What Levels of Agreement Can Be Expected Between Histopathologists Assigning Cases to Discrete Nominal Categories? A Study of the Diagnosis of Hyperplastic and Adenomatous Colorectal Polyps

Simon S. Cross, M.D., F.R.C.Path., Samar Betmouni, B.Sc., M.B.B.S., D.Phil., Julian L. Burton, M.B.B.S., Asha K. Dubé, B.Sc., M.B.B.S., M.R.C.Path., Kenneth M. Feeley, M.B.B.S., M.R.C.Path., Miles R. Holbrook, M.B.B.S., Ph.D., Robert J. Landers, M.B.B.S., M.R.C.Path., Phillip B. Lumb, M.D., M.B.A., F.R.C.Path, Timothy J. Stephenson, M.D., M.B.A., F.R.C.Path.

Department of Pathology, University of Sheffield Medical School (SSC, SB, JLB PBL), and Department of Histopathology, Royal Hallamshire Hospital, Central Sheffield University Hospitals NHS Trust (AKD, KMF, MRH, RJL, TJS), Sheffield, England

Aims: To assess the levels of agreement between histopathologists for a two-class nominal categorization process—the discrimination between hyperplastic and adenomatous colorectal polyps. **Methods:** Fifty hyperplastic and 50 adenomatous polyps received consecutively in the laboratory were categorized by nine histopathologists, and the level of agreement between all observers and the original diagnosis was assessed using kappa statistics. **Results:** For the eight observers with 11 months or more experience in histopathology, there was a high level of agreement with kappa statistics ranging from 0.84 to 0.98. This process was performed rapidly with an average of 13 to 22 seconds spent on each case. One observer with only 6-weeks' experience of histopathology had a lower overall level of agreement with kappa statistics ranging from 0.46 to 0.54, but the performance on the later cases was much higher. **Conclusions:** The level of agreement in the distinction between hyperplastic and adenomatous colorectal polyps is high among histopathologists with at least moderate amounts of experience in histopathology. The one virtually naïve observer showed a marked learning response during the study without feedback on case outcome. This suggests that histopathologists are very reliable in assigning cases to distinct nominal categories and that learning of these processes occurs early in a histopathologist's career.

Histopathology is often used as the basis for diagnosis in the treatment of patients, but we know that it is still an imperfect method of diagnosis (1, 2). Many studies have emphasised the lack of reproducibility of histopathology diagnosis by studying the assignment into arbitrary categories of cases from a biological spectrum, *e.g.*, the grading of *in situ* carcinoma of the breast (3), or by selecting a population of "difficult" cases, such as equivocal melanocytic skin tumors (4). The most common form of assessment of agreement in these studies has been the use of kappa statistics, which give a value of 1 for perfect agreement and a value of 0 for the level of agreement to be expected by chance alone (5, 6). The values of kappa vary according to the area studied, but have included 0.17 and 0.18 for the grade of gastric atrophy (7, 8), 0.23 to 0.37 for grading of dysplasia in colorectal adenomas (9), 0.41 for Gleason scoring of prostatic carcinoma (10), 0.44 for the classification of mammary ductal carcinoma *in situ* (3), and 0.50 for the benign or malignant distinction in cutaneous melanocytic lesions (4). The poor levels of agreement in these studies have been noted by publications making evidence-based recommendations for medical practice that have included titles such as "Pathology as Art Appreciation" (11). We have not identified any published studies in which a clear-cut assignment between two nominal categories has been investigated for interobserver agreement.

In this study, we use the distinction between hyperplastic and adenomatous colorectal polyps (with the exclusion of serrated adenomas) (12) as

an example of such a two-class nominal categorization problem. The distinction does have clinical importance because, although hyperplastic colorectal polyps do have reproducible genetic mutations, the risk of subsequent colorectal cancer in patients who have only hyperplastic polyps is lower than it is for those with adenomatous polyps (13).

MATERIALS AND METHODS

Slides of 50 consecutively received, endoscopically resected colorectal polyps originally reported as hyperplastic polyps and 50 consecutively received colorectal polyps originally reported as tubular, villous, or tubulovillous adenomas were retrieved from the files of the Department of Histopathology, Royal Hallamshire Hospital, from the beginning of April 1998. Serrated adenomas were not represented in this series. There was a single hematoxylin- and eosin-stained slide for each case. By chance, exactly 50 of each type were received during the same time period.

The observers were all histopathology consultants or trainees working in the Department of Histopathology, Royal Hallamshire Hospital. The length of time that each had spent in histopathology was recorded. Each observer was given the 100 slides in the order in which they were received in the laboratory. The observer was asked to assign each polyp to the category of hyperplastic or adenomatous after light microscopic examination, but to make no other assessment (such as grade of dysplasia). The observers were asked to carry out the assignment as accurately as possible. The time taken for each observer to make all the categorizations was recorded. The agreement between all the observers and the original diagnosis was assessed using kappa statistics with 95% confidence intervals (6).

RESULTS

The results are summarized in Tables 1 and 2 and Figures 1 and 2.

DISCUSSION

The results of this study show that for this discrimination task with two nominal categories, the level of agreement between histopathologists of 11 months or more training is very high. Although the study could not be performed under conditions that were exactly the same as the routine working environment, the task was presented in a very similar format. The polyps were a consecutive series presented in the same order as they were received in the histopathology laboratory. It is very probable that similar levels of performance occur in routine practice, so clinicians can be very confident in the histopathology diagnosis of colorectal polyps and can implement different follow-up procedures for patients with solely hyperplastic or adenomatous polyps. Studies would be required to confirm these results for other specific areas of histopathology, but it is likely that where there are distinct nominal categories there will be high level of agreement between histopathologists with a corresponding high degree of confidence in the histopathology diagnosis. This contrasts with studies on agreement in allocation to arbitrarily defined categories within a biological spectrum, such as tumor grade or degree of inflammation (14). The average time taken for all observers, except Observer 7, was short (in
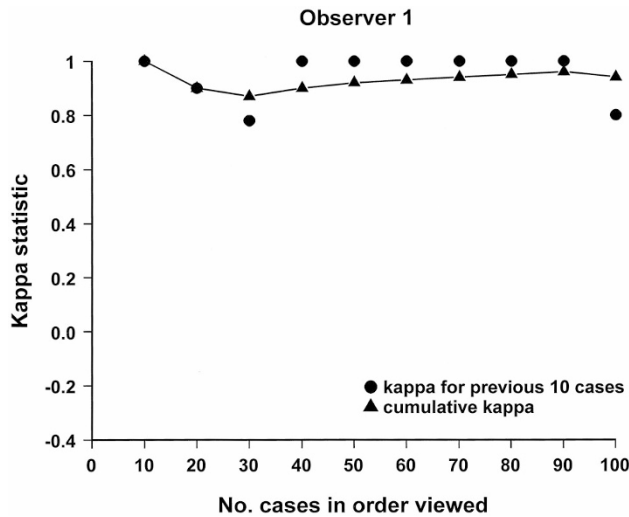
TABLE 1. Kappa Statistics for Each Observer Compared with Each Other Observer and the Original Diagnosis[a]

	Original Diagnosis	Observer 1	Observer 2	Observer 3	Observer 4	Observer 5	Observer 6	Observer 7	Observer 8	Observer 9
Original diagnosis		0.94 (0.87–0.99)	0.88 (0.79–0.97)	0.90 (0.81–0.99)	0.92 (0.84–0.99)	0.86 (0.76–0.96)	0.86 (0.76–0.96)	0.48 (0.31–0.65)	0.90 (0.81–0.99)	0.90 (0.81–0.99)
Observer 1			0.90 (0.81–0.99)	0.92 (0.84–0.99)	0.98 (0.94–0.99)	0.92 (0.84–0.99)	0.88 (0.79–0.97)	0.50 (0.33–0.67)	0.92 (0.84–0.99)	0.96 (0.91–0.99)
Observer 2				0.94 (0.87–0.99)	0.88 (0.79–0.97)	0.90 (0.81–0.99)	0.94 (0.87–0.99)	0.52 (0.35–0.69)	0.90 (0.81–0.99)	0.86 (0.76–0.96)
Observer 3					0.90 (0.81–0.99)	0.88 (0.79–0.97)	0.96 (0.90–0.99)	0.54 (0.37–0.70)	0.92 (0.84–0.99)	0.88 (0.79–0.97)
Observer 4						0.94 (0.87–0.99)	0.86 (0.76–0.96)	0.52 (0.35–0.69)	0.90 (0.81–0.99)	0.98 (0.94–0.99)
Observer 5							0.88 (0.79–0.97)	0.50 (0.33–0.67)	0.88 (0.79–0.97)	0.96 (0.91–0.99)
Observer 6								0.54 (0.37–0.70)	0.88 (0.79–0.97)	0.84 (0.73–0.95)
Observer 7									0.46 (0.28–0.63)	0.50 (0.33–0.67)
Observer 8										0.88 (0.79–0.97)
Observer 9										

[a] 95% confidence intervals in parentheses.

TABLE 2. The Length of Experience in Histopathology of Each Observer and the Average Time Taken to View and Assign Each Case to a Category
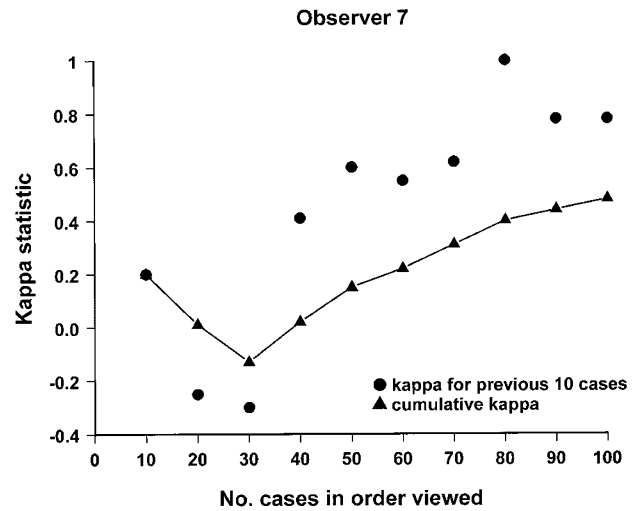
Observer	Months Experience in Histopathology	Average Time Taken per Case (seconds)
1	180	17
2	84	14
3	24	25
4	54	17
5	11	16
6	36	13
7	1.5	122
8	195	13
9	54	13

FIGURE 1. Graph showing the kappa statistics for subgroups of 10 cases and the cumulative kappa statistics for Observer 1.

FIGURE 2. Graph showing the kappa statistics for subgroups of 10 cases and the cumulative kappa statistics for Observer 7.

the range 13 to 25 seconds per case) and includes the time taken to place the slide on the microscope and write down the category. This suggests that the process of category assignment is occurring by a recalled pattern-recognition mechanism rather than a reasoned heuristic (15). The high-level agreement of category assignment at high speed is not surprising given the high level of visio-spatial skills that humans exhibit in many other aspects of life and the volume of the brain that has been evolutionarily allocated to visual interpretation.

The results for the observer with only 6-weeks' experience of histopathology (Observer 7) are interesting. The overall agreement with the other observers and the original diagnosis was only moderate (16) and was significantly lower than all the other results. However, the overall kappa statistic on the 100 cases hides a variation in performance that is revealed by analysis of subsets. Figure 2 shows that the performance of Observer 7 improved significantly as more cases were observed and was approaching the level of the other observers by the end of the observations. This suggests that Observer 7 was learning to discriminate be-

tween the two classes during the observations even without knowing the specific outcome for each case. Observer 7 did know of the existence of hyperplastic and adenomatous polyps before the study and had seen a few examples that had been routine specimens and therefore would have some concept of the entities and would be likely to attach the correct nominal label to easily distinguished examples. It is likely that during the process of making the assignments, Observer 7 adjusted the concepts of the two categories, thus improving performance. The greatly increased time that Observer 7 took to complete the study may reflect this learning process.

The improvement in performance with increasing numbers of cases raises an important issue that should be addressed by all studies that seek to define levels of performance in histopathology (or any other task). Most studies in histopathology are designed to estimate the level of performance that would occur in a routine laboratory situation, but the study itself is not part of routine work and may ask observers to make assignments to categories that they do not routinely use (*e.g.*, when assessing the validity of a new classification or grading system). It is possible that overall kappa statistics from such studies may show low levels of agreement, as the results from Observer 7 did in this study. However, unless the additional analysis of performance with different numbers of observed cases is made, it is not possible to identify situations in which learning is occurring. If this analysis does show that learning is occurring during the course of the study (as illustrated by Observer 7 in Fig. 2), then the results cannot be taken to be the best that will be achieved. In such situations, further training should be undertaken until the performance reaches a steady state, as illustrated by Observer 1 in Figure 1.

Although the level of agreement between experienced histopathologists was high, it was not perfect. The observer with the highest level of agreement with the original diagnosis (Observer 1, kappa statistic 0.94) still disagreed with the original diagnosis on three cases. There is no perfect standard in this study because the distinction between adenomatous and hyperplastic polyps is only made by histopathology examination. However, if the performance of Observer 1 is compared with the original diagnosis, this gives a sensitivity for the diagnosis of adenomatous polyps of 96% (95%; confidence interval, 91 to 99%) and a specificity of 98% (94 to 99%). It could be argued that an apparent false-negative rate of 4% is too high for a condition that has a recognized follow-up protocol with the aim of early detection of colorectal cancer. This rate could be reduced by the observer lowering their internal threshold for the diagnosis of adenomatous polyps, but this would increase the false-positive rate. A higher false-positive result would mean that patients with only hyperplastic polyps would be subject to follow-up procedures (such as colonoscopy) that have a recognized morbidity and small, but definite, mortality rate. The reasons for nonperfect agreement can be understood by analogy with another pattern recognition process: ornithology. Most keen amateur ornithologists would identify common species of birds with perfect accuracy from clear photographs or pictures, but would not always produce a perfect performance in the field. The decrease in performance would be due to imperfect views of the birds due to movement, poor lighting, and obscuring vegetation. In a similar way, the colorectal polyps in this series have factors that make identification difficult, including diathermy artifact, incomplete sampling, and suboptimal planes of sectioning. Another possible confounding factor in this two-class discrimination problem could be serrated adenomas because these may have a morphology that is intermediate between adenomas and hyperplastic polyps. However, these lesions are not as frequent as adenomatous or hyperplastic polyps and none were represented (to the best of our diagnostic expertise) in this study.

REFERENCES

1. Foucar E. Do pathologists play dice? Uncertainty and early histopathological diagnosis of common malignancies Histopathol 1997;31:495–502.
2. Ramsay AD. Errors in histopathology reporting: detection and avoidance. Histopathol 1999;34:481–90.
3. Bethwaite P, Smith N, Delahunt B, Kenwright D. Reproducibility of new classification schemes for the pathology of ductal carcinoma *in situ* of the breast. J Clin Pathol 1998;51:450–4.
4. Farmer ER, Gonin R, Hanna MP. Discordance in the histopathologic diagnosis of melanoma and melanocytic nevi between expert pathologists. Hum Pathol 1996;27:528–31.
5. Silcocks PB. Measuring repeatability and validity of histological diagnosis – a brief review with some practical examples. J Clin Pathol 1983;36:1269–75.
6. Cross SS. Kappa statistics as indicators of quality assurance in histopathology and cytopathology. J Clin Pathol 1996;49:597–9.
7. Tepes B, Ferlan-Marolt V, Jutersek A, Kavcic B, Zaletel-Kragel L. Interobserver agreement in the assessment of gastritis reversibility after *Helicobacter pylori* eradication. Histopathol 1999;34:124–33.
8. Offerhaus GJA, Price AB, Haot J, Ten Kate FJW, Sipponen P, Fiocca R, *et al.* Observer agreement on the grading of gastric atrophy. Histopathol 1999;34:320–5.
9. Brown LJR, Smeeton NC, Dixon MF. Assessment of dysplasia in colorectal adenomas: an observer variation and morphometric study. J Clin Pathol 1985;38:174–9.
10. Lessells AM, Burnett RA, Howatson SR, Lang S, Lee FD, McLaren KM, *et al.* Observer variability in the histopathological reporting of needle biopsy specimens of the prostate. Hum Pathol 1997;28:646–9.
11. Anonymous. Pathology as art appreciation: melanoma diagnosis. The bandolier. 1999. Available at: http://www.jr2.ox.ac.uk/bandolier/band37/b37-2.html.
12. Matsumoto G, Mizuno M, Shimizu M, Manabe T, Iida M. Clinicopathological features of serrated adenoma of the colorectum: comparison with traditional adenoma. J Clin Pathol 1999;52:513–6.
13. Iino H, Jass JR, Simms LA, Young J, Leggett B, Ajioka Y, *et al.* DNA microsatellite instability in hyperplastic polyps, serrated adenomas, and mixed polyps: a mild mutator pathway for colorectal cancer? J Clin Pathol 1999;52:5–9.
14. Cross SS. Grading and scoring in histopathology. Histopathol 1998;33:99–106.
15. Underwood JCE. Introduction to biopsy interpretation and surgical pathology. 2nd ed. London: Springer-Verlag; 1987.
16. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.