Observer accuracy in estimating proportions in images: implications for the semiquantitative assessment of staining reactions and a proposal for a new system

S S Cross

Abstract

***Background*—The results of immunohistochemical staining are often assessed by semiquantitative scoring. However, these scoring systems are usually non-standardised and there has been little evaluation of the accuracy and reliability of this subjective assessment.**

***Aims*—To assess the accuracy of observer estimation of proportions of objects in an image.**

***Methods*—Images were generated that contained known proportions of pink squares in grids of 50×50 and 100×100 squares. Observers were shown each image for five seconds in random order and either estimated the proportion of pink squares or selected the image (from a pair of images) that contained the greater proportion of pink squares. The observers were four consultant histopathologists, seven trainee histopathologists, and six control non-histopathologists.**

***Results*—The raw estimations of proportions showed a close correlation with the real proportions, with correlation coefficients of 0.94 and 0.95 for consultant and trainee histopathologists on the 50×50 grids. However, the performance in the comparison task was much higher, with an almost perfect classification for grids of equal size even when the proportions only differed by 5%.**

***Conclusions*—Histopathologists can estimate proportions of objects in an image with a reasonable degree of accuracy in this abstract test system. All observers, whether histopathologists or not, can discriminate between proportions that are only 5% different in equal sized image grids. This suggests that the generation and use of carefully calibrated reference images could greatly improve the accuracy and reliability of semiquantitative scoring of immunohistochemical or any other staining.**

(*J Clin Pathol* 2001;**54**:385–390)

Keywords: scoring; immunohistochemistry; semiquantitative; interobserver agreement; κ statistics

Section of Oncology and Pathology, Division of Genomic Medicine, University of Sheffield Medical School, Beech Hill Road, Sheffield S10 2UL, South Yorkshire, UK
S S Cross

Correspondence to:
Dr Cross
s.s.cross@sheffield.ac.uk

Accepted for publication
27 October 2000

In many experimental histopathology studies, and some routine laboratory tests,[1 2] an assessment of the proportion of cells stained by a reaction is made. This staining reaction is most commonly produced by immunohistochemistry,[3] but can also be generated by in situ hybridisation[4 5] or non-immunological histochemical staining. There is no standard method of assessing the proportion of stained cells, and the methodology described in publications ranges from rigorous quantitation using computerised image analysis systems[6 7] to undefined subjective categories.[8 9] One of the most common methods is semiquantitative estimation by human observers.[10] In such systems, the proportion of cells that are stained is divided into approximately four arbitrary categories (for example, 0–25%, 26–50%, 51–75%, and 76–100%) and specimens are assigned to a category by an observer.[11–13] There is little standardisation in such schemes, with varying numbers of categories with different boundaries (for example, 0%, 1–10%, 11–50%, 51–100%;[14] < 5%, 5–75%, > 75%,[15] or 0%, < 10%, 11–40%, 41–70%, > 70%[16]). Very few of the published studies[17] that use such semiquantitative systems make any measurement of the interobserver and intraobserver reproducibility of the scoring, so it is often difficult to assess the validity of the results. If the staining assessed in such studies is going to generate information that will be used in the selection of treatment for patients,[1 2 18–20] then better quality control is required.

The reproducibility and accuracy of the semiquantitative scoring of staining reactions is largely unknown because of a lack of published studies. It has been shown that pathologists have high levels of agreement when assigning images to discrete nominal categories (for example, hyperplastic and adenomatous polyps),[21] but far lower levels of agreement when assigning images to categories with arbitrary boundaries within a continuum.[11 22]

The assessment of accuracy and reproducibility in immunohistochemical studies is hampered by the lack of a gold standard for the proportion of cells that should be expressing an antigen and by several peripheral factors, such as selection of the field of view, area of the field of view on different microscopes, and different conditions of illumination. This study has been designed to investigate the accuracy of different observers in estimating the proportions of different elements in images. It uses abstract images that have been generated to give known proportions of differently coloured elements so that accuracy can be measured easily. A new system of image comparison is also tested to investigate whether this could provide better

accuracy and reproducibility than existing semiquantitative systems.

Methods and materials

The observers were presented grids of pink and blue squares on a black background (fig 1) in a presentation program (Powerpoint; Microsoft, Seattle, USA) running on a standard microcomputer. Each image was presented for six seconds. There was a blank black image between each test image and this remained on the screen until the observer had given their response. The observers' responses were recorded by the study coordinator (SSC). There were two example images at the start of each set of test images. The grids of pink and blue squares were generated using pseudorandom numbers in a custom written program (in the Matlab programing language; MathWorks, Natick, USA), which allowed the user to select the proportion of pink or blue squares. The program checked that the final proportions of squares were within 1% of the proportion specified by the user. The seed for the pseudorandom number generator was reset between each image using numbers from random number tables.

Each observer was exposed to three sets of test images. The first set was of grids of 50×50 and 100×100 squares with the proportion of pink squares set at 5% intervals between 0% and 100% for both grid sizes (fig 1A), producing a total of 42 images. The images were presented in a randomised order and the observer was asked to estimate the proportion of pink squares in each image to the nearest whole per cent. The second set was of two grids on each test image set side by side. The grids were either both 50×50 or 100×100. For each grid size there were pairs of grids with a proportion of pink squares of 25%, 50%, or 75% in one and 5%, 10%, 15%, and 20% above and below those percentages in the other grid (fig 1B), producing a total of 48 images. The grids in each pair were randomly assigned easily to the left or right side of the screen; laterality was explicitly indicated on the screen by the letters. The images were presented in a randomised order and the observer was asked to indicate which grid (left or right) contained the greater proportion of pink squares. The observers were told that no pair contained an equal proportion. The third set was also composed of pairs of grids but with a 50×50 and a 100×100 size grid on each image. For each grid size there were pairs with a 50% proportion of pink squares in one and 5%, 10%, 15%, and 20% above and below that percentage in the differently sized grid (fig 1C), producing a total of 16 images. The grids in each pair were assigned randomly to the left or right side of the screen; laterality was explicitly indicated on the screen. The images were presented in a randomised order and the observer was asked to indicate which grid (left or right) contained the greater proportion of pink squares. The observers were told that no pair contained an equal proportion.

The observers consisted of four consultant histopathologists, seven trainee histopathologists, and six controls who were naïve to structured observational tasks.

The least squares method of regression was used to plot a regression line and to derive a correlation coefficient for the relation between the estimated and real proportions of squares. The degree of interobserver agreement was assessed using κ statistics and 95% confidence intervals.[23]

Results

The results are summarised in figs 2 and 3 and tables 1–7.

Discussion

The test system used has many differences to light microscopic estimation of the proportion of cells that is positive by immunostaining. The test images are presented to the observer without the need for selection of a field of view. The test images have a sharp contrast between the pink and blue elements within them and there is no gradation between the elements or within them. There are no other elements within the test images that could confuse the observer. It is probable that all these differences would tend to improve observer performance on the test images relative to light microscopic images of immunostained preparations. Therefore, these results probably represent the upper limits of possible performance in this task.

The correlation coefficients for the real and estimated proportions are all relatively high, although those for trainee and consultant pathologists were higher than for the controls. The scatter graphs show that the correlation is much better at the low and high percentages and that there is a wide spread of estimates in the 40–60% range. The correlation is better for all groups of observers for the 50×50 grids than for the 100×100 grids. In the 50×50 grids the squares were clearly visible as individual objects at the viewing resolution (fig 1A) but in the 100×100 grids individual squares were more difficult to discern (fig 1B). It may be that in the 100×100 grids there was some blurring of perception between differently coloured squares that made estimation of the proportions less accurate. These results suggest that raw estimates, by relatively experienced observers (such as histopathologists), of a proportion of positive cells could be used in research studies as long as repeated observations were made to eliminate occasional aberrant estimations. However, as mentioned this study probably tests for the best level of performance that can be attained, and images with more distracting elements and less discrete object boundaries might be estimated less reproducibly. Nevertheless, a raw estimate by human observers could be a valid research method that would obviate the need for complicated, expensive, and time consuming digital image analysis.[6–7]

Although the correlation between real and raw estimates of proportions are relatively close, performance was much better when two images were compared simultaneously. There

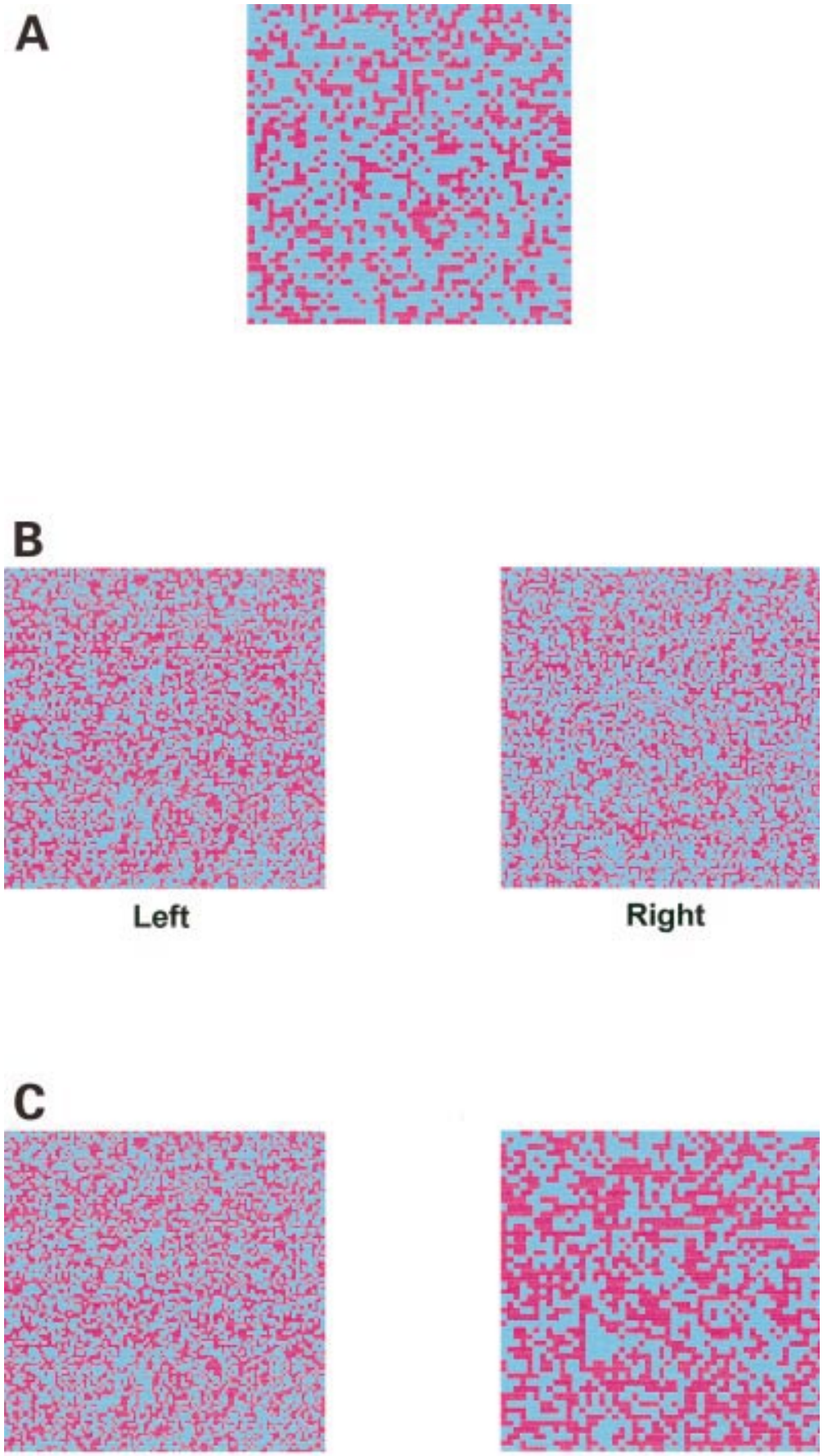
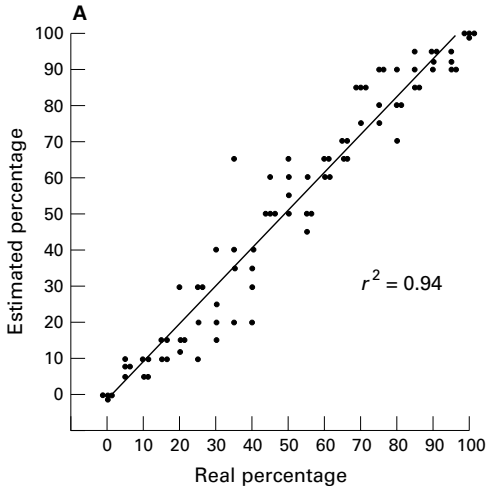
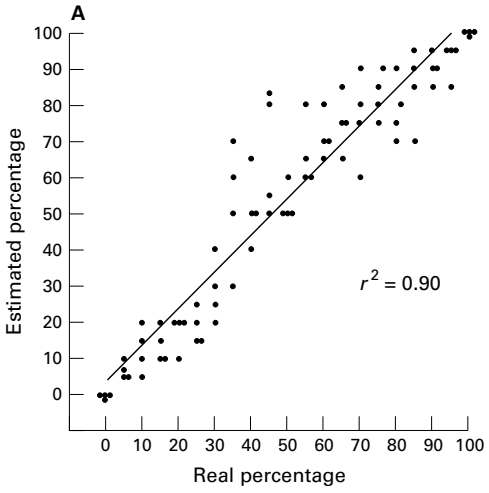
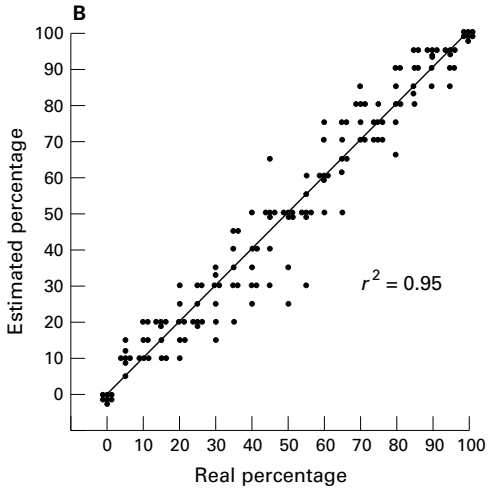
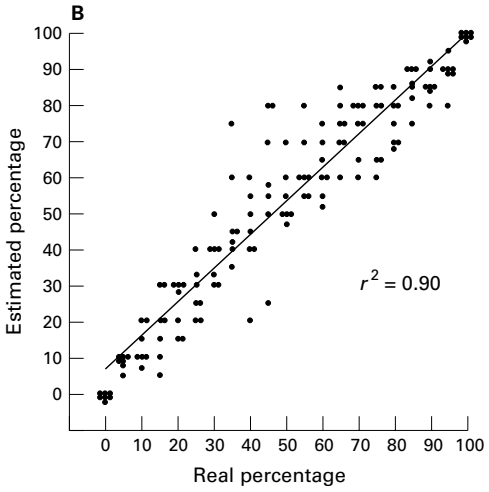


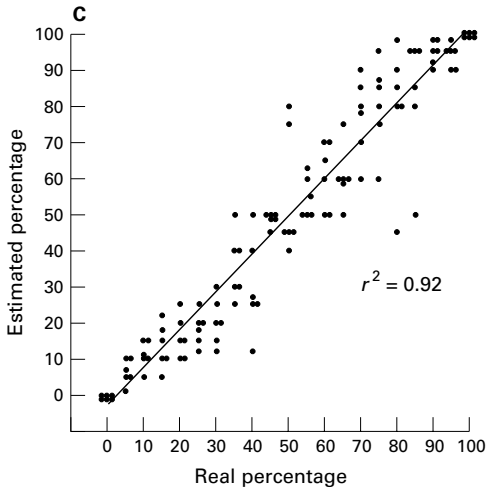
Figure 1 Examples of the test images. (A) A 50 × 50 grid in which 40% of the squares are coloured pink. (B) Two 100 × 100 grids in a comparison image; the left hand grid has 50% pink squares, the right hand grid has 45%. Despite the small difference in the proportion of pink squares between the two grids all the observers correctly selected the left hand grid as the one with the greatest proportion. (C) Two grids in a comparison image. The left hand 100 × 100 grid has 50% pink squares, the right hand 50 × 50 grid has 55% pink squares. Most of the errors made in the entire study related to this image, although 74% of observers still correctly identified the right hand grid as containing the greatest proportion.











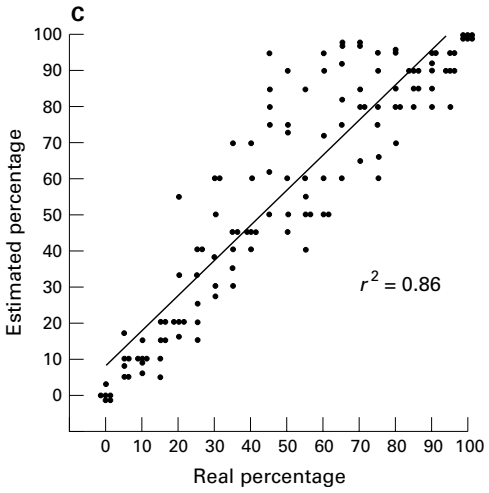


Figure 2 Scattergrams for the estimated proportion of pink squares in a 50 × 50 grid for (A) the consultant histopathologists, (B) the trainee histopathologists, and (C) controls. The regression line and correlation coefficient were derived using the least squares method of regression.

Figure 3 Scattergrams for the estimated proportion of pink squares in a 100 × 100 grid for (A) the consultant histopathologists, (B) the trainee histopathologists, and (C) controls. The regression line and correlation coefficient were derived using the least squares method of regression. It can be seen that the points have a greater scatter around the regression line than for the 50 × 50 grids (fig 2). This scatter is greater in the control group than in the consultant or trainee histopathologist groups and there is a tendency towards overestimation of the proportion.

was only one incorrect response out of 816 comparisons of equal sized grids by all observers. In 204 of these comparisons the two images only differed by 5%. This improvement in accuracy when a visual comparator is present is shown in the κ statistics for a categorical system of percentage estimates by histopathologists (tables 5–7). If the two categories around 50% (from the widely used four category system) are analysed there is a significant difference between the agreement with raw estimates (0.66) and those obtained

Table 1 Results of comparisons between two 100 × 100 square size grids for all observers

Reference (%)	Comparison (%)	No. correct	No. incorrect
25	5	17	0
25	10	17	0
25	15	17	0
25	20	17	0
25	30	17	0
25	35	17	0
25	40	17	0
25	45	17	0
50	30	17	0
50	35	17	0
50	40	17	0
50	45	17	0
50	55	17	0
50	60	17	0
50	65	17	0
50	70	17	0
75	55	17	0
75	60	17	0
75	65	17	0
75	70	17	0
75	80	16	1
75	85	17	0
75	90	17	0
75	95	17	0

Table 2 Results of comparisons between two 50 × 50 square size grids for all observers

Reference (%)	Comparison (%)	No. correct	No. incorrect
25	5	17	0
25	10	17	0
25	15	17	0
25	20	17	0
25	30	17	0
25	35	17	0
25	40	17	0
25	45	17	0
50	30	17	0
50	35	17	0
50	40	17	0
50	45	17	0
50	55	17	0
50	60	17	0
50	65	17	0
50	70	17	0
75	55	17	0
75	60	17	0
75	65	17	0
75	70	17	0
75	80	17	0
75	85	17	0
75	90	17	0
75	95	17	0

Table 3 Results for the comparisons between one grid of 100 × 100 square size and the other of 50 × 50 square size for all observers

Reference (%)	Comparison (%)	No. correct	No. incorrect
50	30	34	0
50	35	33	1
50	40	32	2
50	45	28	6
50	55	25	9
50	60	34	0
50	65	34	0
50	70	33	1

with equally sized comparators (1.00) and even unequally sized comparators (0.96). This suggests that a much more accurate and reproducible method of estimating the proportion of positive cells would be to construct a series of reference images of varying proportions of positive cells (spread evenly from 0% to 100%) and to ask observers to compare the new experimental images to these references.[24–26] The proportions of positive cells in the reference images could be determined by robust objective methods that might include

Table 4 Agreement between four class categorisation of the real proportion of pink squares in the grid and that estimated by all histopathologists (consultants and trainees) for the 100 × 100 square grid

	Real (%)			
	0–25%	26–50%	51–75%	76–100%
Estimated (%)				
0–25%	57	4	0	0
26–50%	9	32	0	0
51–75%	0	19	37	8
76–100%	0	0	18	47

κ Statistic 0.66 (95% confidence interval, 0.59 to 0.74).

Table 5 Agreement between two class categorisation of real proportion (between 26% and 75%) of pink squares in the grid and that estimated by all histopathologists (consultants and trainees) for the 100 × 100 square grid

	Real (%)	
	26–50%	51–75%
Estimated (%)		
26–50%	32	0
51–75%	19	37

κ Statistic 0.60 (95% confidence interval, 0.45 to 0.76).

Table 6 Agreement between two class categorisation of real proportion (between 26% and 75%) of pink squares in the grid and that obtained by comparison of two equal sized grids by all histopathologists (consultants and trainees) for the 100 × 100 square grid

	Real (%)	
	26–50%	51–75%
Estimated (%)		
26–50%	44	0
51–75%	0	44

κ Statistic 1.00.

Table 7 Agreement between two class categorisation of real proportion (between 26% and 75%) of pink squares in the grid and that obtained by comparison of two unequal sized grids by all histopathologists (consultants and trainees)

	Real (%)	
	26–50%	51–75%
Estimated (%)		
26–50%	166	2
51–75%	8	174

κ Statistic 0.94 (95% confidence interval, 0.91 to 0.98).

digital image analysis. Although the level of performance in a real experimental environment might not be as high as in this study, it is still likely that proportions of positive cells could be estimated to within 10% with a high degree of certainty. The performance of the controls on comparison of equal sized grids was almost perfect and suggests that this methodology would be useful when relatively inexperienced observers, such as postgraduate students, are making the estimates.

These results suggest that a robust and accurate method for estimating the proportion of cells that are positive for some staining reaction would be visual comparison of the test image with a series of reference images of known proportions. Future studies of this methodology should be based on real microscopic images with reference images in which the proportions have been measured by a well validated quantitative method. The findings of this study are relevant to any situation in which visual

proportion is required and can be applied to areas outside histopathology, such as the design of human computer interfaces.[27]

The author would like to thank all the observers who participated in this study.

1 Barnes DM, Millis RR, Beex LM, *et al*. Increased use of immunohistochemistry for oestrogen receptor measurement in mammary carcinoma: the need for quality assurance. *Eur J Cancer* 1998;**34**:1677–82.
2 Mohsin SK, Allred DC. Immunohistochemical biomarkers in breast cancer. *Journal of Histotechnology* 1999;**22**:249–61.
3 McNicol AM, Richmond JA. Optimizing immunohistochemistry: antigen retrieval and signal amplification. *Histopathology* 1998;**32**:97–103.
4 Seitzer U, Gerdes J. Measuring immune responses in situ: immunohistology and in situ hybridization. *Methods in Microbiology* 1998;**25**:651–62.
5 Jimenez RE, Wallis T, Tabasczka P, *et al*. Determination of Her-2/neu status in breast carcinoma: comparative analysis of immunohistochemistry and fluorescent in situ hybridization. *Mod Pathol* 2000;**13**:37–45.
6 Matkowskyj KA, Schonfeld D, Benya RV. Quantitative immunohistochemistry by measuring cumulative signal strength using commercially available software Photoshop and Matlab. *J Histochem Cytochem* 2000;**48**:303–11.
7 Lohmann CM, League AA, Clark WS, *et al*. Bcl-2 : Bax and bcl-2 : Bcl-x ratios by image cytometric quantitation of immunohistochemical expression in ovarian carcinoma: correlation with prognosis. *Cytometry* 2000;**42**:61–6.
8 Ordi J, Schammel DP, Rasekh L, *et al*. Sertoliform endometrioid carcinomas of the ovary: a clinicopathologic and immunohistochemical study of 13 cases. *Mod Pathol* 1999; **12**:933–40.
9 Bates AW, Baithun SI. Secondary neoplasms of the bladder are histological mimics of nontransitional cell primary tumours: clinicopathological and histological features of 282 cases. *Histopathology* 2000;**36**:32–40.
10 McCluggage WG, Maxwell P, Hamilton PW, *et al*. High metallothionen expression is associated with features predictive of aggressive behaviour in endometrial carcinoma. *Histopathology* 1999;**34**:51–5.
11 Cross SS. Grading and scoring in histopathology. *Histopathology* 1998;**33**:99–106.
12 Wang S, Saboorian MH, Frenkel E, *et al*. Laboratory assessment of the status of Her-2/neu protein and oncogene in breast cancer specimens: comparison of immunohistochemistry assay with fluorescence in situ hybridisation assays. *J Clin Pathol* 2000;**53**:374–81.
13 van Doorn HC, Burger CW, van der Valk P, *et al*. Oestrogen, progesterone, and androgen receptors in ovarian neoplasia: correlation between immunohistochemical and biochemical receptor analyses. *J Clin Pathol* 2000;**53**:201–5.
14 Kaufmann O, Dietel M. Thyroid transcription factor-1 is the superior immunohistochemical marker for pulmonary adenocarcinomas and large cell carcinomas compared to surfactant proteins A and B. *Histopathology* 2000;**36**:8–16.
15 Bankfalvi A, Terpe H-J, Breukelmann D, *et al*. Immunophenotypic and prognostic analysis of E-cadherin and b-catenin expression during breast carcinogenesis and tumour progression: a comparative study with CD44. *Histopathology* 1999;**34**:25–34.
16 Pernick NL, DaSilva M, Gangi MD, *et al*. "Histiocytic markers" in melanoma. *Mod Pathol* 1999;**12**:1072–7.
17 Bonatz G, Lüttes J, Hamann S, *et al*. Immunohistochemical assessment of p170 provides prognostic information in endometrial carcinoma. *Histopathology* 1999;**34**:43–50.
18 Hall PA, Going JJ. Predicting the future: a critical appraisal of cancer prognosis studies. *Histopathology* 1999;**35**:489–94.
19 Rhodes A, Jasani B, Barnes DM, *et al*. Reliability of immunohistochemical demonstration of oestrogen receptors in routine practice: interlaboratory variance in the sensitivity of detection and evaluation of scoring systems. *J Clin Pathol* 2000;**53**:125–30.
20 Rhodes A, Miller KD, Jasani B, *et al*. Immunohistochemical demonstration of oestrogen and progesterone receptors: correlation of standards achieved on in house tumours with that achieved on external quality assessment material in over 150 laboratories from 26 countries. *J Clin Pathol* 2000;**53**:292–301.
21 Cross SS, Betmouni S, Burton JL, *et al*. What levels of agreement can be expected between histopathologists assigning cases to discrete nominal categories? A study of the diagnosis of hyperplastic and adenomatous polyps. *Mod Pathol* 2000;**13**:941–4.
22 Ramsay AD. Errors in histopathology reporting: detection and avoidance. *Histopathology* 1999;**34**:481–90.
23 Cross SS. Kappa statistics as indicators of quality assurance in histopathology and cytopathology. *J Clin Pathol* 1996;**49**: 597–9.
24 Tufte ER. *The visual display of quantitative information*. Cheshire, Connecticut: Graphics Press, 1983.
25 Tufte ER. *Envisioning information*. Cheshire, Connecticut: Graphics Press, 1990.
26 Tufte ER. *Visual explanations*. Cheshire, Connecticut: Graphics Press, 1997.
27 Walker AJ, Cross SS, Harrison RF. Visualisation of biomedical datasets by use of growing cell structure networks: a novel diagnostic classification technique. *Lancet* 1999;**354**:1518–21.