Pergamon

European Journal of Cancer Vol. 30A, No. 10, pp. 1414–1419, 1994

0959-8049/94 $7.00+0.00

0959-8049(94)00261-4

Clinical Papers

Consistency of Histopathological Reporting of Breast Lesions Detected by Screening: Findings of the U.K. National External Quality Assessment (EQA) Scheme

U.K. National Coordinating Group for Breast Screening Pathology
J.P. Sloane, R. Ellman, T.J. Anderson, C.L. Brown, J. Coyne, N.S. Dallimore, J.D. Davies, D. Eakins, I.O. Ellis, C.W. Elston, S. Humphreys, D. Lawrence, J. Lowe, J. O'D. McGee, R.R. Millis, J. Nottingham, N. Ryley, D.J. Scott, J.M. Sloan, J. Theaker, P.A. Trott, C.A. Wells and H.D. Zakhour

The aim of the scheme was to determine consistency of histopathological reporting in the United Kingdom National Breast Screening Programme. This external quality assessment scheme involved 51 sets of 12 slides which were circulated to 186–251 pathologists at intervals of 6 months for 3 years. Participants recorded their diagnoses on standard reporting forms, which were submitted to the U.K. National Cancer Screening Evaluation Unit for analysis. A high level of consistency was achieved in diagnosing major categories of breast disease including invasive carcinoma and the important borderline lesions, radial scar and ductal carcinoma *in situ* (DCIS), the latter exceeding a national target set prior to the onset of the scheme. Atypical hyperplasia (AH) was reported with much less consistency although, where it was the majority opinion, over 86% of diagnoses were of benign disorders and only 14% were of DCIS. Inconsistency was encountered in subtyping and measuring DCIS, the former apparently due to current uncertainties about classification and the latter to poor circumscription, variation in size in different sections and merging with zones of AH. Reporting prognostic features of invasive carcinomas was variable. Measurement of size was achieved with adequate consistency except in a small number of very poorly circumscribed tumours. Grading and subtyping were inconsistent although the latter was not specifically tested and will be the subject of future study. Members of the National Coordinating Group achieved greater uniformity than the remainder of the participants in all diagnostic categories, but both groups experienced similar types of problem. Our findings suggest that participation in the scheme improves diagnostic consistency. In conclusion, consistency in diagnosing invasive carcinoma and radial scar is excellent, and good in DCIS, but improvements are desirable in diagnosing atypical hyperplasia, classifying DCIS and reporting certain prognostic features of invasive tumours. Such improvements will require further research, the development of improved diagnostic criteria and the dissemination of clearer guidelines.
Eur J Cancer, Vol. 30A, No. 10, pp. 1414–1419, 1994

INTRODUCTION

ALTHOUGH THE primary method of mass screening for breast cancer is radiological, the quality of the pathology services is of crucial importance. There is now widespread acceptance that the use of cytology at the assessment stage, together with clinical examination and imaging, can improve the level of confidence about benign diagnoses and reduce the number of unnecessary surgical operations [1]. The quality of the histopathology service is also important not only to determine if excised mammographic abnormalities are benign or malignant, but also to characterise carcinomas with respect to features of prognostic significance. These features may be important in determining patient management and are useful in monitoring breast screening programmes, as reduction in mortality is reflected by more favourable prognostic features of the tumours detected. It is not likely to be known

Correspondence to J.P. Sloane at the Department of Histopathology, Royal Marsden Hospital, Sutton SM2 5PT, U.K.
Received 8 April 1994; accepted 6 June 1994.

for a number of years whether mortality will be affected by the U.K. Screening Programme. Specimens from screened women have presented histopathologists with greater problems of macroscopic and histological examination than those from symptomatic patients. The former are related to the large number of impalpable abnormalities which necessitate the use of specimen radiography, and the latter to the higher incidence of certain 'borderline' lesions, such as ductal carcinoma *in situ* (DCIS) and radial scar.

With these considerations in mind, a working group was set up jointly by the National Health Service Breast Screening Programme (NHSBSP) and the Royal College of Pathologists at the beginning of the national programme to examine the issue of quality assurance in breast screening pathology. A report was published in the form of two documents, *Guidelines for Pathologists* [2] and *Pathology Reporting in Breast Cancer Screening* [3]. The former stated four major objectives for a quality assurance programme: (1) to improve the identification of lesions producing mammographic abnormalities, (2) to improve the consistency of diagnoses made by pathologists, (3) to improve the quality of prognostic information in pathological reports, (4) to maximise the number of people in whom an accurate diagnosis can be made without an open biopsy. The first objective was addresed in *Pathology Reporting in Breast Cancer Screening* by giving advice on macroscopic examination of specimens from screened women, and the last by recommending the expansion of cytological services accompanied by the necessary training programmes.

Objectives 2 and 3 are concerned largely with improving reporting consistency. To this end, a standard reporting form was devised which would ensure that the same pathological data were collected from each woman using the same terminology, which was defined in *Pathology Reporting in Breast Cancer Screening*. Having taken this initiative, some mechanism was required to assess the degree of consistency among pathologists involved in the screening service. An external quality assessment (EQA) scheme was therefore set up, and the findings of the first 3 years are described in this communication.

DESIGN OF THE SCHEME

After a preliminary, restricted circulation, three sets of 12 slides were sent, at 6-monthly intervals, to each of 17 regional coordinators; one for each of 14 English health regions and one each for Scotland, Wales and Northern Ireland. Two cases were selected from each of the following diagnostic groups chosen because they were considered to represent the major categories of breast disease and/or problems particularly likely to be associated with screening: (1) benign, not otherwise specified, (2) radial scar/complex sclerosing lesion, (3) atypical hyperplasia, (4) *in situ* carcinoma, (5) invasive carcinoma 10 mm or less, (6) invasive carcinoma larger than 10 mm. The blocks were provided by the regional coordinators and selected randomly within the six categories as the first two screened cases to be received by their centres after a specified date.

Fifty-three sections were cut from each block and numbered consecutively in the order they were cut. Every fifth slide, in addition to the first and last, was examined to ensure that the appearances did not change significantly. The first and fifty-third were retained for reference and the remainder sent in groups of three to the regional coordinators, noting the level numbers.

The coordinators then circulated the sets of slides to as many pathologists in their region as possible over a 3-month period. Participants reported the sections anonymously using a form (Figure 1) derived from the reporting form used for the U.K. National Breast Screening Programme. The completed forms were returned to the Cancer Screening Evaluation Unit at Sutton, Surrey. Six rounds had been completed prior to this report, and the number of participants, excluding coordinators, rose from 186 in round 1 to 251 in round 6.

Nine cases were initially selected but later excluded from circulation for the following reasons. Three radial scars and one case of DCIS were too small to obtain 53 sections and in 1 case of atypical ductal hyperplasia and 1 case of DCIS it proved impossible, partly because of calcification, to obtain sections of adequate quality. Three invasive carcinomas (two less and two more than 10 mm) were not used as the blocks as they were considered too thin to obtain the required number of sections.

Results of the coordinators were compared with those obtained for the remainder of the participants. The former group consisted of the 17 regional coordinators mentioned above and four other members of the National Coordinating Group for Breast Screening Pathology. This group thus comprises pathologists, many of whom have a particular interest and experience in breast screening pathology.

Statistical analysis

For each slide, the majority opinion of the coordinators was taken to be the diagnostic category used most frequently on the report forms. (This was not necessarily the same as the majority opinion of all readers, nor the consensus verdict of the coordinators after discussion, nor was it necessarily the median diagnosis. In fact, these different methods of choosing a standard would have given a different result on two of the cases.) The categories of individual diagnoses were then compared with the majority opinion for each slide and the proportion of slides on which each participant agreed with the majority opinion was calculated. A low score could be due either to random error or bias. To investigate bias towards under- or overdiagnosis of malignant neoplasia, a score was calculated as follows, taking the example of a case where the majority diagnosis was *in situ* carcinoma: those reporting *in situ* carcinoma scored 0, microinvasive carcinoma $+\frac{1}{2}$, invasive cardinoma $+1$, atypical hyperplasia -1, radial scar or other benign condition -2. Similarly, for a radial scar, an individual assessing it to be invasive carcinoma would score $+3$. Scores of an individual were then added together and divided by the number of slides reported. A positive score thus indicates a tendency towards overdiagnosis of malignancy and a negative one a tendency towards underdiagnosis.

The consistency of the participants and of subgroups among them was assessed using kappa statistics, a measure of agreement which takes account of the amount of agreement expected due to chance. Cohen's kappa statistic [4] takes the value zero if agreement is no better than expected by chance, a value of one if agreement is perfect and a negative value if there is a consistent tendency to disagree. The kappa statistic is independent of any assumption about the true diagnosis. It has, however, the disadvantage that it is affected by the prevalence, giving an underestimate where the prevalence of the category is below 20% [5]. As a rough guide, it should be noted that Landis and Koch recommended that a kappa of less than 0.4 should be considered poor and a kappa of 0.75 should be considered extremely good [6], although higher values should be expected for the more straightforward than the borderline categories. Variation in results from different circulations is likely to be

BREAST SCREENING HISTOPATHOLOGY

Slide no.

Hospital

Histological diagnosis ☐ Normal ☐ Benign ☐ Malignant

For BENIGN lesions please tick the lesions present

participant's name or no.

☐ Fibroadenoma
Papilloma ☐ Single ☐ Multiple
☐ Complex sclerosing lesion/radial scar
☐ Other (please specify)

☐ 'Fibrocystic change'
☐ Solitary cyst
☐ Periductal mastitis/duct ectasia
☐ Sclerosing adenosis

EPITHELIAL PROLIFERATION

☐ Not present ☐ Present with atypia ('ductal')
☐ Present without atypia ☐ Present with atypia (lobular)

For MALIGNANT lesions please tick any of the following present

NON-INVASIVE

☐ Ductal Subtype ☐ Paget's disease ☐ Lobular

MICROINVASION

☐ Not present ☐ Possible ☐ Present

INVASIVE

☐ 'Ductal' (not otherwise specified) ☐ Tubular or cribriform carcinoma
☐ Medullary carcinoma ☐ Mucoid carcinoma
☐ Lobular carcinoma
☐ Other primary carcinoma (please specify)
☐ Other malignant tumour (please specify)

MAXIMUM DIAMETER (invasive component) mm (in-situ) mm

GRADE ☐ I ☐ II ☐ III ☐ not assessable VASCULAR INVASION? ☐ Present ☐ Not seen

COMMENTS/ADDITIONAL INFORMATION

Figure 1.

affected not only by increasing experience but also by chance variation in the characteristics of the particular slides included in each circulation.

In addition to kappas for individual categories, an overall kappa has been calculated weighting the kappa for individual categories by the proportion of reports in that category. Since the cases selected for circulation included an undue proportion of radial scars, atypical hyperplasias and *in situ* carcinomas, an adjusted kappa has also been calculated using as weights the proportions of each type of lesion as they present in screening work. This case-mix was assessed from the distribution of 4274 lesions reported from 10 of the regions for a period in which first-round screening predominated.

RESULTS

Major diagnostic categories

The level of consistency in diagnosing the major categories in all six circulations is shown in Tables 1 and 2.

In Table 1, the diagnosis made by the majority of the participants is shown in the left-hand column. The rows to the right of these diagnoses show the percentage of readings falling into all categories for each majority diagnosis. The numbers forming a diagonal line from the top left to the bottom right of the table (numbers in bold type) thus show the percentage of individual opinions forming the majority diagnoses. The majority diagnosis of benign, not otherwise specified (NOS) was thus formed by 88.6% of all individual opinions, radial scar by 74.8%, atypical hyperplasia by 41.9%, *in situ* carcinoma by 77.9% and invasive carcinoma by 92.1%.

In Table 2, the data are presented as kappa statistics for the individual categories as well as three major groups of benign, *in situ*/microinvasive and invasive carcinoma. The kappa values are consistently higher for coordinators. The values in parentheses have been adjusted for the case-mix expected in screening practice.

Table 3 shows under/overdiagnosis bias scores and shows that coordinators' scores ranged from −0.1 to +0.09. Among the non-coordinators, some showed a more pronounced bias towards overdiagnosis, and others towards underdiagnosis of malignancy.

Ductal carcinoma in situ

This was the majority diagnosis in 17 cases. Table 1 shows that combining all these cases, the majority diagnosis was selected by 77.9% of participants. The proportion favouring the majority diagnosis, however, varied between cases, from 34 to 97%. A comparison was made by one observer of the histological features of the cases where the majority was formed by more or less than 70% of participants. The most significant distinguishing feature of the cases where agreement was high was the presence of a comedo growth pattern in at least part of the lesion; large nuclear size, prominent nucleoli and low nucleo-cytoplasmic ratio were also characteristic of the group.

A low level of consistency was achieved in subtyping DCIS. Table 4 shows kappa values for the major subtypes; the highest value was obtained for comedo carcinoma at 0.44.

Wide ranges in size were reported for each of the 17 DCIS, but for eight (47%), 80% of the estimates were within ±3 mm of

Table 1. Distribution of individual opinions for slides classified according to the majority opinion

Majority diagnosis	Benign NOS (%)	Radial scar (%)	Atypical hyperplasia (%)	*In-situ* carcinoma (including microinvasive) (%)	Invasive carcinoma (%)	No. of cases	Total readings
Benign NOS	**88.6**	3.6	5.9	1.1	0.1	17	4105
Radial scar	13.1	**74.8**	8.3	0.9	2.9	10	2466
Atypical hyperplasia	37.4	7.0	**41.9**	13.6	0.1	3	701
In-situ carcinoma (including microinvasive)	6.3	0.5	10.5	**77.9**	4.7	21	5045
Invasive carcinoma	0.8	3.3	1.0	2.6	**92.1**	21	5228
						72	17545

In situ and microinvasive carcinoma are grouped together as the latter was never a majority diagnosis. Microinvasive carcinoma was defined, for the purposes of the U.K. Screening Programme, as a predominantly *in-situ* carcinoma with one or more foci of invasion, none exceeding 1 mm in maximum dimension.
NOS, Not otherwise specified.
The majority diagnosis is that of the coordinators.

Table 2. Kappa statistics after six circulations

Category	Coordinators		Non-coordinators	
Benign (NOS)	0.76		0.70	
Radial scar	0.78	0.84	0.63	0.79
Atypical hyperplasia	0.25		0.17	
In situ carcinoma	0.75	0.81	0.62	0.70
Microinvasive carcinoma	0.30		0.28	
Invasive carcinoma	0.94	0.94	0.83	0.83
Overall	0.75	0.86	0.64	0.78
	(0.85)	(0.89)	(0.75)	(0.85)

Values in brackets adjusted for screening case-mix.
NOS, not otherwise specified.

the median. More widely distributed readings were apparently explained by (1) significant changes in the size of the lesion from sections 1–53 (4 cases), (2) lack of a clear boundary, the DCIS merging with areas of atypical hyperplasia (3 cases), (3) the DCIS appeared as widely dispersed foci (1 case) and (4) the lesion was fairly well circumscribed but exhibited different dimensions in different axes (1 case).

Table 3. Bias scores indicating overdiagnosis/underdiagnosis of malignancy

Score*	Coordinators (%)	Non-coordinators (%)
−0.2 to −0.9	0	15.1
−0.1 to −0.19	27	20
−0.09 to +0.09	73	54
+0.1 to +0.19	0	5.4
+0.2 to 0.5	0	5.4

*Bias score = $\frac{\text{sum of deviations}}{\text{no. slides reported}}$.

Table 4. Kappa statistics after six circulations

DCIS subtype (all participants)			
NOS	0.09	Solid	0.22
Comedo	0.44	Mixed	0.06
Cribriform	0.22	Other	0.11
Papillary	0.30	Overall	0.23

DCIS, ductal carcinoma *in situ*; NOS, not otherwise stated.

Invasive carcinoma

Observer consistency in recognising the major subtypes of invasive carcinoma is shown in the form of kappa statistics in Table 5. Kappa values are low for all categories. There was no evidence of better agreement among coordinators than non-

Table 5. Kappa statistics for six circulations (all participants)

Invasive subtype			
Ductal NOS	0.21	Mucoid	0.00
Lobular	0.22	Medullary	0.31
Tubular	0.31	Other	0.05
		Overall	0.21

NOS, not otherwise specified.

Table 6. Kappa statistics after six circulations

Grade	Coordinators	Non-coordinators
1	0.58	0.36
2	0.40	0.18
3	0.38	0.21
Overall	0.46	0.26

coordinators. Table 6 shows that consistency of grading was only moderately good even among coordinators whose overall kappa score was 0.46.

As with DCIS, there was a wide scatter of size measurements in all cases, although histograms generally showed tight grouping and, in 13 of the 16 cases, over 80% of participants' measurements were within ±3 mm of the median value. All these cases exhibited a single focus of tumour of variable circumscription. In the remaining three cases, there were small microscopic foci of tumour lying outside the main tumour mass. In all 16 cases, the variation in size between the first and last sections to be cut from the block was no greater than 2 mm as measured by one observer.

Effect of participation on consistency

Table 7 provides evidence that consistency in diagnosing the six major categories improves as a consequence of participating in the scheme. In each circulation, agreement between new readers was less good than among those who had previously participated. In order to exclude the possibility that this difference was due to new participants making deviant diagnoses and subsequently leaving the scheme, old participants were also compared with new participants who persevered, i.e. subsequently became old participants. The differences in kappa values were largely maintained (Table 7).

DISCUSSION

The scheme described in this paper has shown that it is possible for large numbers of pathologists to achieve a very high or acceptable level of agreement in diagnosing the major categories of breast disease studied with the exception of atypical hyperplasia. More variability was encountered in reporting prognostic features of carcinomas. The measurement of tumour size was consistent where lesions were relatively well-circumscribed. Greater consistency was achieved by the 21 coordinators than by other participants, although both groups experienced similar types of problem. Participation in the scheme itself appeared to improve diagnostic consistency.

Table 7. Overall kappa statistics of new participants compared with old participants (excluding coordinators)

Circulation	New participants No.	New participants Kappa	Old participants No.	Old participants Kappa
1	185 (164)	0.53 (0.54)	15	0.58
2	64 (51)	0.52 (0.54)	144	0.59
3	29 (29)	0.65 (0.65)	187	0.74
4	33 (27)	0.69 (0.69)	212	0.72
5	15 (9)	0.61 (0.63)	230	0.69
6	11	0.55	240	0.57

Figures in parentheses refer to new participants who persevered with the scheme, i.e. subsequently became old participants.

The low level of agreement in diagnosing atypical hyperplasia is not surprising and in keeping with that encountered in most previous reports [7, 8]. Much greater concordance was achieved in one recent study [9], although this was at the expense of a lower level of agreement on *in situ* carcinoma than was found in the present communication. Furthermore, the study exhibited several major differences for the scheme described here: (1) only six pathologists, chosen by the scheme organiser, participated, (2) the cases were selected by the organiser according to their histological appearance and the technical quality of the sections, (3) slides were covered with masking tape so that only lesions of interest were visible, (4) participants were provided with a written summary of diagnostic criteria and a set of teaching slides prior to the study. The first three of these different aspects would be inappropriate for an EQA scheme, whose aim is to determine diagnostic consistency with a minimum of artificiality. The last, however, may be relevant. Although detailed diagnostic criteria for hyperplasia of usual type and ductal carcinoma *in situ* were laid down in the booklet, *Pathology Reporting in Breast Cancer Screening*, prior to the initiation of the scheme, atypical hyperplasia was simply defined as a group of intermediate proliferations which did not fit easily into either of the other two categories. Detailed architectural and cytological criteria for diagnosing atypical ductal hyperplasia have recently been published [10] and could form the basis of a more precise definition in the next edition of the booklet, an illustrated version of which is presently being prepared. An encouraging finding of the U.K. scheme, however, is that, although the level of agreement on atpyical hyperplasia was low, over 86% of diagnoses were benign where it formed the majority opinion (Table 1).

The majority diagnosis of ductal carcinoma *in situ* was formed by just under 78% of all opinions. Kappa values were 0.75 for coordinators and 0.62 for non-coordinators or 0.81 and 0.70, respectively, if microinvasive carcinomas were also included. These results are satisfactory given that a national target of 0.60 was set at the beginning of the quality assurance programme. Not surprisingly, the major benign diagnosis in this category was atypical hyperplasia, improved definition of which could also lead to greater consistency in diagnosing DCIS.

The proportion of opinions agreeing with the majority diagnosis of DCIS varied greatly from case to case from 34 to 97%. Only a preliminary attempt was made to identify histological features associated with diagnostic consistency but necrosis emerged as a very important factor. Certain cytological features were also associated with diagnostic consistency but were not independent of the presence of necrosis.

A low level of agreement was achieved in classifying DCIS, the highest kappa statistic being 0.44 for the comedo variant. Table 4 shows that lesions were classified solely according to growth pattern, but evidence is emerging that cytological features, regardless of growth pattern, may be associated with an increased risk of recurrence after treatment [11]. The criteria for classifying DCIS are being reviewed for the purposes of the EQA scheme with the intention of producing a classification of greater biological and clinical relevance and, hopefully, reproducibility.

Agreement on invasive carcinoma was, as expected, very good. Difficulty was occasionally encountered in distinguishing low grade carcinomas from radial scars (2.9% of all readings) and identifying small foci of invasion in what were mainly *in situ* carcinomas (4.7% of readings). Nearly 2% of diagnoses were,

however, benign for reasons which are difficult to explain by examining the sections. The possibility that at least some of the deviant diagnoses may be due to clerical error cannot be excluded.

In contrast, the level of consistency achieved in reporting prognostic factors of invasive carcinoma was disappointing. However, the kappa statistics for histological subtypes should be interpreted with some caution as the scheme has so far not specifically focused attention on this aspect, and several readers made no attempt to specify subtype. This will be the subject of further investigation.

Consistency of grading was low even among coordinators. This was a particularly disappointing finding given the precision with which the grading criteria were defined at the beginning of the scheme and the undoubted prognostic significance of grade in some studies [12]. It was not possible to determine whether lack of consistency is the result of inherent subjectivity of current methods of histological examination or of a failure on the part of participants to refer to the published criteria.

Unexpectedly, there were invariably large variations in tumour size measurements, although generally the majority of estimates were tightly clustered with just a few outlying results. Again, it was not possible to determine to what extent these wide variations were due to histological misinterpretation, mismeasurement or simply clerical error. Not surprisingly, greater variations were encountered in measuring *in situ* than invasive carcinomas for reasons outlined earlier. Measurements of invasive carcinomas were generally within acceptable limits (±3 mm—a range which grudgingly accepts readings which are rounded up or down to the nearest 5-mm mark), except where there was more than one focus of tumour. Clearer criteria and guidance for measuring tumours could result in greater consistency in determining size.

The diagnosis of radial scars was surprisingly consistent, the majority diagnosis being selected by nearly 75% of all individual opinions. Misdiagnosis of radial scars as low-grade carcinomas had been perceived as a significant problem of breast screening pathology by the working group prior to the commencement of the EQA scheme, but less than 3% of opinions fell into the category of invasive carcinoma where radial scar was the majority diagnosis. Over 96% of diagnoses were benign.

The scheme described in this report is a development of that funded by the Medical Research Council in association with a trial of early detection of breast cancer (TEDBC) in which kappa values of 0.16, 0.67 and 0.76 were obtained for atypical hyperplasia, DCIS/microinvasive carcinoma and invasive carcinoma, respectively [7]. All these values are lower than those achieved in the present U.K. national scheme even by the non-coordinators, an achievement which is all the more significant given that only nine pathologists participated in the TEDBC scheme and the same slides were examined by all participants. It thus appears that efforts in recent years from various quarters to improve diagnostic consistency in breast pathology have met with some degree of success. The findings also compare very favourably with those obtained from similar studies of observer consistency in reporting cervical and bladder biopsies, particularly as these schemes also involved small numbers of pathologists (12 or less) examining the same set of slides [13–15].

The scheme is, of necessity, artificial to some extent. Only one block per case was used and slides were unaccompanied by clinical details or information about macroscopic appearances. (This was felt not to be undesirable given that the scheme was designed primarily to investigate the reporting of histological features.) Participants were obliged to report specific categories, and there was no opportunity to express uncertainty. Furthermore, the sections were not cut and stained in the participants' own laboratories and may thus have exhibited appearances somewhat different from those to which they were accustomed. Finally, participants may not devote the same attention to EQA cases as to those which form part of their surgical workload. Nevertheless, the first six rounds of the scheme have generated important information about consistency of reporting breast specimens in the U.K. National Breast Screening Programme and suggest that participation itself can improve diagnostic consistency.

1. Lamb J, Anderson TJ, Dixon MJ, Levack PA. Role of fine needle aspiration cytology in breast cancer screening. *J Clin Pathol* 1987, **40**, 705.
2. Royal College of Pathologists Working Party. *Guidelines for Pathologists*. Screening Publications ISBN 1 871997 65 8.
3. Royal College of Pathologists Working Party. *Pathology Reporting in Breast Cancer Screening*. Screening Publications, ISBN 1 871997 70 4.
4. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Measurements* 1960, **20**, 37–46.
5. Altman DG. *Practical Statistics for Medical Research*. Chapman and Hall, 1990, 403–409.
6. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977, **33**, 159–174.
7. Swanson Beck J and members of the Medical Research Council Breast Tumour Pathology Panel. Observer variability in reporting of breast lesions. *J Clin Pathol* 1985, **38**, 1358–1365.
8. Rosai J. Borderline epithelial lesions of the breast. *Am J Surg Pathol* 1991, **15**, 209–221.
9. Schnitt SJ, Connolly JL, Tavassoli FA, *et al.* Interobserver reproducibility in the diagnosis of ductal proliferative breast lesions using standardized criteria. *Am J Surg Pathol* 1992, **16**, 1133–1143.
10. Page DL, Rogers LW. Combined histologic and cytologic criteria for the diagnosis of mammary atypical ductal hyperplasia. *Human Pathol* 1992, **23**, 1095–1097.
11. Lagios MD. Duct carcinoma *in situ*. *Surg Clin North Am* 1990, **70**, 853–871.
12. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology* 1991, **19**, 403–410.
13. Ismail SM, Colclough AB, Dinnen JS, *et al.* Observer variation in histopathological diagnosis and grading of cervical intraepithelial neoplasia. *Br Med J* 1989, **298**, 707–710.
14. Robertson AJ, Anderson JM, Swanson Beck J, *et al.* Observer variability in histopathological reporting of cervical biopsy specimens. *J Clin Pathol* 1989, **42**, 231–238.
15. Robertson AJ, Swanson Beck J, Burnett RA, *et al.* Observer variability in histopathological reporting of transitional cell carcinoma and epithelial dysplasia in bladders. *J Clin Pathol* 1990, **43**, 17–21.

Acknowledgements—In addition to all the participants, the authors thank the staff of the Department of Histopathology, Royal Marsden Hospital, Sutton for technical and secretarial support and members of the National Cancer Screening Evaluation Unit for their help in the data analysis. We are indebted to Dr J.A. Harvey for his comments. Funding of the scheme was provided by the National Health Service Breast Screening Programme.