Variations in histopathological evaluation of non-neoplastic colonic mucosal abnormalities; assessment and clinical significance

R.W.M. GIARD*, J. HERMANS†, D.J.RUITER* & PH.J.HOEDEMAEKER*

*Departments of *Pathology and †Medical Statistics, University Medical Centre, State University of Leiden, The Netherlands*

Accepted for publication 6 July 1984

GIARD R.W.M., HERMANS J., RUITER D.J. & HOEDEMAEKER PH. J. (1985) *Histopathology* **9**, 535–541

Variations in histopathological evaluation of non-neoplastic colonic mucosal abnormalities: assessment and clinical significance

The variation between three pathologists examining histological features seen in non-neoplastic colonic mucosa from 40 biopsies was analysed. Several procedures to express observer variation were used and compared, with emphasis on kappa statistics. Only five features, the presence of ulceration, villous surface, epithelioid granulomas, severe mucus depletion and crypt abscesses were sufficiently reproducible by the three pairs of pathologists. These findings suggest that other criteria used for the classification of inflammatory bowel disease are potentially unsatisfactory. When results from different studies on biopsies are being compared, influence of observer variation should be identified. Comparison of statistical techniques showed overall variation to be less useful than other statistical procedures. There was little difference between results from kappa statistics and other measures of agreement (overall agreement excluded).

Keywords: colon, inflammatory bowel disease, observer variation, kappa statistics

Introduction

Pathologists are required to examine an increasing number of colorectal mucosal biopsies from patients with inflammatory bowel disease, for histological classification. Non-infective colonic inflammatory bowel disease is often difficult to classify in the absence of pathognomonic histological features. Histological diagnosis most often depends on a combination of several mucosal abnormalities, especially when key histological features such as epithelioid granulomas (for Crohn's disease) and severe mucus depletion (for ulcerative colitis) are absent. Epithelioid granulomas

Address for correspondence: R.W.M. Giard, Pathologisch Laboratorium, St. Clara Ziekenhuis, Olympiaweg 350, 3078 HT Rotterdam, Netherlands.

for example are absent in 75% of biopsies from 'proven' cases (Petri *et al.* 1982). Attempts have been made to find additional discriminating morphological features. More subtle abnormalities such as 'microgranulomas' (focal histiocytic aggregates), histiocytic-eosinophil cryptitis and focal inflammation have been identified as useful (Rotterdam, Korelitz & Sommers 1977, Yardley & Hamilton 1980, Yardley & Hamilton 1981).

Ideally a histological diagnosis should be based on both reliable morphological features and uniform criteria. If these requirements are fulfilled, a reliable diagnosis can be made and results from different pathologists are comparable and interchangeable. However, for many diagnostic situations in pathology, these criteria are difficult to fulfil, especially histopathological classification of inflammatory bowel disease. To study reproducibility and reliability inter- and intraobserver study is a convenient tool. The methodology and analysis of these observer-variation studies has been greatly developed during the last two decades (Cohen 1960, Koran 1975, de Dombal 1976, Komaroff 1979, Fleiss 1981, Holman *et al.* 1982). Many studies on this subject in diagnostic histopathology have recently been published (Thomas *et al.* 1983, Stenkvist, Bengtsson & Eriksson 1983, Holman *et al.* 1983).

Assessment of the degree of observer variation may identify diagnostic problem areas and may have great bearing on both histopathological and clinical practice. For example Thomas *et al.* (1983) examined differences between pathologists when grading rectal adenocarcinomas on biopsies. They concluded that it could be hazardous to make surgical decisions based on such gradings alone since grading, especially with poorly differentiated adenocarcinoma, was far from consistent.

How reproducible are colonic mucosal abnormalities in inflammatory bowel disease and what is their contribution to histological diagnosis? We have examined the observer variation when scoring these abnormalities. Several measures are used to express the magnitude of variation (Fleiss 1981), and we have tried to evaluate the bearing of observer variation on the histological classification of inflammatory bowel disease. Different features have been evaluated in order of reliability of reproducibility and compared one with another.

Materials and methods

A set of 40 sections (4 μm thick, paraffin-embedded, H & E stained) was selected. The biopsies were from normal or inflamed mucosa only. The specimens were derived from patients with a clinical diagnosis of idiopathic inflammatory bowel disease. All slides were independently examined by three different experienced consultant pathologists who have a special interest in gastrointestinal histopathology. Before histological examination, a list of histological definitions or morphological descriptions was provided to the observers. The list of histological features together with scoring conventions is given in Table 1. Most features were nominal variables. Items such as mucus cell depletion, epithelial atypia and laminal cellularity were scored semi-quantitatively. All data were entered in a computer file

Table 1. Histological features examined by the three pathologists

Luminal features
Inflammatory exudate (present, absent)
Epithelial features
Mucosal architecture (flat, irregular, villous)
Integrity of mucosal surface (intact, erosion, ulceration)
Crypt architecture (normal, irregular, branching)
Mucous-cell depletion (absent, slight, severe)
Mucosal atrophy (absent, present)
Adenomatous change (absent, present)
Atypia (absent, if present grade 1 to 3)
Paneth-cell metaplasia (present, absent)
Laminal features
Cellularity (normal, increased, severely increased)
Composition of infiltrate (mononuclears only, admixture of neutrophils)
Eosinophilia (present, slight, severe)
Granuloma (present, absent)
Focal histiocytic aggregates (present, absent)
Increased mucosal vascularity (present, absent)
Neutrophil cryptitis (absent, slight, severe)
Crypt abscesses (absent, slight, severe)
Histiocytic-eosinophil cryptitis (present, absent)
Mucosal distribution of infiltrate (diffuse, patchy, focal)
Submucosal extension of inflammation (present, absent, no submucosa)

(DEC PDP 11/70) and subsequently processed using the SPSS-package for preparation of frequency and cross tables. For each pair of observers data from cross tables was used to calculate the following indices of agreement for each feature: 1 overall agreement; 2 kappa value; 3 specific agreement; 4 lambda; and 5 proportional agreement. For definition and further explanation of these indices please see appendix. In addition all observers were asked to give a histological diagnosis for each biopsy.

For evaluation of possible discrepancy between different measures for expressing observer variation, a Pearson correlation coefficient was calculated.

Results

When overall agreement was used alone, more features would show sufficient reproducibility then when the kappa score was used. Since influence of chance is not excluded in overall agreement, it is considered to be a poor indicator of observer variation (Koran 1975). We have therefore discarded this parameter for this study.

Using kappa values, we compared results of three pairs of observers. According to Landis & Koch (1977) features having a kappa value of less then 0.40 were regarded as representing poor agreement beyond chance. A value of more than

Table 2. Features with consistent kappa values in all three observers

Kappa>0.40	Kappa<0.40
Normalities	*Normalities*
Normal crypt architecture	Intact mucosal surface
Normal cellularity	No submucosal extension
	No submucosal tissue
Abnormalities	*Abnormalities*
Mucosal ulceration	Inflammatory exudate
Villous surface	Mucosal erosion
Epithelioid granulomas	Irregular crypt spacing
Severe mucus depletion	Less then severe atypia
Crypt abscesses	Distribution of infiltrate
	Submucosal extension

0.40 was considered as fair to good agreement and a score of 0.75 was excellent. Three groups of features were seen. The first group (Table 2) consists of sufficient agreement between all three pairs. The second group (Table 3) comprises poor agreement between two of the three pairs. The third group (Table 2) showed low kappa scores between all three pairs.

Histological diagnosis of ulcerative colitis showed kappa values ranging from 0.38 to 0.64 (with two pairs above the value of 0.40). For Crohn's disease these figures were 0.27 and 0.50 respectively (with two pairs under the value of 0.40). The lowest kappa value was scored for diagnosis of normal colonic mucosa (−0.13!). In this category only one pair had a kappa of 0.44.

We compared the results of all different scores by means of a Pearson-correlation coefficient (Table 4). A strong correlation is shown between all indices except overall agreement. Use of indicators other than kappa value (and overall

Table 3. Histological features with inconsistent kappa values between observers (overall kappa value below 0.40)

Irregular surface
Crypt branching
Mucosal atrophy
Adenomatous change
Mild increase of cellularity
Absent mucus depletion
Moderate mucus depletion
Eosinophilia, mild
Increased vascularity
Histiocytic-eosinophil cryptitis

Table 4. Pearson correlation–coefficients between various indices of observer variation

	Overall agreement	Kappa	Specific agreement	Lambda	Proportional agreement
Overall agreement	—	0.79	0.43	0.43	0.45
Kappa	0.79	—	0.81	0.81	0.82
Specific agreement	0.43	0.81	—	0.99	0.98
Lambda	0.43	0.81	0.99	—	0.98
Proportional agreement	0.45	0.82	0.98	0.98	—

agreement) therefore made little difference with regard to the choice of reproducible criteria.

Discussion

The value of a diagnostic test is determined by its precision, accuracy and clinical usefulness (Komaroff 1979). Observer variation is a measure of test precision. Very few studies examining the importance of observer variation when classifying inflammatory bowel disease so far have been published. Cook & Dixon (1973) examined the magnitude of observer variation on macroscopical and microscopical examination of resection specimens from patients with either Crohn's disease or ulcerative colitis but using only overall agreement. From a total of 34 separate criteria, nine proved to be reliable and discriminating. In our study only five abnormal histological features showed sufficient reproducibility. Among these are two traditional 'key' discriminant features proven to be of great value for the distinction between ulcerative colitis (severe mucus depletion) and Crohn's disease (epithelioid granulomas) (Hywel Jones, Lennard Jones & Morson 1973). The other three have some or no discriminative value, although a villous mucosal surface favours a diagnosis of ulcerative colitis.

Subtle features with alleged diagnostic value for identifying Crohn's disease such as microgranulomas and histiocytic-eosinophil cryptitis showed poor reproducibility. This may imply that a pathologist is dependent on the presence of only very few abnormalities for reliable classification of inflammatory bowel disease. Unfortunately such abnormalities are often absent.

Although this study was primarily aimed at determining the magnitude of observer variation when scoring different histological features, the differences between the pathologists with regard to their diagnoses were also impressive. Diagnosis of ulcerative colitis seemed fairly consistent, whereas diagnosis of Crohn's disease and especially of normal mucosa were not.

We still know very little about the causes for discrepancies between observers. All features were described on paper and discussed. All observers came from the same institute. Since a 'gold standard' is lacking, we cannot determine the 'true state' of a biopsy and thus we do not know whether under- or overdiagnosis (or both) is contributing to the disagreement between observers. Further fundamental research in this area is badly needed.

References

COHEN J. (1960) A coefficient of agreement for nominal scales. *Educational & Psychological Measurement* **20**, 37–46

COOK M.G. & DIXON M.F. (1973) An analysis of the reliability of detection and diagnostic value of various pathological features in Crohn's disease and ulcerative colitis. *Gut* **14**, 255–262

DE DOMBAL F.T. (1976) How objective is medical data? In *Decision Making and Medical Care*, eds F.T.de Dombal & F.Gremy, pp. 33–37, North Holland Publishing Company, Amsterdam

FLEISS J.L. (1981) *Statistical Methods for Rates and Proportions* pp. 212–234, Wiley and Sons, New York

HOLMAN C.D.J., JAMES I.R., HEENAN P.J., MATZ L.R., BLACKWELL J.B., KELSALL G.R.H., SING. A. & TEN SELDAM R.F.J. (1982) An improved method of analysis of observer variation between pathologists. *Histopathology* **6**, 581-589

HOLMAN C.D.J., MATZ L.R., FINLAY-JONES L.R., WATERS E.D., BLACKWELL J.B., JOYCE P.R., KELSALL G.R.H., SHILKIN K.B., CULLITY G.J., WILLIAMS K.E., MATTHEWS M.L.V. & ARMSTRONG B.K. (1983) Inter-observer variation in the histopathological reporting of Hodgkin's disease: an analysis of diagnostic subcomponents using kappa statistics. *Histopathology* **7**, 399–407

HYWEL JONES J., LENNARD JONES J.E. & MORSON B.C. (1973) Numerical taxonomy and discriminant analysis applied to non-specific colitis. *Quarterly Journal of Medicine* **42**, 715–732

KOMAROFF A.L. (1979) The variability and inaccuracy of medical data. *Proc. IEEE* **67**, 1196–1207

KORAN L.M. (1975) The reliability of clinical methods, data and judgment. *New England Journal of Medicine* **293**, 642–646, 695–701

LANDIS J.R. & KOCH. G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174

PETRI M., POULSON S.S., CHRISTENSEN K. & JARNUM S. (1982) The incidence of granulomas in serial sections of rectal biopsies from patients with Crohn's disease. *Acta Pathologica et Immunologia Scandinavica Section A* **90**, 146–147

ROTTERDAM H., KORELITZ B.I. & SOMMERS S.C. (1977) Microgranulomas in grossly normal rectal biopsy in Crohn's disease. *American Journal of Clinical Pathology* **67**, 550–554

STENKVIST B., BENGTSSON E. & ERIKSSON O. (1983) Histopathological systems of breast cancer classification: reproducibility and clinical significance. *Journal of Clinical Pathology* **36**, 392–398

THOMAS G.D.H., DIXON M.F., SMEETON N.C. & WILLIAMS N.S. (1983) Observer variation in the histological grading of rectal carcinoma *Journal of Clinical Pathology* **36**, 385–391

YARDLEY J.H. & HAMILTON S.R. (1980) Pathological aspects of diagnosis, pathogenesis and etiology of idiopathic inflammatory bowel disease. In *Inflammatory Bowel Disease. Developments in Gastroenterology vol 3.* pp. 3–10, Martinus Nijhoff, The Hague

YARDLEY J.H. & HAMILTON S.R. (1981) Focal non-specific inflammation (FNI) in Crohn's disease. In *Recent Advances in Crohn's Disease.* pp. 5–12, Eds A.S.Pena & I.T.Weterman Martinus Nijhoff, The Hague

Appendix

All parameters described below are extensively reviewed in FLEISS 1981. For all parameters the 2×2-contingency table shown is used.

Table A1. 2×2-contingency table

	Second rater		
	+	–	Total
First rater			
+	a	b	p_1
–	c	d	q_1
Total:	p	q_2	1.0

All quantities are expressed as rates of the total number of observations, so the total sum is 1.0.

OVERALL AGREEMENT

This is the sum of rates for complete agreement between both observers ($P_0=a+d$).

KAPPA VALUE

The kappa value shows the agreement rate, whereby correction has been made for chance influences. For its calculation the overall agreement is used together with the chance expected agreement P_e (were $P_e=p_1\times q_1+p_2\times q_2$) in the following formula: kappa=$(P_0-P_e)/(1-P_e)$.

SPECIFIC AGREEMENT

This measure is especially suitable when the category under study is relatively rare. It is calculated by the formula: $P_s=2a/(2a+\mathrm{b}+c)$.

LAMBDA

This is another measure, which takes account of the discrepancies and concordance of the feature under study using the following formula: lambda=[2a−(b+c)]/[2a+(b+c)]

PROPORTIONAL AGREEMENT

This is calculated: $P_a=a/(p_1+P_2)+d/(q_1+q_2)$