies labelled poor, agreement with the corresponding main tumour grade dropped as low as 40%.

When the grade based on multiple biopsy samples was compared with the grade of the corresponding main tumour (MFD), the level of agreement (53%, $k = 0{\cdot}172$, $n = 30$) was apparently worse than with a single biopsy, although this did not reach statistical significance.

Discussion

In a paper published in 1953 Qualheim and Gall[10] addressed themselves to the question "Is histologic grading of colon carcinoma a valid procedure?" On demonstrating that only 28% showed a consistent histological pattern throughout the tumour, these authors concluded that most biopsy samples would not be representative of the main tumour and that there was no basis for using histological grade as a means of determining prognosis. However, heterogeneity in itself need not prevent uniformity of grading, as this should be based on the least differentiated area represented. Failure to achieve individual consistency arises from the subjective nature of the diagnostic criteria and the inevitable overlap between categories. When interobserver uniformity is sought, however, more serious differences in basic interpretation become apparent. This point was amplified very forcibly by the large survey of histopathology reporting of large bowel cancer involving pathologists at 22 hospitals conducted by Blenkinsopp *et al.*[8] The proportion of carcinomas placed in each category was, well differentiated 3–93%, moderately 8–82% and poorly 5–30%, variations which can only be the result of widely different standards of assessment.

Our findings indicate that the level of intra- and interobserver agreement in grading can be improved when the observers acquaint themselves with the criteria laid down by Dukes and other experts. When "non-specialists" are involved agreement rates tended to fall and in some cases failed to reach statistically significant levels. We have demonstrated that agreements between observers which can be identified as "haphazard" are very high for some observer pairs. These are disturbing findings as the study was performed with experienced pathologists from the same department and one would have expected "mutual education" to have resulted in a more uniform standard of grading. The analysis reveals that observers show wide variations in their bias towards one or other end of the differentiation spectrum and in their willingness to depart from the middle ground and allocate extreme grades. Interestingly, with some observers this bias appeared to differ between the grading of biopsies and main tumours.

The relation between biopsy and main tumour grades has not been adequately explored. Dukes[4] undertook a similar exercise on 20 consecutive rectal cancers and found agreement in only four cases, although his classification included a fourth category—colloid carcinoma, which would inevitably diminish agreement. We found a 56–69% agreement (depending on the observer) between biopsy and the corresponding main tumour grade. Whilst in general terms this is an improvement over the agreement found by Dukes, our surgical colleagues can draw little assurance from these levels when dealing with individual patients. In terms of surgical management, poorly differentiated carcinoma is the most important group to identify. If we consider the "specialist" observer (MFD), although an overall level of agreement between biopsy and tumour of 69% was obtained, for main tumours graded as poorly differentiated only 52% of the corresponding biopsies were in agreement. In other words, almost one half of poorly differentiated carcinomas were not identified as such in the preoperative biopsy even when graded by an experienced observer. Furthermore the levels of agreement between observers over the allocation of poor differentiation were surprisingly low, the worst levels being 33% for biopsies and 40% for main tumours.

It has been suggested that rectal carcinomas tend to be better differentiated towards the surface[4] and our own experience would seem to support this observation. It could be anticipated therefore that higher degrees of differentiation would be found in biopsies than in main tumours and that this may account for the poor correlation between the two. Unexpectedly, in cases where there was a discrepancy between the biopsy and tumour grades allocated by the principal observers, we found that in a slight majority the biopsies had been given a worse grade than the corresponding main tumours (32/62 cases).

It was hoped that such discrepancies could be minimised by sampling several parts of the tumour

but there was no improvement in the level of agreement when the grade was based on multiple biopsies; indeed, if anything, it was somewhat worse. This apparent paradox may be partly explained by the "quality" of the biopsies obtained; some were fragments of necrotic tumour, others were traumatised and many were composed of uninvolved mucosa which had been inadvertently sampled (positive biopsies 1–9, mean = 4·6). Thus, in a few cases, the amount of tumour on which grading had to be based was less than in the biopises used in the retrospective series. However, we do not feel that this detracts from the conclusion that multiple biopises will not substantially improve the level of agreement. Taken overall, our findings would appear to negate the value of preoperative biopsy of rectal carcinomas for grading purposes.

Perhaps pathologists should be exploring novel methods of neoplastic grading such as that proposed by Zajicek[11] which describes cell proliferation in terms of Galilean geometry. On the other hand, it may well be that histological grading will ultimately be abandoned in favour of the immunohistochemical detection of tumour products which are quantitatively related to the invasive or metastatic potential of the carcinoma.[12] In this way, immunohistochemical studies on biopsies of rectal carcinomas may provide much more accurate guidance in the planning of surgical procedures.

We would like to thank our colleagues Dr RW Blewitt, Dr PN Cowen and Dr I Lauder for participating in this study and Miss Helen J Swiercz for typing the manuscript.

References

[1] Lock MR, Cairns DW, Ritchie JK, Lockhart-Mummery HE. The treatment of early colorectal cancer by local excision. *Br J Surg* 1978;**65**:346–9.

[2] Localio SA, Eng K. Current concepts in cancer: sphincter-saving operations for cancer of the rectum. *N Engl J Med* 1979;**300**:1028–30.

[3] Williams NS, Dixon MF, Johnston D. Distal intramural spread and local recurrence of rectal cancer. *Br J Surg* 1979;**66**:890.

[4] Dukes CE. The classification of cancer of the rectum. *J Pathol Bacteriol* 1932;**35**:323-32.

[5] Dukes CE. Histological grading of rectal cancer. *Proc R Soc Med* 1937;**30**:371–6.

[6] Dukes CE. The relation of histology to spread in intestinal cancer. *Br J Cancer* 1950;**4**:59–62.

[7] Grinnell RS. The grading and prognosis of carcinoma of the colon and rectum. *Ann Surg* 1939;**109**:500–33.

[8] Blenkinsopp WK, Stewart-Brown S, Blesovsky L, Kearney G, Fielding LP. Histopathology reporting in large bowel cancer. *J Clin Pathol* 1981;**34**:509–13.

[9] Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Measurement* 1960;**20**:37–46.

[10] Qualheim RE, Gall EA. Is histologic grading of colon carcinoma a valid procedure? *Arch Pathol* 1953;**56**:466–72.

[11] Zajicek G. On neoplastic grading. *Med Hypotheses* 1981;**7**:503–10.

[12] Buckley CH, Fox H. An immunohistochemical study of the significance of HCG secretion by large bowel adenocarcinomata. *J Clin Pathol* 1979;**32**:368–72.

Requests for reprints to: Dr MF Dixon, Department of Pathology, University of Leeds, Leeds LS2 9JT, England.