J Clin Pathol 1989;**42**:231–238

Observer variability in histopathological reporting of cervical biopsy specimens

A J ROBERTSON, J M ANDERSON, J SWANSON BECK, R A BURNETT, S R HOWATSON, F D LEE, A M LESSELLS, K M McLAREN, S M MOSS, J G SIMPSON, G D SMITH, H B TAVADIA, F WALKER

From the Department of Pathology, Ninewells Hospital and Medical School, Dundee, Scotland

SUMMARY Sections from 100 cervical biopsy specimens were studied by 12 consultant histopathologists to determine the robustness of the existing pathology terminology and classification. Analysis by κ statistics showed good agreement in the diagnosis of CIN 3 and squamous carcinoma but an inability to distinguish accurately between the lesser grades of CIN.

It is recommended that the classification be changed to low grade (present CIN 1 and 2) and high grade (present CIN 3) categories alone. There was very poor agreement in the identification of cellular changes associated with human papilloma virus (HPV) infection.

Several novel analytical methods of assessing the severity of uterine cervical intraepithelial neoplasia (CIN) have been proposed,[1 2] but histological assessment remains the basis for determination of treatment, clinical management, and subsequent follow up of patients. Although clear criteria for the diagnosis and grading of CIN have been described,[3] such assessments are subjective and prone to intra- and inter-observer variation.[4 5] The problems of histological assessment have been further complicated by the increasing recognition of human papilloma virus (HPV) infection[6 7] which may be an aetiological factor in the development of CIN.[8 9] HPV infection may be indicated by koilocytosis and other changes that distort cellular appearances and so may apparently exaggerate the severity of the premalignant appearances of the cervical epithelium, particularly in the higher layers—making grading more difficult.

It is reasonable that efforts should be made to establish the degree of confidence which can be given to the histological reporting of cervical biopsy lesions by pathologists and to determine the robustness of the existing terminology and classification. We describe the findings of a study of cervical biopsy specimens conducted by a group of 12 pathologists, all of consultant grade, but with varying degrees of experience.

Accepted for publication 1 September 1988

Material and methods

COMPOSITION OF PANEL

Twelve histopathologists were invited to join the study with a deliberate attempt by the organisers to obtain a composition representative of Scottish pathology as a whole. The members came from pathology laboratories in Aberdeen (n = 2), Dundee (n = 2), Edinburgh (n = 2), Airdrie (n = 1), Perth (n = 1), Stirling (n = 1) and Glasgow (n = 3) and varied in years of consultant experience (five to 25 years) and nature of substantive post (university staff n = 5: NHS staff n = 7). All the members of the group had undertaken their postgraduate training in Scotland.

CLASSIFICATION OF CERVICAL HISTOPATHOLOGY

At the initial meeting current cervical pathology terminology was reviewed and following discussion a proforma was designed for completion after examination of each slide in the circulation. This was modified in a minor way after the first circulation and the final form is shown in the figure. It was decided to keep the classification simple, but to relate it as closely as possible to everyday practice. The following disease categories were agreed: clinically important inflammation; immature squamous metaplasia; viral features; CIN grades 1–3; invasive squamous carcinoma; endocervical glandular dysplasia; and adenocarcinoma. During the course of the study the pathologists met after each circulation of slides to discuss the diagnostic problems and use of terminology.

SLIDE NO: ☐

QUALITY: Satisfactory = 1
Unsatisfactory = 2 ☐

NORMAL: Yes = 1
No = 2 ☐

SIGNIFICANT INFLAMMATION: Stromal = 1
Epithelial = 2
Stromal & Epithelial = 3
None = 4 ☐

IMMATURE SQUAMOUS METAPLASIA: Present = 1
None = 4 ☐

VIRAL FEATURES: Outwith CIN only = 1
Within CIN only = 2
Both outwith & within CIN = 3
None = 4 ☐

CIN: Grade 1 = 1
Grade 2 = 2
Grade 3 = 3
None = 4 ☐

INVASION: Possible = 1
Definite = 2
None = 4 ☐

If possible or definite, depth in mm = ☐

ENDOCERVICAL GLANDULAR DYSPLASIA: Yes = 1
None = 4 ☐

ADENOCARCINOMA: Yes = 1
None = 4 ☐

PATHOLOGIST CODE NO: ☐

DATE: ☐☐☐☐☐☐

Figure *Cervical histopathology.*

SELECTION OF HISTOLOGICAL MATERIAL

The study comprised a series of 100 consecutive cervical biopsy specimens from the surgical files of the laboratories at Perth Royal Infirmary and Ninewells Hospital, Dundee. Since colposcopic cervical biopsy specimens are usually examined at several levels, each set of slides was initially screened by one of the group (AJR), who selected the slide with most surface epithelium. The study was designed to test the robustness of terminology and the reproducibility of individual feature assessment; the fact that one of the group had previewed the slides was not considered likely to bias unduly the findings. The study was coordinated from Dundee by an independent organiser who initially allocated a confidential code to each participant. Four circulations of 30 slides were studied consecutively with a meeting of the group at the end of each circulation. At these meetings selected slides on which there had been substantial disagreement were discussed using a microscope projection system. The third circulation included six repeat slides from the previous two circulations, and in the fourth circulation a further 14 slides were repeated. The participants in the study were not informed of the inclusion of slides previously circulated. The slides selected for repeat study were chosen because of their apparent difficulty when reported in the previous circulation.

A seven point diagnostic scale was constructed. Where no CIN was recorded, slides were classified as normal, inflammatory only, or showing immature squamous metaplasia. CIN 1, 2, and 3 were treated as separate categories and it was also noted whether invasion (squamous carcinoma) was absent, possible, or definite (figure). The recording of viral features was also analysed using four categories: none; present outwith CIN only; present within CIN only; and present both outwith and within CIN.

A problem with the analysis of studies of this kind is the lack of knowledge of the "correct" diagnosis of each slide. This necessitates some decision concerning the reference categories against which individual diagnoses are to be compared. In some studies an "expert peer group" diagnosis is used.[10] In the present study a "majority diagnosis" for the CIN category was taken as that which occurred most frequently for each slide. In most cases this was equivalent to taking the median category. The comparison of individual diagnoses with this "majority" opinion gives a measure of the interobserver agreement. The relative performance of the individual pathologists was assessed by calculating the mean deviation of his or her diagnoses from the majority, assuming the difference between each category to be one unit. Although the categorical nature of the data and the fact that while the categories are ranked in increasing severity the importance of the difference between adjacent categories is not uniform means that the use of parametric statistics is not strictly valid, the mean deviation is used as a descriptive statistic. A positive value indicates a tendency to overdiagnosis and a negative value a tendency towards underdiagnosis compared with the majority.

Kappa statistics are a measure of overall agreement which do not require any assumption concerning the "correct" diagnosis and which include a correction for the amount of agreement which would be expected by chance alone.[11] The overall value of κ for more than two categories is defined as a weighted average of the values for the individual categories.[12] The value of κ can range from $-1{\cdot}0$ to $+1{\cdot}0$. A value of 0 indicates chance agreement only, while a value of $+1{\cdot}0$ indicates perfect agreement. A negative value would imply systematic disagreement between observers. It is generally accepted that a value of 0·75 or above reflects excellent agreement, with 0·4–0·75 suggesting fair to good agreement and values less than 0·4 meaning agreement is poor.[13]

All the analyses were carried out for each separate batch of slides and for the batches combined. Batch 1, however, was excluded as this was regarded as a "pilot" survey, and equally the second circulation of the repeated slides was excluded. The pathologists' diagnoses of those slides circulated twice were com-

Table 1 *Kappa statistics: batch 1*

Seven categories:

	Normal	*Inflammatory*	*Immature squamous metaplasia*	*CIN1*	*CIN2*	*CIN3*	*CIN3 and possible invasion*	*Overall*
κ	0·1	0·17	0·45	0·13	0·14	0·41	0·17	0·27

Four categories:

	Normal/inflammatory immature squamous metaplasia	*CIN1*	*CIN2*	*CIN3/CIN3 and possible invasion*	*Overall*
κ	0·45	0·13	0·14	0·52	0·34

Viral features:

	None	*Outwith CIN only*	*Within CIN only*	*Both outwith and within CIN*	*Overall*
κ	0·30	0·08	0·15	0·13	0·18
			*0·27		0·24

*Reanalysis of combined categories.

Table 2 *Kappa statistics: batches 2–4*

Four categories:	*Normal/inflammatory immature squamous metaplasia*	*CIN1*	*CIN2*	*CIN3 with/ without invasion*	*Overall*
Batch 2	κ 0·41	0·24	0·11	0·51	0·35
Batch 3	κ 0·61	0·28	0·19	0·71	0·51
Batch 4	κ 0·38	0·16	0·23	0·51	0·36

Viral features:	*None*	*Outwith CIN only*	*Within CIN only*	*Both outwith and within CIN*	*Overall*
Batch 2	κ 0·25	0·14	0·03	0·24	0·17
			* 0·31		0·25
Batch 3	κ 0·18	0·09	0·09	0·12	0·12
			* 0·30		0·21
Batch 4	κ 0·1	0·08	0·03	0·07	0·07
			* 0·13		0·11

*Reanalysis of combined categories.

pared to give a measure of the intraobserver variability.

Results

Table 1 gives the κ statistics for batch 1, for agreement on the seven categories on the scale, with an overall value of 0·27. Overall, 60% of diagnoses were in agreement with the majority view; consensus was greatest in CIN 3 and least on CIN 1 and CIN 2. Reducing the scale to four categories by combining the three benign categories of normal, inflammatory, and immature squamous metaplasia into one, and also combining the CIN 3 categories with or without invasion, has the effect of increasing the overall value of κ from 0·27 to 0·34. Of the individual categories, agreement was best on immature squamous metaplasia and CIN 3, with only poor agreement on CIN 1 and CIN 2. Table 1 gives the κ statistics for the recognition of viral features. Agreement is generally poor, with an overall value of κ of 0·18, which only increases to 0·24 if the categories of viral features within CIN and both within and outwith CIN are combined.

Table 2 gives the equivalent κ statistics for batches 2, 3 and 4, respectively. Although the overall value of κ for the CIN categories is improved for batch 3, this is to some extent due to the presence in that batch of several slides on which there was a high level of agreement on definite invasion. A smaller number of such slides may explain the fall in overall agreement on batch 4, which included a number of slides recirculated because they had previously been the subject of disagreement. Table 2 shows that the agreement on the presence of viral features remained consistently low.

Table 3 shows the comparison of individual reports with the majority diagnosis for batches 2–4 combined (with repeat readings excluded). The overall agreement with the majority was 76%, with highest agreement in the CIN 3 category and the poorest agreement on CIN 1 and CIN 2. The κ statistics for batches 2–4

Table 3 *Agreement with majority diagnosis: batches 2–4*

		Majority category			
	No CIN	*No CIN*	*CIN1*	*CIN2*	*CIN3*
Individual reports		106 (74%)	19	20	8
	CIN1	25	35 (49%)	18	5
	CIN2	12	17	88 (52%)	63
	CIN3	1	1	42	512 (87%)
	Total reports (No of slides)	144 (12)	72 (6)	168 (14)	588 (49)
	Overall agreement with majority = 76% (741/984)				

Table 4 *Kappa statistics: batches 2–4*

Four categories:	*Normal/inflammatory immature squamous metaplasia*	*CIN1*	*CIN2*	*CIN3 with/ without invasion*	*Overall*
κ	0·52	0·24	0·20	0·61	0·44
Viral features:	*None*	*Outwith CIN only*	*Within CIN only*	*Both outwith and within CIN*	*Overall*
κ	0·21	0·11	0·05	0·15	0·14
			* 0·27		0·21

*Reanalysis of combined categories.

combined presented a similar picture (table 4). The overall value for the four CIN categories was 0·44; the agreement on viral features was poor ($\kappa = 0{\cdot}21$).

To determine the effect of the presence of virus on the diagnosis of CIN the slides in batches 2–4 were divided into three groups: 44 slides where the majority of pathologists reported viral features within CIN (group A), 16 slides where the majority reported either no viral features, or virus outwith CIN only (group B), and 22 slides on which there was either a 6–6 or a 7–5 split (group C). Table 5 shows the κ statistics for the three groups. Although the overall agreement was best for group B, this was probably due to the fact that the reporting of viral features was highest where slides were reported as CIN 1 or CIN 2, which are those categories on which agreement was poor. Agreement on CIN 1 and CIN 2 was, in fact, slightly better in group A than in group B. When the κ statistics for the clinically important groupings of normal, inflammatory, and immature metaplasia compared with CIN or viral features were compared for batches 2–4, a combined value of $\kappa = 0{\cdot}39$ was obtained, this still being a relatively poor value.

Table 6 gives the results of the repeat readings on the 21 recirculated slides. The overall agreement on rereading was only 51% for the eight diagnostic categories, but 60% if only four categories were used. For viral features the agreement was 48% for four categories, but 59% if the last two categories were combined.

Table 7 shows the level of agreement with the majority diagnosis for each participating pathologist in batches of slides 2–4. In most cases the same pathologists tended to underrate (indicated by a negative mean deviation) or overrate (positive mean deviation) in all batches of slides.

Discussion

Decisions affecting clinical management, treatment, and follow up are often based on a histopathological diagnosis, this being regarded as a "gold standard". If this is to continue to be the case, however, it is important for histopathogists to examine critically current histopathological classifications and their reliability in practice. Quality control programmes have been started in histopathology[14]: these entail circulation of a series of slides round a group of pathologists with subsequent comparison of the reported overall diagnosis. The intention of these programmes is to allow individual pathologists to compare their diagnosis with the majority diagnosis and hence, in essence, obtain information about their diagnostic capability when compared with that of their peers. None of these circulations, however, addresses the central issue of how pathologists actually reach a final diagnosis, or of how robust terminology is, and no information is obtained on the expected confidence which can be placed on a diagnosis. Inevitably, the need for such information is becoming more urgent as the problems of finance allocation in the health service demand greater diagnostic accuracy and the ability to decide whether treatment is cost effective.

It is clear from our study on cervical histopathology

Table 5 *Kappa statistics: for CIN in presence or absence of viral features: batches 2–4*

		Normal/inflammatory immature squamous metaplasia	*CIN1*	*CIN2*	*CIN3/CIN/with possible or definite invasion*	*Overall*
A (viral features present)	κ	0·24	0·25	0·17	0·50	0·33
B (no viral features)	κ	0·60	0·04	0·07	0·73	0·53
C (no consensus)	κ	0·63	0·07	0·24	0·76	0·58

Table 6 *Repeat reading of slides*

Disease categories:

		Second reading							
		Normal	*Inflammatory*	*Immature squamous metaplasia*	*CIN1*	*CIN2*	*CIN3*	*CIN3 with possible invasion*	*CIN3 with definite invasion*
First reading	Normal	0	2	1	0	0	0	0	0
	Inflammatory	2	7	4	9	4	2	1	0
	Immature squamous metaplasia	0	7	15	6	3	0	1	0
	CIN1	0	7	4	15	6	4	1	0
	CIN2	1	2	3	11	29	9	0	0
	CIN3	0	4	2	2	13	55	1	0
	CIN3+	0	0	0	0	0	5	2	0
	CIN3++	0	0	0	0	0	0	0	0

Per cent agreement = eight categories 51% (123/240), four categories 60% (145/240)

Viral features:

		None	*Outwith CIN*	*Within CIN*	*Both outwith and within CIN*
First reading	None	59	23	9	5
	Outwith CIN	22	11	7	12
	Within CIN	6	6	27	13
	Both	9	11	12	19

Per cent agreement = 48% (116/240)

that pathologists are reasonably precise in their ability to distinguish benign lesions (normal/inflammatory/immature squamous metaplasia) from serious, clinically important, lesions (CIN 3 with or without microinvasion). Their ability to distinguish between CIN grades 1 and 2, however, is poor and there is little agreement about the presence of HPV infection. In clinical practice it is now standard to treat all patients who have CIN with laser treatment or cold coagulation, taking no account of the degree of CIN present, and therefore we propose that pathologists should consider no longer using the three-grade system of CIN diagnosis. The findings suggest that it would be more appropriate to have low grade CIN (present CIN 1 and 2) and high grade CIN (current CIN 3) categories. This simplification in classification would have parallel advantages for cytopathologists[15] who are now being encouraged to apply CIN grading to cytological specimens where there is good agreement about pronounced cytological changes, but not the lower grades of abnormality. In most laboratories pathologists reporting on cervical biopsy specimens would normally examine sections from several levels, unlike in the present study where the pathologists were asked to base their diagnosis on one section alone. This aspect of the study caused concern to some of the participants, but although multiple sections may in practice give greater security to diagnosis, the conditions of our study were the same for all observers and hence the results still provide a critical consideration of the strength of the existing terminology.

Table 7 *Measures of agreement of individual pathologists with majority diagnosis in grading of CIN: batches 2–4*

Pathologist	*Mean deviation*	*Variance*	*Per cent agreement with consensus*
1	−0·04	0·63	63
2	−0·23	0·53	73
3	−0·13	0·24	78
4	+0·28	0·38	73
5	−0·05	0·27	87
6	−0·38	0·63	68
7	0·05	0·34	76
8	−0·15	0·37	76
9	+0·01	0·41	81
10	+0·10	0·21	82
11	−0·23	0·48	72
12	−0·13	0·44	76

Our study of cervical specimens has produced similar levels of agreement between pathologists to that described in breast pathology.[16] It is interesting to note that in a recently published study of assessment of dysplasia in chronic inflammatory bowel disease[17] the overall agreement with the mean (four categories) was 66%. In common with some other investigations, this study also used analysis of variance techniques to estimate interobserver variation; for the most part, however, these rely on the assumption of a continuous diagnostic scale and do not take account of chance agreement. A recent paper which addressed these issues has criticised the κ statistic for depending on the prevalence of the categories of the classification scale.[18] Some evidence of this has been found in the present study, and the effect has been borne in mind when comparing the overall agreement for different series.

The κ statistics for individual categories, however, are also of considerable interest.

Our findings are somewhat similar to those of Ringsted *et al*,[4] who found that there was good agreement in the diagnosis of benign cervical disease and invasive cervical cancer. The reliability of the diagnosis of dysplasia and carcinoma in situ (CIN grade 3) was unsatisfactory. Although we found difficulty in the diagnosis of low grade CIN, this was not the case for CIN 3. A feature which these authors did not consider in their study and indeed may not have been encountered or recognised at that time but which nowadays is a common finding is the presence of squamous cellular changes associated with infection by HPV. In our study we tried to assess the presence or absence of these features and to state whether it was present outwith or in association with CIN, or both. This aspect of the study resulted in some of the poorest levels of agreement—the ability to distinguish these features being little better than what would be expected by chance. It was interesting to note that when split into two groups—those who reported viral features frequently and those who did not—pathologists produced similar results for intraobserver variability. HPV infection may induce cytological changes which may exaggerate the apparent severity of dysplastic changes in cervical epithelium. These appearances, however, must not negate the attempt to assess dysplasia as the virus itself may have an aetiological role. The standard histological methods of assessment do not permit identification of virus types which may be of greater oncogenic potential,[19] moreover, cytopathic effects may not reflect viral integration, a phenomenon of possibly greater importance in malignant transformation. It is thus evident that the cytological changes produced by HPV, their importance, and their contribution to the degree of dysplastic changes are areas fraught with interpretative difficulty. It is clear that additional methods, perhaps immunohistochemical, will need to be used to assess HPV changes with any degree of accuracy. A recent study with a polyclonal antibody to HPV has not been very promising.[20] We are currently examining the use of a new monoclonal antibody MC2 which is a marker of squamous differentiation.[21]

The slides that were recirculated were chosen because of their wide variation of results on the first reading. An unselected group of slides would probably have given a better intraobserver agreement than that of our study. These slides were examined further on conclusion of the study, and it seemed that there were many reasons for the variation found, including the presence of virus, pronounced inflammation partly obscuring the epithelium, a very small area of epithelial abnormality and specimen orientation problems.

In conclusion, it is evident from the findings of this study that there is considerable inter- and intra-observer variation in the histopathological reporting of cervical biopsy specimens using the existing classification and we propose that there is a need for a simplification of the classification and further analysis of the difficulty of grading CIN in the presence of HPV.

We are grateful to Mrs Joyce Duncan for secretarial assistance. This study is funded by the Scottish Home and Health Department Grant K/RES/D3/18.

References

1 Sincock AM, Evans-Jones J, Partington CK, Steele SJ. Quantitative assessment of cervical neoplasia by hydrolysed DNA assay. *Lancet* 1987;ii:942–3.
2 Singer A, Tay SK, Griffin JFA, Wickens DG, Dormandy TL. Diagnosis of cervical neoplasia by the estimation of octadeca-9-11-dienoic acid. *Lancet* 1987;i:537–9.
3 Buckley CH, Butler GB, Fox H. Cervical intraepithelial neoplasia. *J Clin Pathol* 1982;**35**:1–13.
4 Ringsted J, Antroup F, Asklund C, *et al.* Reliability of histopathological diagnosis of squamous epithelial changes of the uterine cervix. *Acta Pathol Microbiol Scand (Sect A)* 1978;**86**: 273–8.
5 Cocker J, Fox H, Langley FA. Consistency in the histological diagnosis of epithelial abnormalities of the cervix uteri. *J Clin Pathol* 1968;**21**:143–65.
6 McCance DJ, Campion MJ, Clarkson PK, Chester PM, Jenkins D, Singer A. The prevalence of human papilloma virus type 16 DNA sequences in cervical intraepithelial neoplasia and invasive carcinoma of the cervix. *Br J Obstet Gynaecol* 1985; **92**:1001–5.
7 Zur Hausen H, Gissman L, Schlehoffer JR. Viruses in the aetiology of human genital cancer. *Papers in Medicine and Virology* 1984;**30**:170–86.
8 Burne P, Woodman C, Meanwell C, Kelly K, Jordan J. Koilocytes and cervical human papilloma virus infection. *Lancet* 1986;i: 205–6.
9 Jenkins D, Tay SK, Dyson JL. Koilocyte frequency and the presence of human papilloma virus infection. *Lancet* 1986;i: 557–8.
10 Patchefsky AS, Potek J, Hoch WS, Libsketz HI. Increased detection of occult breast carcinoma after more thorough histological examination of breast biopsies. *Am J Clin Pathol* 1973;**60**:799–804.
11 Cohen JA. A coefficient of agreement for nominal scales. *Educational Psychological Measurements* 1960;**20**:37–46.
12 Fleiss JL. *Statistical methods for rates and proportions.* 2nd ed. New York: John Wiley, 1981.
13 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74.
14 Lee FD, Burnett RA. Quality assurance in histopathology. *J Pathol* 1987;**152**:247–51.
15 Evans DND, Hudson EA, Brown CL, *et al.* Terminology in gynaecological cytopathology: report of the Working Party of the British Society for Clinical Cytology (Review Article). *J Clin Pathol* 1986;**39**:933–44.
16 Beck J Swanson and Members of MRC Breast Tumour Pathology Panel. Observer variability in reporting of breast lesions. *J Clin Pathol* 1985;**38**:1358–65.

17 Dundas SAC, Kay R, Beck JS, *et al.* Can histopathologists reliably assess neoplasia in chronic inflammatory bowel disease? *J Clin Pathol* 1987;**40**:1282–6.
18 Kjaersgaard-Andersen P, Christensen F, Schmidt SA, Pedersen NW. A new method of estimation of interobserver variation and its application to the radiological assessment of osteoarthrosis in hip joints. *Statistics in Medicine* 1988;7:639–47.
19 Deknezian R, Chen X, Kuo T, Ordonez N, Katz RL. DNA hybridisation for human papilloma virus (HPV) in cervical lesions. *Arch Pathol Lab Med* 1987;**111**:22–7.
20 Jenkins DJ, Tay SK, Maccox MH. Routine papilloma virus antigen staining of cervical punch biopsy specimens. *J Clin Pathol* 1987;**40**:1212–6.
21 Sanders DSA, Kerr MA, Hopwood D, Coghill G. Expression of the CD 15 antigen is a marker of cellular differentiation in cervical intraepithelial neoplasia (CIN). *J Pathol* 1988;**155**: 207–12.

Requests for reprints to: Professor J Swanson Beck, Department of Pathology, Ninewells Hospital and Medical School, PO Box 120, Dundee DD1 9SY, Scotland.