J Clin Pathol 1983;**36**:1269–1275

Measuring repeatability and validity of histological diagnosis—a brief review with some practical examples

PBS SILCOCKS

From the MRC Environmental Epidemiology Unit, Southampton General Hospital, Southampton SO9 4XY

SUMMARY Evaluation of histological diagnosis requires an index of agreement (to measure repeatability and validity) together with a method of assessing bias. Cohen's kappa statistic appears to be the most suitable tool for measuring levels of agreement, which if unsatisfactory may be caused by bias. Further study of bias is possible by examining levels of agreement for each diagnostic category or by searching for categories of disagreement in which more observations occur than would be expected by chance alone.

This article gives reasons for choosing the kappa statistic, with examples illustrating its calculation and the investigation of bias.

The intention of this article is to encourage wider use of the kappa statistic as a method for evaluating histological diagnosis which is feasible even under routine conditions and will be useful to those interested in epidemiological or quality control aspects of histopathology. Examples might be a pathologist wishing to assess his consistency in grading large bowel cancer by reviewing a sample of past cases, or wishing to compare his gradings with those of a colleague. Although the method suggested here has been applied before, it seems timely to present an account which emphasises the ease of the basic calculations.

EVALUATION OF DIAGNOSIS

Histological diagnosis is fundamental in the study of many diseases and of tumours in particular, providing a yardstick against which many other tests are evaluated. This does not preclude evaluation of histological diagnosis itself however, and this may be done by examining repeatability, validity and bias.

Repeatability is the level of agreement between replicate measurements. Within-observer repeatability is measured by allowing one observer to examine a specimen on two or more occasions, whilst between-observer repeatability is measured when two or more observers examine the same specimen. If the same observer examines two or more specimens taken from the same subject then differences in diagnosis will reflect both within-observer repeatability and the repeatability of specimen collection (partly sampling error and partly biological variation). Repeatability in this sense is distinct from group repeatability in which the overall proportions of individuals said to be diseased are compared between observers. Group repeatability may be high despite poor agreement on individual cases and so in spite of its uses[1 2] it need not further concern us.

Validity is the extent to which the measurements reflect the truth. In histopathology this question may be divided into the validation of an individual's diagnosis against that of an expert and the validation of the concepts on which the expert's opinion is based. The latter is beyond the scope of this article except to observe that since the classification of tumours for example is based in part on theoretical considerations of histogenesis, partly on histological appearance and partly on likely behaviour[3] validation of these concepts will be performed in different ways: Experimental studies tested the validity of the APUD concept,[4] numerical taxonomy was used to study oral leukoplakia and carcinoma[5] and correlation with survival tested the validity of tumour grading.[6]

Bias is a systematic difference or error in measurement, not accounted for by chance variation. For example an observer might consistently grade carcinomas less severely than a colleague.

Accepted for publication 28 June 1983

Such a process of evaluation is similar to quality control. The concept of quality control as applied to histopathology has been criticised[7] but nevertheless various studies have evaluated both technical and diagnostic aspects of the subject, using methods of varying general applicability and complexity.[7–11] Whatever method is used it should be simple to interpret, easy to apply to histopathological diagnoses and statistically testable.

POSSIBLE APPROACHES

Some examples of studies on repeatability are those of Cocker,[13] Sissons,[14] and Feinstein,[15] which are all based on assigning scores either to disease categories or to the degree to which certain histological features are present. These methods can be criticised because the scores, though discrete, are treated as continuous variables, consequently the results are presented either as a mean score or as a mean difference in score, together with the appropriate standard error or variance. It is difficult to reinterpret such results in terms of disease categories, particularly if the mean difference in scores is a fraction of the difference in score between disease categories and if the scores corresponding to disease categories are separated by unequal intervals (which might be the case if the diseases had very different prognoses). Feinstein's approach is further complicated by allowing for differences in terminology that pathologists use. However these methods do assess repeatability, validity and bias, are statistically testable and do not necessarily require very large numbers of cases.

In contrast the overall proportion or percent agreement is a simple and intuitively obvious measure of repeatability (Table 1). Unfortunately this index can give spuriously high estimates of repeatability: for example if two observers assigned subjects at random to the different categories but in similar proportions (perhaps suggested by knowledge of the actual prevalences of the categories in the sample) then the overall proportion of agreement might be surprisingly high even though determined by chance alone in this instance. Some of the alternative measures proposed to avoid this difficulty[16–18] cannot be calculated if certain entries in the table are zero, or are restricted to a 2 × 2 table or only measure agreement in the diagnosis of positives.

Table 1 *Observed and expected proportion of agreement*

Observer 1	*Observer 2*		
	Category 1	*Category 2*	
Category 1	a	b	a + b
Category 2	c	d	c + d
Total	a + c	b + d	N = (a + b + c + d)

Observed proportion of agreement = $\frac{a + d}{N}$.

Expected proportion of agreement =

$$\left[\frac{(a + c)(a + b)}{N} + \frac{(b + d)(c + d)}{N}\right] / N .$$

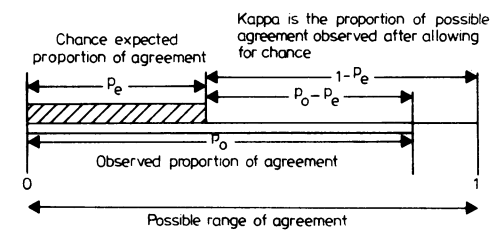
Since repeatability may be regarded as a special case of association indices of association can be used to measure repeatability provided the direction of association is known. However those based on χ^2 statistics do not indicate direction and the maximum value can vary with the number of categories being examined or may only be achieved when the numbers assigned by each observer to corresponding categories are equal, whilst other statistics such as Yule's Gamma are affected by zero entries too.[19]

A more fundamental objection however to the use of these measures of association and scores is that some underlying ordering of the diagnostic categories is necessary. Unless an unequivocal ordering is possible based for example on epithelial dysplasia[13] the degree of association can be changed by an arbitrary reordering of categories. This potential difficulty is avoided if no ordering is assumed to exist.

In addition, those measures with a probabilistic interpretation, like the overall proportion of agreement, would seem to be the easiest to understand. Cohen's kappa statistic[20] has been shown to unify several of these approaches by allowing for chance expected agreement[21] and it is easy to conceptualise as the proportion of agreement having allowed for that expected by chance (see Figure). It ranges from negative values (indicating disagreement) through 0 (chance agreement) to +1 (perfect agreement) like a correlation coefficient, of which it is a form. From the Figure it can be seen that for agreement since $\kappa \leq p_o$, the widely used measure, proportion of agreement, tends to overestimate repeatability.

ADVANTAGES AND DISADVANTAGES OF KAPPA

The kappa statistic has several advantages as a



Interpretation of kappa (ignoring negative values)

measure of repeatability: it attempts to correct for chance expected agreement, it is not affected by a few zero cell entries, it indicates both direction and magnitude of agreement, its maximum positive value is not affected by the number of subjects examined nor by unequal marginal totals or by the number of categories. It is easy to calculate and is not restricted to a 2 × 2 table.

The main disadvantage of the kappa statistic is the number of cases required. Although the number needed to assume a normal sampling distribution may be as small as $2C^2$, where C is the number of categories,[22] the sample size needed to achieve a particular error rate may be considerably larger (see appendix). Exact probability methods are also available[23] which apply to any sample size but they are more difficult to calculate, do not yield confidence limits and only test the null hypothesis that $\kappa = 0$. Whilst this is the usual assumption in significance testing, as Cohen[20] has observed it makes little sense to test against zero since one would expect at least better than chance agreement between say two trained pathologists. As a yardstick Landis and Koch[24] have suggested that a value of kappa $\geq +0{\cdot}75$ should be taken as an arbitrary index of excellent agreement beyond chance expectation, whilst values $\leq +0{\cdot}40$ should be regarded as poor. Confidence intervals provide an assessment of whether a particular kappa differs significantly from such a predetermined value. A second disadvantage is that the minimum negative value of kappa is not a fixed quantity.[20] In practice this is not a serious defect as when studying agreement only positive values are of primary interest and it may be outweighed by other advantages of kappa which is flexible enough to be used to compare multiple observers diagnoses[21 25] or, if desired, to take into account the lesser severity of minor disagreements by calculating the weighted form.[21 26] The weights roughly correspond to Feinstein's "discrepancy scores"[15] and could be assigned on the lines suggested by Owen and Tighe.[27] This has not been adopted here though because the precise values for weights may be disputed whereas exact concordance of diagnosis is fact.

It appears therefore that the value of kappa is a suitable statistic for evaluating histological diagnosis but despite this only relatively recently has it been applied to histopathology.[28–32] Even in these examples however, the usual approach has still been to test an obtained value of kappa against the null hypothesis value of zero, rather than against some desired standard of agreement.

Table 2 *Correlation between visual gradings of elastosis in breast carcinoma*

		Observer 1				
	Grade	*0*	*1*	*2*	*3*	*Total*
Observer 2	0	10	4	0	0	14
	1	2	19	5	0	26
	2	1	6	14	3	24
	3	0	1	3	12	16
	Total	13	30	22	15	80

$$p_o = \frac{10+19+14+12}{80} = 0{\cdot}69$$

$$p_e = \left[\frac{(13\times14)}{80} + \frac{(30\times26)}{80} + \frac{(22\times24)}{80} + \frac{(15\times16)}{80}\right]/80 = 0{\cdot}27$$

$$\kappa = \frac{0{\cdot}69-0{\cdot}27}{1-0{\cdot}27} = +0{\cdot}57$$

95% confidence limits = $\kappa \pm 1{\cdot}96 \times$ standard error of κ = +0·43 to +0·71

From: Robertson AJ, Brown RA, Cree IA, *et al.* Prognostic value of measurement of elastosis in breast carcinoma. *J Clin Pathol* 1981;**34**:738–43.

EXAMPLES OF KAPPA IN PRACTICE

(i) *Repeatability*

Table 2 shows how kappa may be calculated, in this case indicating only moderate interobserver repeatability in the assessment of elastosis in breast carcinoma. The value p_e is simply the sum of the expected numbers in each diagonal cell (calculated exactly as for a χ^2 test of association) divided by the total number of cases.

Another example may be given for the main diagnostic categories of lung cancer using the results published in Feinstein's paper[15] from which the average within and pairwise between observer kappa may be calculated by Fleiss' method.[21] The values which were obtained perhaps not surprisingly show that within observer repeatability was significantly greater than that between observers ($0{\cdot}01 < p < 0{\cdot}05$):

average within observer kappa = +0·81 95% confidence limits +0·68 to +0·94.
(based on two readings of each slide)

Table 3 *Validation: sensitivity and specificity of a test*

		True or expert diagnosis	
		Present	*Absent*
Observer's opinion	*Present*	a	b
	Absent	c	d

$$\text{Sensitivity} = \frac{a}{a+c}.$$

$$\text{Specificity} = \frac{d}{b+d}.$$

Table 4 *Validation of a test with more than two diagnostic categories*

		True or expert diagnosis		
		Severe	*Moderate*	*Absent*
Observer's opinion	*Severe*	a	b	c
	Moderate	d	e	f
	Absent	g	h	i

Three separate estimates of both sensitivity and specificity are possible, depending on whether moderate and severe, absent and moderate or absent and severe categories are pooled.

average pairwise between observer kappa = +0·47 95% confidence limits +0·29 to +0·64

(ii) *Validity*

This is usually described in terms of sensitivity and specificity (Table 3). These terms only apply to dichotomous classifications, of which a 2 × 2 table is an example but they are useful because tests for different purposes require different values for sensitivity and specificity and a single index figure would not indicate whether a "good" test had high sensitivity or high specificity. On the other hand for a 3 × 3 table (Table 4) estimates of sensitivity and specificity do not make sense unless one pools different categories (a problem illustrated by Lambourne and Lederer's study[12]). It is here that a summary index is useful and exactly as with repeatability, kappa is preferable to the overall proportion of agreement. Some studies on agreement between observers and an expert or "official" diagnosis such as the one conducted by the Royal College of Pathologists of Australia[33] have been flawed because in effect they have only examined sensitivity, without examining the ability of pathologists to agree on negative cases (specificity).

If the kappa statistic is used it may be calculated in the usual way, for example Table 5 in which sputum cytology is validated against histological appearances. Agreement is excellent since the upper confidence limit exceeds +0·75, although since the lower limit falls in the "moderate" range a case could be made that agreement is "moderate to excellent" and that a larger sample would reduce this uncertainty. A word of caution however: the formulae given presuppose that neither of the observers (expert or not) selects the subjects to be representative of the categories. Instead, either a third party should select the subjects or the series should comprise all (or a random sample of all) cases as received by one or both observers, in effect making the assessment "blind" for both.

(iii) *Bias*

Since bias will tend to cause unequal marginal totals it is advisable to examine these first. If marginal totals are unequal then the maximum possible amount of agreement must be less than 100% and Cohen[20] has indicated how to assess the extent of disagreement as a ratio of observed kappa:maximum possible kappa given the different marginal totals. Even if marginal totals are equal, however, there may still be some bias, detectable by a closer examination of the table.

A simple approach is to tabulate the standardised residual frequencies for each cell of the table, that is:

$$\frac{(\text{observed frequency} - \text{expected frequency})}{\sqrt{(\text{expected frequency})}},$$

and to observe which are significantly different from zero in either direction, thus highlighting areas of consistent disagreement, or bias. Following Bishop[34] the critical frequency is given by $(\sqrt{\chi^2})/C$ where C is the number of categories and χ^2 is the critical value for a C × C table at a given significance level and is determined from a standard table of χ^2 values. For Table 6 at the 5% significance level, the critical frequency is 1·03. The agreement cells will all more or less have residuals >+1·03 because kappa differs significantly from zero and is positive. The disagreement cells will either be non-significantly different from zero or will be >±1·03. Those with significant *negative* values are also to be expected if agreement is significant but cells with a value >1·03 would indicate a category of disagreement in which significantly more observations occurred than would

Table 5 *Accuracy of sputum cytology*

		Biopsy diagnosis of cell type				
		Squamous	*Small cell*	*Adeno*	*Large cell*	*Total*
Sputum cytological diagnosis of cell type	Squamous	111	1	6	2	120
	Small cell	1	11	0	0	12
	Adeno	0	0	16	1	17
	Large cell	4	0	3	5	12
	Total	116	12	25	8	161

p_o = 0·89 p_e = 0·56 κ = +0·75 95% confidence limits +0·64 to +0·86.

From: Payne CR, Hadfield JW, Stavin PG, *et al.* Diagnostic accuracy of cytology and biopsy in primary bronchial carcinoma. *J Clin Pathol* 1981:**34**:773–8.

Table 6 *Indices of bias from data in Table 5*
(1) *Table of residuals*

		Biopsy diagnosis of cell type			
	Category	*Squamous*	*Small cell*	*Adeno*	*Large cell*
Sputum diagnosis of cell type	Squamous	+2·6	− 2·64	−2·93	−1·62
	Small cell	−2·59	(+10·7)	−1·36	(−0·77)
	Adeno	−3·5	(− 1·12)	+8·22	(+0·17)
	Large cell	−1·58	(− 0·94)	(+0·83)	(+5·74)

No residual in a disagreement cell is >+1·03 and so no serious problem seems to exist. Those residuals in brackets are derived from cells with an expected frequency <5 and are therefore unreliable.

(2) *Kappa, p_s and p_h statistics for diagnostic categories*

Category	*Kappa*	*p_s (systematic error)*	*p_h (haphazard error)*
Squamous	0·78	0·025	0·062
Small cell	0·91	0·000	0·012
Adeno	0·73	0·050	0·012
Large cell	0·47	0·025	0·037
Overall	0·75	0·049	0·062

Kappa values should ideally be high, whilst p_s and p_h statistics which measure error should be low. Unfortunately as yet no general standard exists by which to evaluate p_s and p_h statistics.

be expected by chance alone. This method does have some limitations however: since the critical frequency is the square root of χ^2 for an individual cell it would be prudent to apply it only to those cells with an expected frequency of 5 or more and also the method is rather insensitive.

A second approach is that of Fleiss[21] in which the extent of agreement is calculated for each diagnostic category. This is done by collapsing the data for each agreement cell into a separate fourfold table from which a kappa statistic may be found. For Table 5 this means the dichotomous distinctions: squamous carcinoma *v* all others, small cell carcinoma *v* all others, adenocarcinoma *v* all others and large cell carcinoma *v* all others. The results are then tabulated as a list of diagnostic categories with corresponding levels of agreement. Those categories for which agreement is poor would require further study.

Yet another technique which can usefully be combined with estimation of overall agreement, is to distinguish systematic and haphazard components of poor agreement by means of p_s and p_h statistics;[31] but there is no reason why they should not be applied like kappa to separate diagnostic categories as well.

A third approach is that suggested by Holman[29] who has advocated (i) the calculation of conditional probability levels for each cell (that is, the probability that pathologist B will make a given diagnosis, supposing that pathologist A has made the same diagnosis) and (ii) calculation of kappa values for *each cell*. Under these circumstances of course many of the kappa values occur in disagreement cells and so these values are interpreted as the degree of *association* between pathologist A's diagnosis and pathologist B's different diagnosis. An alternative method would be to extend Fleiss' approach by further collapsing the original table into separate dichotomous distinctions such as (for Table 5) "squamous carcinoma by observer 1 but adenocarcinoma by observer 2" *v* "all other categories". The result by either method is a matrix of kappa values, some positive and some negative. Exactly as with residuals, the high positive values should lie in the agreement cells whilst the disagreement cells should show strongly negative values or at most be only weakly positive. A disadvantage of Holman's method is that it assumes the marginal totals are equal. Extending Fleiss' technique does allow for unequal marginal totals but seems to offer little over calculation of residuals for each cell.

The results from these approaches are tabulated in Tables 6 and 7. Whatever method is chosen to evaluate bias, the real value lies not just in its detec-

Table 7 *Kappa values for each cell derived from Table 5*
(1) *Method of Holman et al.*

		Biopsy diagnosis of cell type			
	Category	*Squamous*	*Small cell*	*Adeno*	*Large cell*
Sputum diagnosis of cell type	*Squamous*	0·83	−0·07	−0·01	−0·01
	Small cell	−2·32	0·91	−0·04	−0·03
	Adeno	−0·81	−0·08	0·31	0·14
	Large cell	−1·41	−0·08	0·19	0·43

NB. Negative values <−1 are possible with kappa, unlike a correlation coefficient.

(2) *Extending Fleiss' method*

	Biopsy diagnosis of cell type			
Category	*Squamous*	*Small cell*	*Adeno*	*Large cell*
Squamous	0·78	−0·14	−0·23	−0·04
Small cell	−0·14	0·91	−0·11	−0·06
Adeno	−0·23	−0·10	0·73	0·01
Large cell	−0·08	−0·80	0·07	0·47

The overall patterns of results are similar although the correlation with residual values is greater for method (2) than (1), to be expected since both the former and the method of residuals take differing marginal totals into account.

tion but in discussing possible reasons. For example, poor agreement on a particular diagnostic category may indicate a problem of differential diagnosis. In Table 6(2), the relatively poor agreement on large cell carcinoma may be because it is difficult to distinguish from adenocarcinoma, for the residuals in the adenocarcinoma/large cell carcinoma disagreement cells are positive, although not significantly so. On the other hand if all other types of carcinoma were easily confused with adenocarcinoma then no particular disagreement cells might be prominent. In addition, if indices of bias for corresponding cells are unequal, this might be the result of new knowledge acquired between the two occasions or as in Table 5 by the availability of extra information in the biopsy: Compare the residual in the cell "sputum diagnosis of adenocarcinoma/biopsy diagnosis of squamous carcinoma" with that in "biopsy diagnosis of adenocarcinoma/sputum diagnosis of squamous carcinoma" (Table 6).

Once repeatability and validity have been measured it may be necessary to improve them since a classification showing poor agreement, for example Hartveit's Tumour grading system[35] is surely hard to justify. More objective methods such as morphometry and immunohistochemistry may help but unsuspected pitfalls may exist even in the simplest of techniques as Ellis[36] pointed out, whilst lack of standardisation of immunological reagents and procedures may account for varying reports of CEA negativity in malignant mesotheliomas.[37 38]

In conclusion, epidemiological surveys, laboratory quality control and clinical work all require some measure of repeatability and validity. This applies to histopathology perhaps more than to other specialties for there is a danger that histological diagnosis, often regarded as definitive by the non-specialist, may be seen as somehow beyond question, and this is plainly not so.

The variety of methods used for assessment of histological diagnosis and their unsystematic application do suggest some uncertainty as to the best approach, however. The kappa statistic is a tool which is applicable to all forms of histological diagnosis and which, because of its simplicity, can provide a uniform criterion of repeatability from which all may benefit.

The data in Tables 2 and 5 together with the data used to illustrate within and between observer variability appear with the permission of the authors and the editors of the *Journal of Clinical Pathology* and the *American Review of Respiratory Disease.*

Appendix

Estimating sample size

Sample size can be roughly estimated using the formula:

$$N \simeq \left[\frac{Z_\alpha}{\kappa_L\,(1-f)}\right]^2 \times \left[\kappa_L + \frac{f}{c-1}\right] \times \left[f-\kappa_L\right]$$

where

κ_L = the minimum value of kappa we wish to detect at a given significance level (this need only be one-tailed).

Z_α = the standard normal deviate for the desired significance level.

f = the fraction of the true value of kappa which κ_L is specified to be.

c = the number of diagnostic categories being used.

Thus if there are three categories, $f = 0{\cdot}9$, $\kappa_L = 0{\cdot}75$ and $Z_\alpha = 1{\cdot}64$ (representing a 5% significance level with one tail)

$$N \simeq \left[\frac{1{\cdot}64}{0{\cdot}75\,(1-0{\cdot}9)}\right]^2 \times \left[0{\cdot}75 + \frac{0{\cdot}9}{2}\right] \times \left[0{\cdot}9 - 0{\cdot}75\right]$$

$$\simeq \underline{86}$$

References

1. Blenkinsop WK, Stewart-Brown S, Blesovsky L, *et al.* Histopathology reporting in large bowel cancer. *J Clin Pathol* 1981;**34**:509–13.
2. Macartney JC, Henson DE, Codling BW. Quality assurance in anatomic pathology. *Am J Clin Pathol* 1981;**75**, 3 Suppl: 467–75.
3. Walter JB, Israel M. *General pathology* 5th ed. Edinburgh: Churchill-Livingstone, 1979.
4. Gould RP. In: Anthony PP, Woolf N, eds. *Recent advances in histopathology* No 10. Edinburgh: Churchill-Livingstone, 1978.
5. Kramer IRH. Computer-aided analyses in diagnostic histopathology *Postgrad Med J* 1975;**51**, 600:690–4.
6. Fisher ER, Redmond C, Fisher B. Histologic grading of breast cancer. *Pathol Annu* 1980;**15** pt 1:239–51.
7. Wright EA. Quality control in histopathology. *Proc R Soc Med* 1975;**68**:619–22.
8. Penner DW. In: Sommers SC, ed. *Pathology Annual* Vol 8. New York: Appleton Century Crofts, 1973.
9. Barr WT, Williams ED. Value of external quality assessment of the technical aspects of histopathology. *J Clin Pathol* 1982;**35**:1050–6.
10. Barr WT. Technical quality control in histopathology. *J Clin Pathol* 1978;**31**:996–8.
11. Iversen OH, Sandnes K. The reliability of pathologists. A study of some cases of lymph node biopsies showing giant follicular hyperplasia or lymphoma. *Acta Pathol Microbiol Scand* 1971;**79**:330–4.
12. Lambourne A, Lederer H. Effects of observer variation in population screening for cervical carcinoma. *J Clin Pathol* 1973;**26**:564–9.

[13] Cocker J, Fox H, Langley FA. Consistency in the histological diagnosis of epithelial abnormalities of the cervix uteri. *J Clin Pathol* 1968;**21**:67–70.

[14] Sissons HA. Agreement and disagreement between pathologists in histological diagnosis. *Postgrad Med J* 1975;**51** no 600:685–9.

[15] Feinstein AR, Gelfman NA, Yesner R, *et al.* Observer variability in the histopathologic diagnosis of lung cancer. *Am Rev Respir Dis* 1970;**101**:671–84.

[16] Carpenter RG. In: Bullpitt CJ, Dollery CT, Carne S, eds. Change in symptoms of hypertensive patients after referral to hospital clinic. *Br Heart J* 1976;**38**:121–8.

[17] Rogot E, Goldberg ID. A proposed index for measuring agreement in test—retest studies. *J Chron Dis* 1966;**19**:991–1006.

[18] Rose GR, Barker DJP. *Epidemiology for the uninitiated.* London: British Medical Association, 1979.

[19] Blalock HM. *Social statistics* 2nd (revised) edition. McGraw-Hill, 1979.

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

[21] Fleiss JL. Statistical methods for rates and proportions 2nd ed. New York: John Wiley & Sons, 1981.

[22] Cicchetti DV, Fleiss JL. Comparison of the null distribution of weighted kappa and the C ordinal statistic. *Appl Psychol Measurement* 1977;**1**:195–201.

[23] Wackerley DD, McClave JT, Rao PV. Measuring nominal scale agreement between a judge and a known standard. *Psychometrika* 1978;**43**:213–23.

[24] Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74.

[25] Light RJ. Measures of response agreement for qualitative data: some generalisations and alternatives. *Psychol Bull* 1971;**76**:365–77.

[26] Spitzer RL, Cohen J, Fleiss JL, Endicott J. Quantification of agreement in psychiatric diagnosis. *Arch Gen Psychiat* 1967;**17**:83–7.

[27] Owen DA, Tighe JRT. Quality evaluation in histopathology. *Br Med J* 1975;i:149–50.

[28] Bain GO, Koch M, Hanson J. Feasibility of grading prostatic carcinomas. *Arch Pathol Lab Med* 1982;**106**:265–7.

[29] Holman CDJ, James IR, Heenan PJ, *et al.* An improved method of analysis of observer variation between pathologists. *Histopathology* 1982;**6**:581–9.

[30] Stanley KE, Matthews MJ. Analysis of a pathology review of patients with lung tumours. *J Natl Cancer Inst* 1981;**66**:989–92.

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

[32] 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.

[33] Royal College of Pathologists of Australia (1969–71). Reports of surveys by the Board of Education of the college: Sydney.

[34] Bishop YMM, Fienberg SE, Holland PW. Discrete multivariate analysis: theory and practice. MIT Press, 1975.

[35] Stenkvist B, Westman-Naeser J, Vegelius J, *et al.* Analysis of reproducibility of subjective grading systems for breast carcinoma. *J Clin Pathol* 1979;**32**:979–85.

[36] Ellis PSJ, Whitehead R. Mitosis counting, a need for reappraisal. *Hum Pathol* 1981;**12**:2–3.

[37] Carson JM, Pinkus GS. Mesothelioma: profile of keratin proteins and CEA — an immunoperoxidase study of 20 cases and comparison with pulmonary adenocarcinoma. *Am J Pathol* 1982; **108**:80–8.

[38] Kwee WS. Histologic distinction between malignant mesothelioma benign pleural lesion and carcinoma metastasis. Evaluation of the application of morphometry combined with histochemistry and immunostaining. *Virchows Archiv A* 1982;**397**:287–300.

Requests for reprints to: Dr PBS Silcocks, Department of Clinical Epidemiology & Social Medicine, St George's Hospital Medical School, Cranmer Terrace, Tooting, London SW17 0RE, England.