Letters to the Editor

Observer variation in histological diagnosis

I was most interested in the recent articles on observer variation in histological diagnosis,[1 2] a subject which has fascinated me for several years. Perhaps the kappa statistic, which is both versatile and easy to conceive, will now become a standard measure of reproducibility in histopathology. It is also refreshing to see that Landis and Koch's "benchmarks" have been adopted for evaluating levels of agreement,[2] a far more realistic procedure than merely testing for statistical significance as Cohen himself observed.[3] I would like to point out however that the denominator in the formula given by Stenkvist for the standard error of kappa is incorrect and should be $N(1-p_c)^2$.

In particular Thomas' paper is of value in tackling the problem of bias in diagnosis (bias being a systematic difference in results not attributable to random error). I wonder, though, whether the idea behind the CEN coefficient could be approached more simply by finding kappa for the central category *v* all others?[4] Also the CEN coefficient would be inapplicable if more than three categories were distinguished. In any case neither the CEN nor the OPT coefficient would apply to truly nominal data whereas finding residual for each cell[5] would give a measure of bias applicable under any circumstances. Another minor criticism is that no comparison is given between levels of reproducibility before and after familiarisation of the observers with the grading criteria, although since the results showed poor reproducibility even after training, this may have been thought superfluous.

Oddly enough the results in Grinnell's own paper[6] suggest that grading of large bowel carcinoma as then practised was in any case not worthwhile: using the published data I have performed an approximate analysis of variance on the proportions of patients surviving to five years in each combination of grade and stage (assuming that there was no imbalance of other prognostic factors between the different grade/stage combinations). There is a highly significant ($p < 0.001$) main effect for stage but no significant effect ($p > 0.05$) for grade. On the other hand a similar analysis of Bloom's results for breast cancer[7] indicates highly significant ($p < 0.001$) main effects for both stage and grade. Given that both grading systems have low reproducibility these results, whilst not conclusive, do suggest that the lack of association between grade and five year survival in large bowel cancer is not simply the result of poor reproducibility but that grading of this tumour is indeed grossly invalid, lacking both reproducibility and predictive value.

The value of grading depends on the kind of tumour, the degree of reproducibility and how well it independently predicts outcome. However even in the simplest of schemes there are traps for the unwary[8 9] and as Stenkvist has demonstrated[10] grade (at least in breast cancer) combines different variables not necessarily in the most rational way, whilst in gastric carcinoma grade shows no significant association with aneuploidy.[11] Is it now time to abandon the concept of grade altogether to replace it by defined "histological features affecting prognosis" specific to each tumour?

PAUL SILCOCKS
MRC Environmental Epidemiology Unit
University of Southampton,
Southampton General Hospital
Southampton SO9 4XY

References

[1] Thomas GDH, Dixon MF, Smeeton NC, Williams NS. Observer variation in the histological grading of rectal carcinoma. *J Clin Pathol* 1983;**36**:385–91.
[2] Stenkvist B, Bengtsson E, Eriksson O, Jarkrans T, Nordin B, Westman-Naeser S. Histopathological system of breast cancer classification: reproducibility and clinical significance. *J Clin Pathol* 1983;**36**:392–8.
[3] Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Measurement* 1960;**20**:37–46.
[4] Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York: John Wiley & Sons, 1981.
[5] Bishop YMM, Fienberg SE, Holland PW. *Discrete multivariate analysis: theory and practice.* MIT Press, 1975.
[6] Grinnell RS. The grading and prognosis of carcinoma of the colon and rectum. *Ann Surg* 1939;**109**:500–33.
[7] Bloom HJG. Prognosis in carcinoma of the breast. *Br J Cancer* 1950;**4**:259–88.
[8] Ellis PSJ, Whitehead R. Mitosis counting, a need for reappraisal. *Hum Pathol* 1981;**12**:2–3.
[9] Graem N, Helweg-Larsen K. Mitotic activity and delay infixation of tumour tissue. *Acta Pathol Microbiol Scand* [*A*] 1979;**87**:375–8.
[10] Stenkvist B, Westman-Naeser S, Vegelius J *et al.* Analysis of the reproducibility of subjective grading systems for breast carcinoma. *J Clin Pathol* 1979;**32**:979–85.
[11] Inokuchi K, Kodama Y, Sasaki O, Kamegawa T, Okamura T. Differentiation of growth patterns of early gastric carcinoma determined by cytophotometric DNA analysis. *Cancer* 1983;**51**:1138–41.

Dr Dixon comments on behalf of his colleagues:

We thank Dr Silcocks for his comments on our attempts to define bias in our observer variation study of rectal cancer[1] and would like to make the following points.

Firstly the CEN coefficient is not directly related to the kappa coefficient for central category *v* all others and we would argue that the calculation of this coefficient represents a simpler approach than a comparison of kappa values.

Another important point is that in using Cohen's kappa statistic we assumed that different types of disagreement carry equal weighting. Disagreements of the type well *v* poor could be considered more serious since classifications are not from adjacent categories. These disagreements should, perhaps, be given a higher weight. In practice, however, less than 1% of all pairings in our study were of this type so that any difference would be negligible. Although OPT and CEN can be generalised to the case of more than three categories, having more categories would increase the proportion of disagreeing pairs with categories not adjacent. Some form of weighting would then need to be introduced and the analysis would be more complicated (see Cohen, 1968).[2] Furthermore, the weighting used would have to be subjective. How much worse is one form of disagreement than another?

Since we did not achieve outstanding agreement with three categories there seems little point in increasing the number of categories. Certainly in histological practice few observers would attempt (or even want) to increase the number of tumour grades. The argument would also apply to most other histological abnormalities which are conventionally graded as mild, moderate, and severe. With three categories an optimal situation is reached. Although the data are effectively nominal, OPT and CEN may be used. We do agree, however, that OPT and CEN may not be used with nominal data in general since it is not ordered.

M F DIXON
Department of Pathology,
University of Leeds,
Leeds LS2 9JT

References

[1] Thomas GDH, Dixon MF, Smeeton NC, Williams NS. Observer variation in the histological grading of rectal carcinoma. *J Clin Pathol* 1983;**36**:385–91.

[2] Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968;**70**:213–20.

Dr Stenkvist replies as follows:

In relation to Dr Paul Silcocks comments on observer variation I wish to make the following comment:

The breast cancer classification systems we analysed not only have a poor reproducibility but they also lack clinical value in that they do not take into account rate of recurrence. However, parameters whose measurement is easily reproducible, such as "tumour markers" (CEA etc), DNA distribution among tumour cell nuclei, measures of tumour size, etc, although clearly correlated with prognosis when you compare *groups* of patients, are of little or no value for *prognosis in the individual patient* unless they are combined in a reliable way into a malignancy index that can be obtained by step-wise logistic regression analysis. We have recently addressed this problem,[1] and we think it is time that pathologists and clinicians start to look for optimal combinations of parameters to create "malignancy indices" for tumour diseases in order to have practicable methods in daily clinical work.

Breast cancer patients deserve the greatest possible assurance in estimates of the severity of their disease and in the selection of their treatment, just as the clinician should receive the most complete information possible about his patient's condition. Our study indicates that it will be possible to give the clinician such complete information (superior to so-called clinical staging), provided that the significant variables are recorded in a meticulous way and combined into a risk curve. Patients with low malignancy grade as demonstrated in our study could then be confidently reassured that they will not suffer from recurrent disease, and resources for adjuvant therapy and follow-up can be saved for those patients who really need them (ie patients with a high malignancy grade).

Our study further emphasises that there is no evidence that THE factor, a single entity of absolute prognosis, exists or is likely to exist.

We regret our mistake in proof-reading the standard error of kappa of Dr Jacob Cohen's formula. The computation, however, was performed using the correct formula.

BJÖRN STENKVIST

Department of Clinical Cytology, University Hospital, S-751 85 Uppsala, Sweden

Reference

[1] Stenkvist B, Bengtsson E, Dahlqvist B, *et al.* Predicting breast cancer recurrence. *Cancer* 1982;**50**:2884–93.

Antinuclear antibody-negative systemic lupus erythematosus—how common?

We read with interest the paper in the October 1982 issue of the Journal.[1] We feel that two factors require further consideration before accepting that 8·9% of patients with SLE may be ANA negative. Firstly, the authors did not test their ANA-negative sera to see whether they would react with human substrates. It is well recognised that a small number of such sera will react only with human tissues such as peripheral granulocytes.

Secondly, our own experience over several years has led us to become increasingly concerned about results obtained with radioimmunoassay kits supplied by RC Amersham. We have found positive anti-DNA results at levels of 30 U/ml or greater, not uncommonly, in sera of patients who have no reliable clinical evidence of SLE as determined by the American Rheumatism Association criteria. Reasons for this may be several and include DNA reactions with basic proteins, binding of C1q to DNA or to low density lipoproteins as well as the possibility that some preparations may contain some single stranded DNA. Binding to basic proteins is not uncommon in patients with malignant disease and may be reversed with sodium dodecyl sulphate. Resulting from our dissatisfaction with RIA kits we compared RIA results with those obtained by immunofluorescence using either *Crithidia luciliae* or human metaphase chromosomes as substrate. Our findings indicate that both immunofluorescent procedures gave a far better correlation with clinical manifestations of SLE and we have now standardised our test procedures on *C luciliae* as substrate.

KC WATSON

EJC KERR

Central Microbiological Laboratories, Western General Hospital, Crewe Road, Edinburgh EH4 2XU

Reference

[1] McHardy K, Horne CHW, Rennie J. Antinuclear antibody-negative systemic lupus erythematosus—how common? *J Clin Pathol* 1983;**35**:1118–21.

Professor Horne and Dr McHardy reply as follows:

Although we did not screen the ANA-negative sera against human tissues we agree with Watson and Kerr that a small number of sera will react with such tissues. However, we must point out that we never claimed that the radioimmunoassay kits supplied by RC Amersham did not give false positive results. They most certainly do. It was for this reason that we only studied sera where a value of 40 U/ml or greater had been obtained. The main point of our argument in this paper is simply that if clinicians rely solely on the conventional immunofluorescence screening procedure for antibodies to nuclear constituents they are likely to be misled if they hold the traditional view that ANA are present in over 98% of SLE cases.

Finally, we accept that an immunofluorescence screening test based on *Crithidia luciliae* is of diagnostic value and we are currently comparing it with the RIA kits.

CHW HORNE

K McHARDY

Department of Pathology, University Medical Buildings, Foresterhill, Aberdeen AB9 2ZD

IgA pyroglobulinaemia in lymphoma

Pyroglobulins are abnormal immunoglobulins, usually IgG or IgM, which when heated to 56°C form a gel irreversible by changes of temperature, pH or dilution. They were first recognised in 1953,[2] and since then have been reported in a variety of conditions, mainly multiple myeloma, Waldenström's macroglobulinaemia,[4] and lymphoma.[5] We would like to report a patient with an unusual pyroglobulin. The patient had abnormal bleeding associated with defective ristocetin-induced platelet aggregation. This appeared to be related to the presence of the abnormal globulin. When the patient's serum was fractionated to obtain the abnormal pyroglobulin and when this fraction was added to normal plasma it produced a similar defect in platelet aggregation.

Case report

The patient was a 72-year-old woman. Six weeks prior to admission, she complained of arthralgia, involving both ankles, wrists and knees. She was admitted after the development of a purpuric rash.

On examination, she had a marked purpuric rash over both legs. There was no lymphadenopathy and no hepatomegaly but her spleen was palpable 4–5 cm below