Department of
Pathology, University
of Wales College of
Medicine, Heath Park,
Cardiff, CF4 4XN, UK
A G Douglas-Jones
J M Morgan

Department of
Pathology, Worthing
Hospital, Worthing
West Sussex,
BN11 2DH, UK
M A C Appleton

Department of
Pathology, Llandough
Hospital, Penarth,
CF64 2XX, UK
R L Attanoos
M Cotter
N S Dallimore

Department of
Pathology, Ysbyty
Gwynedd,
Penrhosgarnedd,
Bangor, Gwynedd
LL57 2PW, UK
A Caslin
M Hughes

Department of
Pathology, East
Glamorgan General
Hospital, Church
Village, Nr
Pontypridd,
CF38 1AB, UK
C S Champ
D Stock
A Verghese

Department of
Pathology, Morriston
Hospital, Swansea,
SA6 6NL, UK
A P Griffiths
C O'Brien
A Dawson

Department of
Pathology, Royal
Gwent Hospital,
Cardiff Road,
Newport, Gwent,
NP9 2UB, UK
A M Rashid
R W Fortt

Department of
Pathology, Royal
United Hospital,
Combe Park, Bath,
BA1 3NG, UK
P A Kitching

Department of
Pathology, Prince
Philip Hospital,
Llanelli, Dyfed,
SA14 8QF, UK
D W Williams

Department of
Pathology, Singleton
Hospital, Swansea,
SA2 8QA, UK
N W Williams
S Williams

Correspondence to:
Dr Douglas-Jones
email: ag.dj@virgin.net

Accepted for publication
7 March 2000

Consistency in the observation of features used to classify duct carcinoma in situ (DCIS) of the breast

A G Douglas-Jones, J M Morgan, M A C Appleton, R L Attanoos, A Caslin, C S Champ, M Cotter, N S Dallimore, A Dawson, R W Fortt, A P Griffiths, M Hughes, P A Kitching, C O'Brien, A M Rashid, D Stock, A Verghese, D W Williams, N W Williams, S Williams

Abstract

Aim—To determine interobserver and intra-observer agreement in the assessment of cytological grade and intraduct necrosis in pure duct carcinoma in situ (DCIS) of the breast.

Methods—Sixty unselected cases with illustrated diagnostic criteria were circulated to 19 practising histopathologists.

Results—Overall agreement was moderate for cytological grade in three categories: 71% agreement; weighted κ (κ_w), 0.36; intraduct necrosis in three categories (absent, present, extensive): 76% agreement; κ_w , 0.57; and the Van Nuys classification system: 73% agreement; κ_w , 0.48. Agreement was no better among observers participating in the National External Quality Assurance Programme. Intra-observer agreement for cytological assessment (69.6% agreement; κ_w , 0.52) and intraduct necrosis (68.3% agreement; κ_w , 0.48) was moderate, suggesting that individual variation rather than precision of criteria contributes to the lack of agreement.

Conclusions—Moderate agreement on observations can be achieved by non-specialist pathologists, with better agreement on necrosis than cytological grade. There was evidence of consistent individual bias towards over or under scoring cytological grade, which could be corrected with adequate and prompt feedback.

(*J Clin Pathol* 2000;53:596–602)

Keywords: classification; duct carcinoma in situ; breast; intraduct necrosis

Duct carcinoma in situ (DCIS) has been classified according to architecture,¹ cytological appearance,² presence of necrosis,³ or a combination of these features.⁴ A classification of DCIS should be simple, quick, and reproducible, with low interobserver variation. All interpretations of microscopic features are subjective, but precisely defined discriminating features should result in less interobserver variation. The definitions used in our study are based on those published after the consensus conference on the classification of DCIS.⁵

The aim of our study was to assess interobserver variation in the observation of individual components used in the classification

of DCIS, and to determine which features gave rise to the greatest problems in discrimination.

The results of cytological grade alone served to assess the interobserver variation in the National Health Service Breast Screening Programme (NHS BSP) grading system,⁶ which is similar to the classification proposed by Holland *et al.*²

In the Van Nuys system, group 1 comprises low grade (cytology 1 or 2) without necrosis, group 2 comprises low grade (cytology 1 or 2) with necrosis (minimum of debris with five pyknotic nuclei), and group 3 comprises cytology grade 3 with or without necrosis.⁴

Consistency of observation (intra-observer variation) was assessed by circulation of the same slides to all the observers without their knowledge.

Materials and methods

CASE MATERIAL

Pathologists were invited to participate and to submit two unstained sections from 10 cases of pure DCIS from their routine practice. The slides were stained centrally to ensure uniform staining and all slides were checked for suitability for inclusion in our study. Slides were batched as sets of 10 cases. Sixty cases of pure DCIS were submitted. A seventh set was assembled by taking spare sections at random from the other six sets to assess intra-observer variation.

The material was unselected and unmarked 5 μ m haematoxylin and eosin stained sections were used to simulate routine practice.

TRAINING

The observers were 19 qualified practising histopathologists. Some of the participants (11 of 19) in this study receive the National Breast External Quality Assessment (EQA) slide circulations as part of their contracts to the NHS BSP. These participants were designated as "specialists", who should be familiar with grading of invasive breast carcinomas and cytonuclear grading of DCIS, and have received feedback of peer group opinion. Specialist and non-specialist histopathologists were compared to determine whether interobserver agreement is more dependent on the observer or on the precision of the criteria.

OBSERVATIONS

Participants were requested to make two sets of observations on each case. One slide of each

Table 1 Summary of criteria used for cytonuclear grading

Cytological grade	Pleomorphism	Size	Max. diam. of lymphocyte nuclei	Mitoses	Polarity
Low	Slight	Small	<3	Few	Present
Intermediate	Moderate	Moderate	≤4	Present	Residual polarity
High	Extensive	Large	≥5	Easy to find	None

Important note: ignore architectural pattern.
Max. diam. lymphocyte nuclei refers to the maximum size of nuclei compared with lymphocyte nuclei.

case was circulated, but there was no marking or masking of any part of the slide to draw attention to any particular area.

HETEROGENEITY
It has been recognised that the cytonuclear features of DCIS can be heterogeneous; that is, that populations of cells with different cytonuclear grades can be present. Observers were asked to state whether or not they considered the cell population to be heterogeneous. No criteria for the diagnosis of heterogeneity were circulated, but if the cell population was considered to be heterogeneous observers were asked to state which cytonuclear grades were present. For all analyses the highest cytonuclear grade was used.

CYTONUCLEAR GRADE
Nuclear grades were defined as in the pathology reporting guidelines for the NHS BSP⁶ and in a protocol given to each participant (box 1; table 1). Photomicrographs illustrating the three cytological grades were sent to each participant (fig 1).

NECROSIS
Necrosis was defined as absent, present, or extensive “comedo” according to criteria laid out in box 2, and these instructions were accompanied by black and white photomicrographs to illustrate the lesions (fig 2).

Box 1 Recommended criteria for the cytological grading of ductal carcinoma in situ (DCIS)

Important note: ignore architectural pattern.

The nuclear grade is defined as low, intermediate, or high according to the variability of both size and shape of carcinoma nuclei.

HIGH NUCLEAR GRADE
High nuclear grade DCIS has pleomorphic, irregularly spaced nuclei, with a very irregular nuclear outline and multiple nucleoli. The largest nuclei present are more than five times the diameter of lymphocyte nuclei and mitoses are frequently present.

INTERMEDIATE NUCLEAR GRADE
These lesions have mild to moderate pleomorphism but are lacking the uniformity of low nuclear grade lesions. The largest nuclei present are up to four times the diameter of lymphocyte nuclei.

LOW GRADE
Low grade lesions have small, rounded, and uniform nuclei, with the largest nuclei present being no more than three times the diameter of adjacent lymphocyte nuclei.

POLARITY
The orientation of tumour cells around or towards luminal spaces (that is, polarisation of cells) is also assessed. In the cases where it is difficult to distinguish between nuclear grades 1 and 2 and between nuclear grades 2 and 3, presence of polarity is used to assign the DCIS to the lower (better differentiated) grade.

VAN NUYS CLASSIFICATION
This was derived by the project coordinator (AGD-J) from the observations of cytonuclear grade and necrosis, cytonuclear grades 1 and 2 being regarded as low grade and cytonuclear

Figure 1 Photomicrographs of examples of three cytonuclear grades of ductal carcinoma in situ (DCIS) as circulated to participants in conjunction with written descriptions. (A) Low grade, (B) intermediate grade, (C) high grade.

Figure 2 Photomicrographs of examples of necrosis as circulated to participants in conjunction with written descriptions. (A) Intraluminal macrophages, not to be regarded as necrosis; (B) minimum lesion for necrosis present: eosinophilic debris with five or more pyknotic nuclei; (C) duct containing ductal carcinoma in situ (DCIS) with less than 50% diameter showing necrosis: insufficient for extensive “comedo” necrosis; (D) duct containing DCIS with more than 50% diameter showing necrosis: designated as extensive “comedo” necrosis.

grade 3 as high grade.⁴ It is acknowledged that this derived classification may not reflect how

the pathologists themselves would have applied the Van Nuys classification de novo.

Participants were asked to record the time spent assessing the slides and were urged not to spend excessive time looking at them, but were asked to regard the sections as they would routine work.

Box 2 Recommended criteria for the assessment of intraduct necrosis in ductal carcinoma in situ (DCIS)

Important note: ignore architectural pattern.

Necrosis is defined as eosinophilic cellular debris containing five or more pyknotic nuclei. Calcification is regarded as surrogate necrosis if it is accompanied by five or more pyknotic nuclei. The lesions are classified into three categories.

NO NECROSIS

If no duct in any of the sections shows necrosis then the case is classified as without necrosis. The presence of intraluminal macrophages is not regarded as necrosis.

NECROSIS PRESENT

If any duct in any of the sections shows any necrosis an assessment is made regarding the percentage diameter of the duct showing necrosis. If less than 50% of the diameter of any duct shows necrosis it is designated as necrosis present.

EXTENSIVE “COMEDO” NECROSIS

If 50% or more of any duct diameter contains material or calcification with evidence of necrosis (five or more pyknotic nuclei present within the duct) it is designated as extensive necrosis. The presence of five pyknotic nuclei with eosinophilic debris, calcification, or shrinkage artefact is regarded as necrosis, and is scored according to the diameter of duct involved.

STATISTICAL ANALYSIS

κ Statistic

Agreement between observers was determined using the *κ* statistic with correction for expected chance agreement (unweighted *κ*, *κ*_u). Because disagreement by one grade is less important than two grades, the weighted *κ* (*κ*_w) was also calculated, which compensates for degree of error in category assessment.^{7 8} Using this method, each observer was compared with every other, and *κ* values were recorded for each pairing. An overall value for *κ*_u was calculated using the method of Fleiss.⁹ An overall *κ*_w was calculated as the arithmetic mean of individual *κ*_w values (171 pairs of comparisons). The level of agreement achieved as indicated by a particular *κ* was interpreted using a subjective scale as follows: 0.8–1, excellent; 0.6–0.8, good; 0.4–0.6, moderate; 0.2–0.4, fair; and 0–0.2, poor.

Table 2 Interobserver agreement for three categories: *κ* values

	Cytology		Necrosis		Van Nuys	
	<i>κ</i> _u	<i>κ</i> _w	<i>κ</i> _u	<i>κ</i> _w	<i>κ</i> _u	<i>κ</i> _w
<i>kappa</i>						
Overall	0.36	0.47	0.47	0.57	0.40	0.48
Specialists	0.36	0.46	0.43	0.53	0.39	0.48
Non-specialists	0.38	0.49	0.51	0.61	0.44	0.51

*κ*_u, unweighted *κ*; *κ*_w, weighted *κ*.

Table 3 Interobserver agreement for three categories: percentage agreement

Category	Cytoneuclear			Necrosis			Van Nuys		
	1	2	3	1	2	3	1	2	3
Consensus	15	21	24	14	7	39	13	23	24
All agree	285	399	456	266	133	741	247	437	456
No disagree	75	137	119	24	44	169	49	138	136
% Disagreement	26	34	26	16	33	23	20	32	30
Overall % disagreement	28.6			24.0			27.3		
% Agreement	74	66	74	84	67	77	80	68	71
Overall % agreement	71.4			76			72.7		

Table 4 Interobserver agreement for two categories: κ values

	Cytology low (1+2) v high	Necrosis present v absent (1) v (2+3)	Cytology low (1) v high (2+3)	Extensive necrosis "comedo" (1+2) v 3
Overall	0.39	0.64	0.57	0.50
Specialists (n = 11)	0.37	0.62	0.58	0.45
Non-specialists (n = 8)	0.43	0.66	0.56	0.56

Cytology, high v low.
Necrosis, present/absent.
 κ , unweighted κ .

Percentage agreement

The consensus opinion was that given by most of the observers. In those cases in which the consensus was cytological grade 1 (n = 14), total agreement would produce 14 × 19 (266) grade 1 opinions. If there are 75 observations that are other than grade 1, the per cent disagreement is $(75/266) \times 100 = 28\%$ (agreement 72%). This method of measuring the level of agreement does not take into account the contribution of chance alone.

Bias

The consistency of performance of individual observers was examined in two ways. All instances where the individual disagreed with the consensus opinion were identified and scored as +1 if the case was placed in a higher category (or +2 if placed in the highest category), -1 or -2 if in a lower category. The overscore and underscore events were summed to give an overall measure of individual bias. Hence, an individual who scored +9 and -1 appears to have a bias towards over scoring, whereas an individual who scored +3 and -2 is in good agreement with the consensus and shows no such bias.

Consistency

Intra-observer agreement was assessed by comparing the results of two separate observations on the same slide at different times (unknown to the observer) using a simple per cent agreement and the κ statistic.

Table 5 Intraobserver agreement for three categories: κ values and percentages

	Cytology			Necrosis			Van Nuys		
	κ κ_u	κ_w	% (SD)	κ κ_u	κ_w	% (SD)	κ κ_u	κ_w	% (SD)
Overall	0.47	0.52	69.6 (14)	0.4	0.48	68.3 (14)	0.58	0.63	77 (13.7)
Specialists	0.44	0.49	68.6 (14)	0.33	0.44	65 (13)	0.58	0.63	78.6 (9.5)
Non-specialists	0.49	0.55	70.9 (15)	0.45	0.53	72.8 (15)	0.57	0.62	75.4 (18.2)

κ_u , unweighted κ ; κ_w , weighted κ .

Results

INTEROBSERVER AGREEMENT

Data were available on cytological grade in three categories, necrosis in three categories, cytological grade in two categories (high and low), necrosis in two categories (present or absent), and cytological homogeneity in two categories (homogeneous or not). From the observations the Van Nuys classification of each case was derived as described above.

Pair wise interobserver comparisons of agreement for the cytological grade of DCIS were calculated (data not shown). A wide range of κ values ranging from very high (0.85, excellent agreement) to almost 0 (0.06, poor agreement) was obtained. The overall κ for all observers is 0.36, which represents only a fair degree of agreement. When the calculation of the κ values is weighted for degrees of disagreement the overall κ value (κ_w) becomes 0.47, reflecting a moderate degree of agreement.

Table 2 shows overall κ_w and κ_u values for cytological grade, necrosis, and Van Nuys classification, with data for specialists and non-specialists. The κ_w values are consistently higher than the unweighted values, reflecting the degree of error in category assignment. Table 3 shows the corresponding agreement of assignment into three categories expressed as percentages.

When the number of categories is reduced to two the κ_w values increase relative to those for three categories (table 4). The figures show that there is better agreement separating grade 1 cytology from higher grades ($\kappa_u = 0.57$) than grade 3 cytology from lower grades ($\kappa_u = 0.39$). There is greater agreement identifying cases with no necrosis ($\kappa_u = 0.64$) than identifying those with extensive (comedo) necrosis ($\kappa_u = 0.50$).

INTRA-OBSERVER AGREEMENT

When the same slides were assessed on two separate occasions by one observer there was only moderate agreement (κ_w for cytology, 0.47; κ_w for necrosis, 0.40; κ_w for Van Nuys, 0.58) (table 5).

Overall mean agreement for cytological assessment was 70% (SD, 14%), with a range of 50–100%, and for assessment of necrosis 68% (SD, 14%), with a range of 50–90%.

HETEROGENEITY

The mean cytological grade of DCIS was recorded as heterogeneous in 21% (SD, 16%) of cases (overall mean with a wide range of 0–63%).

Figure 3 Observer bias: histograms showing overgrading and undergrading for 60 cases on (A) cytonuclear grade and (B) intraduct necrosis for each observer.

The level of agreement for the assessment of cytological heterogeneity was poor (κ , 0.105; range 0.52 to -0.15; data not shown).

BIAS

Figure 3 shows the analysis of observer bias. In fig 3A the positive scores for the assessment of cytology (assignments in higher grades than the consensus) are plotted in ascending order for all the observers with the corresponding negative scores. Corresponding scores for necrosis are plotted in fig 3B.

Discussion

The aim of our study was to assess interobserver variability of routine unselected cases of DCIS by a group of general histopathologists assessing the slides in their own laboratories under normal working conditions. The study was designed to look at interobserver variability in the assessment of nuclear grade and necrosis. The criteria to be used in the assessment were circulated to all participants, accompanied by black and white photomicrographs illustrating typical appearances. For cytology, the criteria were those of the NHS BSP, which are already used in routine screening practice in the UK.⁶ Criteria for the assessment of necrosis are less well defined and none are universally accepted. The minimum lesion accepted as evidence of necrosis in our study is the presence of eosinophilic debris with at least five pyknotic nuclei, corresponding to the "punctate" necrosis of other authors.⁵ Intraluminal macrophages and eosinophilic secretions were specifically mentioned as not representing necrosis. Extensive necrosis or "comedo-type necrosis" was defined as the presence of eosinophilic debris, containing at least five pyknotic nuclei and occupying more than 50% of the diameter of one or more duct structures, corresponding to the "linear intraductal necrosis" of other authors.⁵ These criteria for necrosis were used to allow assessment of small numbers of duct structures and to avoid the necessity for percentage criteria, which are more difficult and complex to apply.¹⁰

The κ statistic has been adopted as a means of quantifying the degree of interobserver variation. A κ value of 1 indicates perfect agreement

between all the observers and a κ value of 0 or below indicates a level of agreement that could be achieved by chance. It is important to note that the κ statistic depends on the number of observers participating and categories being used (in most studies of this kind, three). Thus when comparing κ statistics from different studies the figures must be regarded as general indicators of agreement, and the numbers of participants must be borne in mind.

In our study, the architecture of the DCIS was ignored because it is well recognised that DCIS is frequently architecturally heterogeneous, and there is no simple system to record this heterogeneity in a reproducible manner. Previous studies on reproducibility of architectural classification in DCIS have shown poor agreement between observers: even though a κ statistic of 0.44 was achieved in one study classifying DCIS by architecture, the overall κ score was 0.23.¹¹⁻¹³

In our study the κ statistic for agreement between 19 pathologists on cytological grading into three categories in unselected cases simulating routine practice was 0.36, with a κ of 0.47 (70% agreement). For necrosis using three categories (no necrosis, minimal necrosis (punctate), and extensive (comedo) necrosis), κ was 0.47, with a κ of 0.57 (75% agreement). These findings indicate that variation in the recognition and categorisation of necrosis using the criteria defined here is less than that for the grading of cytology. This is despite the fact that most pathologists are more familiar with cytonuclear grading (as used in the grading of breast carcinoma) than they are in the assessment of necrosis. Agreement among specialists was no better than among the non-specialists. There is no evidence in our study of more disagreement between cytological grades 1 and 2 than between grades 2 and 3, in contrast to previous work indicating more disagreement between grades 1 and 2.¹⁰

A further feature incorporated into our study was the assessment of the cytological heterogeneity of the DCIS. Observers were asked to state whether they thought that the cytology of the lesion was homogenous and, if not homogenous, they were asked to indicate

which nuclear grades were present. It has been recognised that architectural heterogeneity is a problem in the assessment of DCIS, and it has been suggested that there is less cytological heterogeneity in DCIS. Two previous studies have shown identical results, with architectural heterogeneity in 61% of cases of DCIS and cytological heterogeneity in 16%^{14,15}; however, there are no data on observer variation. In our study, the overall frequency of heterogeneity was 20% (close to the 16% quoted in other studies), but with a range of individual opinion from 0% to 63%, and agreement on heterogeneity was poor (mean κ , 0.105 with a wide range of 0.52 to -0.15). No precise guidelines for the diagnosis of heterogeneity were given, and some participants may have used this option as a means of indicating their uncertainty as to how the lesion should be graded cytologically. A working definition for heterogeneity could be the presence of ducts containing separate and distinct cell populations of different cytonuclear grade within one area of DCIS. For the purposes of the analysis, the higher grade was taken as the reading for that case, in accordance with the criteria set out by Holland, who recognised the presence of cytological heterogeneity.² These findings indicate that cytological heterogeneity in DCIS requires further evaluation, by means of microdissection and genetic analysis, to prove whether or not morphologically distinct populations are different clones of cells.

There have been several studies of interobserver variation in the grading of invasive breast carcinomas. This has led to precise criteria for the assessment of tubules, cellular pleomorphism, and mitotic counts in invasive breast carcinoma, and κ statistics have been derived for agreement in the observation of each of these categories. The κ value measuring interobserver variation in the assessment of cellular pleomorphism in invasive breast carcinoma has been quoted at 0.40, which is very close to the figure in our study; κ statistics were 0.64 and 0.52 for tubules and mitoses, respectively.¹⁶ In the assessment of intraobserver variation in the grading of invasive cancers with experienced observers using precise grading guidelines, an agreement in the range 80–84% was considered to be the optimum achievable.¹⁷

Several other studies have examined the assessment of interobserver variation in the classification of DCIS. The design of these studies has varied in detail and none has looked at the interobserver variation in the observation of cytological heterogeneity, assessment of necrosis, or intraobserver variation.

Earlier work on agreement between two observers at the same institution had indicated most disagreement using the traditional architectural classification (34.4% disagreement) and least using extent of necrosis (17.7% disagreement). We modified the criteria for extent of necrosis to avoid percentage criteria and to allow assessment of necrosis on a small number of ducts. There was 30.5% disagreement using cytonuclear features (Holland) alone. Of the classifications studied, the Van Nuys was associated with the least disagreement (21.1%).⁷

These figures are very similar to those in our study (table 3), which involved larger numbers of observers from different institutions, although there is more disagreement on necrosis (24.0%) and Van Nuys (27.3%).

In the study of Bethwaite *et al*,¹¹ 11 pathologists examined selected cases in which slides were masked to focus the pathologists' attention on individual areas of the slide, and they admitted that their κ statistics might be better than those achievable in routine practice.¹¹ Cases used were associated with an invasive component, whereas the cases in our study were pure DCIS. Agreement was poorest using the architectural scheme (κ , 0.44) and was better using the cytonuclear grade (κ , 0.57), with most consistency achieved using the Van Nuys scheme (κ , 0.66). They identified the recognition of intraluminal necrosis as a source of disagreement in the Van Nuys scheme, which is not supported by our findings. The agreement on absence of necrosis in our study was 84% (κ (two category), 0.64) (tables 3 and 4).

In another study by Sneige *et al*,¹⁸ six experienced breast pathologists in four institutions examined 125 DCIS lesions. A training set of 12 slides was circulated with criteria, and working sessions were used for education on the criteria. Slides were marked, and in this study the overall κ for cytological grading was 0.46, with a range of 0.30–0.61; the κ value for grade 1 versus 2 was 0.29, and the κ value for grade 2 versus 3 was 0.48. The assessment of the extent of necrosis in this study was based on percentage criteria (extensive, > 33% of ducts containing necrosis; focal, < 33% of ducts containing necrosis). No data on interobserver variation for necrosis were available in this study.¹⁸

Sloane *et al*,¹² reporting the results of the national EQA circulations to 250 pathologists in the UK, gave a κ statistic of 0.23 for agreement on architectural type. There are no data on assessment using newer classifications in this publication. In this paper, grading of invasive breast carcinomas yielded a κ statistic of 0.46 for the coordinating group of experienced specialist pathologists, whereas the κ statistic for grading among NHS BSP participating pathologists (designated as "specialists") was 0.26. There is some evidence from the analysis of national EQA statistics that participation in such EQA schemes can improve diagnostic consistency.¹²

Our results showed no difference between the 11 participants who had taken part in the UK EQA study and the eight who had not.

Unknown to them, participants were sent a set of 10 slides that they had assessed before to assess intra-observer variation. Overall, there was approximately 70% agreement with the first observation for both cytological grade (mean, 69.6%; SD, 14%; range, 50–100%) and necrosis (mean, 68.3%; SD, 14%; range, 50–90%). The results presented as κ values indicate only moderate consistency when the same observer reads the same section on two separate occasions (table 5). It is important to note that these κ values are not directly comparable with those given in table 2 because of the difference in numbers of cases observed

(60 v 10) and the number of observers involved (19 v two).

Our study included an analysis of the individual performance of the participants compared with the consensus reading of each case. Those who undergraded from the consensus scored -1 or, if they were two points below the consensus, -2; similarly, they scored +1 or +2 for readings above the consensus. This allows analysis of individual bias and whether this bias is consistent or not. Hence, in fig 3, a pathologist with a score of 0 agrees with the consensus on every case. A pathologist with small but equal negative and positive scores is generally close to the consensus, and without bias, and a pathologist with a high score above and below shows no bias but is randomly disagreeing with the consensus. A pathologist with a high positive score and low negative score appears to be consistently overgrading, whereas one with a small positive score and large negative score is consistently undergrading. For cytological grading, fig 3A shows that some individuals (observers 3, 6, 7, 15, 19) appear to be consistently overgrading and others consistently undergrading (observers 4, 5, 9, 10, 13, 14, 17). These data suggest that with further training, experience, and feedback it may be possible to improve agreement by correcting the overscorers and underscorers. By contrast, in assessment of necrosis there is less overdiagnosis (maximum positive score of 9), and a wide range of underscoring representing failure to find the ducts containing necrosis.

For individual observers there was no significant correlation between overscoring or underscoring bias in cytological grading and necrosis assessment, $r(0.05)(19) = 0.333$.

The highest κ between two individual observers of 0.85 was achieved by two pathologists working in the same department who, when questioned, said that they worked very closely together in routine practice, often presenting each others work at meetings. This observation also suggests that good agreement is possible given sufficient continuous feedback. This is a possible mechanism by which participation in EQA circulations improves consistency, but the feedback is less effective in this situation, because the number of relevant cases is small and feedback of peer opinion may be delayed for weeks or months.

Participants were asked to record the time spent looking at the sections to assess the practicality of incorporating these scoring systems into routine practice. The overall mean time spent on 10 slides was 20 minutes (SD, 5.9; range, 10–30). This shows that assessment of a case takes between one and three minutes (average, two minutes) under normal reporting conditions, which would be acceptable in routine practice.

Summary and conclusions

(1) The classification of necrosis into three simple categories (absent, present, and extensive) using precisely defined criteria shows the least interobserver variation, overall κ of 0.57 (75% agreement).

- (2) The Van Nuys classification system shows the least interobserver variation, with overall κ of 0.48 (73% agreement).
- (3) Intraobserver variation in observation of features of DCIS has been assessed for the first time and individual observers show only moderate agreement with themselves (for cytonuclear features: κ = 0.52; 70% agreement; for necrosis: κ = 0.48; 68% agreement). These data indicate that intraobserver variation in slide reading contributes to interobserver variation.
- (4) There is evidence for a consistent individual observer bias in cytological grading, which might be reduced through continuous feedback.
- (5) The agreement for recognition of cytonuclear heterogeneity is poor (κ = 0.105) and its evaluation requires further study.
- (6) The time to report cytonuclear grade and assessment of necrosis into three categories is about two minutes for each case.
- (7) Levels of agreement obtained by "specialists" using marked slides can be achieved by non-specialists in routine practice.

We are indebted to Dr C Richardson, Mr M Stanton, and Mr N Garrahan for valuable assistance in the preparation of the figures.

- 1 Bellamy CO, McDonald C, Salter DM, et al. Non-invasive ductal carcinoma of the breast: the relevance of histological categorisation. *Hum Pathol* 1993;24:16–23.
- 2 Holland R, Peterse JL, Millis RK, et al. Ductal carcinoma in situ: a proposal for a new classification. *Semin Diagn Pathol* 1994;11:167–80.
- 3 Poller DN, Silverstein MJ, Galea M, et al. Ductal carcinoma in situ of the breast: a proposal for a new simplified histological classification association between cellular proliferation and c-erbB-2 protein expression. *Mod Pathol* 1994;7:257–62.
- 4 Silverstein MJ, Poller DN, Waisman JR, et al. Prognostic classification of breast ductal carcinoma in situ. *Lancet* 1995;345:1154–7.
- 5 Consensus conference on the classification of ductal carcinoma in situ. *Cancer* 1997;80:1798–802.
- 6 National Coordinating Group for Breast Screening Pathology. Classifying malignant non-invasive lesions. In: *Pathology reporting in breast cancer screening*, 2nd ed. Sheffield: NHSBSP Publications, 1997:23–7.
- 7 Altman DG. Some common problems in medical research. In: *Practical statistics for medical research*. London: Chapman and Hall, 1991:404–9.
- 8 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1997;33:159–74.
- 9 Fleiss JL. The measurement of interrater agreement. In: *Statistical methods for rates and proportions*. New York: John Wiley and Sons, 1981:212–36.
- 10 Douglas-Jones AG, Gupta S V, Attanoos RL, et al. A critical appraisal of six modern classifications of ductal carcinoma in situ of the breast: correlation with grade of associated invasive carcinoma. *Histopathology* 1996;29:397–410.
- 11 Bethwaite P, Smith N, Delahunt B, et al. Reproducibility of new classification schemes for the pathology of ductal carcinoma in situ of the breast. *J Clin Pathol* 1998;51:450–4.
- 12 Sloane JP, Ellman R, Anderson TJ, et al. Consistency of histopathological reporting of breast lesions detected by screening: findings of the UK national external quality assessment scheme (EQA). *Eur J Cancer* 1994;30:1414–19.
- 13 Sloane JP, Amendoreia I, Apostolikas N, et al. Consistency achieved by 23 European pathologists in categorising ductal carcinoma in situ of the breast using five classifications. *Hum Pathol* 1998;29:1056–62.
- 14 Harrison M, Coyne JD, Gorey T, et al. Comparison of cytomorphological and architectural heterogeneity in mammographically detected ductal carcinoma in situ. *Histopathology* 1996;28:445–50.
- 15 Quinn CM, Ostrowski JL. Cytological and architectural heterogeneity in ductal carcinoma in situ of the breast. *J Clin Pathol* 1997;50:596–9.
- 16 Frierson HF, Wolber RA, Berean KW, et al. Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histological grading system for infiltrating duct carcinoma. *Am J Clin Pathol* 1995;103:195–8.
- 17 Robbins P, Pinder S, De Klerk N, et al. Histological grading of breast carcinomas: a study of interobserver agreement. *Hum Pathol* 1995;26:873–9.
- 18 Sneige N, Lagios MD, Schwartz R, et al. Interobserver reproducibility of the Lagios nuclear grading system for ductal carcinoma in situ. *Hum Pathol* 1999;30:257–62.