J Clin Pathol 1985;**38**:1358–1365

Observer variability in reporting of breast lesions

J SWANSON BECK AND MEMBERS OF THE MEDICAL RESEARCH COUNCIL BREAST TUMOUR PATHOLOGY PANEL ASSOCIATED WITH THE UNITED KINGDOM TRIAL OF EARLY DETECTION OF BREAST CANCER

From the Epidemiology Section, Institute of Cancer Research, Sutton, Surrey

SUMMARY The consistency of histological diagnosis of breast lesions by members of the panel was studied using a simplified six point classification system that covered the range from normal tissue to invasive carcinoma. In a representative set of 40 sections of the range of breast disease included in the Trial for Early Detection of Breast Carcinoma (the consecutive series), there was, overall, 82% agreement between the diagnoses submitted by members of the panel: diagnosis was virtually consistent (99% agreement) for frankly invasive carcinoma, but there was much greater variability in the diagnosis of borderline lesions. Diagnostic consistency was greatly improved when two categories only (benign and malignant) were considered (94% agreement). The diagnostic consistency of individual pathologists was studied by recirculating the sections: the overall agreement was 78% and, again, most of the inconsistencies occurred when borderline lesions were diagnosed.

Additional studies were undertaken on borderline lesions that had been flagged during the first two years of the trial (the borderline series). Unsurprisingly, there was less agreement among the pathologists when all six diagnostic categories were used (70% and 77% for the specimens from the first and second years, respectively), but consistency was greatly improved when classification was simplified to either benign or malignant (86% and 91%, respectively).

The diagnoses submitted by individual pathologists were found to deviate from the "majority" diagnosis to a relatively minor extent: each pathologist was generally consistent in either underdiagnosing or overdiagnosing in both the consecutive and borderline series.

Breast cancer has a poor prognosis if neoplastic cells have disseminated widely before treatment, but has a relatively favourable prognosis if, at presentation, the tumour is small and the axillary lymph nodes are not affected. There has been widespread interest in developing screening programmes to attempt to diagnose the disease at a presymptomatic stage when it should be "earlier" and therefore more amenable to treatment. Two screening trials suggested that mortality from breast cancer could be reduced by about one third,[1,2] and further evidence suggesting a similar benefit was reported from Holland.[3] The Department of Health and Social Security and the Scottish Home and Health Department set up a large multicentre trial to evaluate the alternative approaches of clinical and mammographic screening and of breast self examination (BSE).[4] In two health districts every woman aged 45 to 64 was invited to be screened annually for seven years; in two further districts women in the same age range were invited to classes to learn breast self examination and were provided with self referral clinics; and in four districts no special services for early diagnosis of breast cancer were provided, so that the results could be compared. In all eight districts records are kept on all breast tissue examined histologically from the study population.

Clearly, the Trial for Early Detection of Breast Carcinoma is heavily dependent on the validity of diagnoses of breast lesions in the participating districts: the difficulties are aggravated by each district having its own group of diagnostic histopathologists with differing previous experience in breast disease. At the start of the trial, a panel of nine pathologists from eight districts was established, with the support of the Medical Research Council to agree on a common classification of breast pathology, to study variability in reporting histological diagnoses, and to work towards improving consistency. To evaluate the results of screening it is particularly important to agree on the classification of lesions that fall between benign disease and cancer—for example, if one district has a systematic bias towards classifying benign epithelial dysplasia as non-invasive carcinoma—comparison of survival of patients with carcinoma in this district with

Accepted for publication 20 August 1985

that in others will give spuriously favourable results. Although this potential bias can be overcome by studying breast cancer mortality in the study population rather than fatality or survival in those with breast cancer, it is still necessary to assess the incidence of non-invasive and invasive cancerous lesions, as one of the side effects of screening may be over selection of non-progressive lesions. It is therefore imperative that efforts should be made to establish the confidence that can be given to the histological reporting of breast lesions by pathologists and that their consistency should be studied. This paper describes the results of studies of consistency that have been carried out as part of the Trial of Early Detection of Breast Cancer.

Material and methods

CLASSIFICATION OF BREAST HISTOPATHOLOGY

Since the planning stages of the trial nine participating pathologists from the eight districts have met regularly to discuss problems of breast pathology and approaches to classification of breast disease, and two pilot studies of observer variability have been carried out. It was agreed that a simplified classification of disease was preferable because of the wide range of alternative diagnostic labels and lack of exclusive diagnostic criteria for all cancer types.[5] Breast disease was classified into 19 categories, which were condensed into the six categories: normal, benign calcification, benign neoplasm, "other benign" (A); benign epithelial dysplasia (B); proliferative epithelial dysplasia with cellular atypia suggestive of borderline malignancy (C); non-invasive lobular carcinoma in situ (D); non-invasive Paget's disease, non-invasive intraduct carcinoma, microinvasive carcinoma (E); infiltrating carcinoma (F).

SELECTION OF HISTOLOGICAL MATERIAL

To obtain a measure of the consistency of histological reporting across the whole spectrum of biopsied breast disease occurring among women in the age range studied, the first survey done by the panel included typical specimens collected from all the districts. The participating pathologists in each centre were instructed to select a representative section from the first five consecutive breast biopsies submitted to the laboratory after a given date from women aged 45 to 64. This yielded 40 sections that were considered to be a representative sample of the breast disease occurring in the patient group of the trial. These 40 sections comprised the consecutive series.

These 40 typical slides were assembled by the coordinating centre for the trial, relabelled with a code number, and circulated in batches of five to all the participating pathologists. Each pathologist was asked to report each batch of slides within one week of receipt, to post the slides to the next participant, and to send his reports to the coordinating centre. Six months later the same 40 slides were renumbered and circulated in the same way. The pathologists were not informed that the slides in the previous circulation were being recirculated with different code numbers. These two circulations enabled an analysis of both interobserver and intraobserver agreement on the consecutive series of 40 typical slides.

It was clearly impossible to arrange for the slides of every woman in the trial who had a breast biopsy to be assessed by the whole panel in this way. It was agreed, however, that all reports of borderline lesions classified by the original pathologist as proliferative dysplasia with cellular atypia suggestive of borderline malignancy; non-invasive carcinoma; or microinvasive carcinoma; (categories C, D, and E) should be "flagged" and that typical slides from all of these borderline lesions should be independently assessed by each member of the panel. These slides were therefore routinely assembled by the coordinating centre, relabelled with a code number, and circulated in batches to members of the panel. Those seen by the panel in the first year of this system (borderline series 1) were mixed with the second circulation of the consecutive series; those in the second year were circulated alone (borderline series 2). The results of these two borderline series are reported separately overleaf. For both the consecutive and borderline series, the term "individual category of diagnosis" was used to denote the category in which each slide was placed by each pathologist before any discussion of the relevant slides by the panel had taken place.

STATISTICAL ANALYSIS

For each slide, the "majority diagnosis" was taken to be that which occurred most often on the nine report forms. (Other methods of estimating the true diagnosis were, of course, possible. One that was considered was the use of the median diagnosis for each slide; in fact, this differed from the majority diagnosis for only two of the circulated slides). The categories of the individual diagnoses were then compared with the majority category for each slide to give a measure of the agreement among observers. For the consecutive series this analysis was done separately for each of the two circulations, and each observer's diagnosis of each slide in the two circulations were also compared to give a measure of the consistency of repeat readings. Finally, the performance of individual pathologists was measured by calculating the deviation of each of his or her diagnoses from the majority. The data were categorical, and although the categories had been ranked according to increasing severity, the extent of the differences—for example, between catego-

Table 1 *Individual diagnoses compared with majority diagnosis for first circulation of consecutive series (figures in parentheses are numbers per cent)*

Individual category of diagnosis	*A*	*B*	*C*	*D*	*E*	*F*	*Total*
A	59 (82)	13 (10)	0	0	0	0	72 (20)
B	13 (18)	95 (70)	2 (11)	0	1 (6)	0	111 (31)
C	0	17 (13)	8 (44)	0	1 (6)	0	26 (7)
D	0	6 (4)	0	0	0	0	6 (2)
E	0	4 (3)	6 (33)	0	16 (89)	1 (1)	27 (8)
F	0	0	2 (11)	0	0	116 (99)	118 (33)
Total	72	135	18	0	18	117	360
Total No of slides	8	15	2	0	2	13	40

ries 1 and 2 and categories 3 and 4 were not necessarily the same. Consequently, the use of parametric statistics was not strictly valid, and the mean deviation and the variance were used merely as descriptive statistics and not for formal testing. A positive mean deviation indicated a tendency towards overdiagnosis of malignancy and a negative mean deviation towards a tendency underdiagnosis. The variance gave an indication of the extent of deviation from the majority in either direction. The results of individual pathologists were not disclosed to the participants.

A measure of overall agreement that does not require the use of a majority or true diagnosis is the kappa statistic,[6] which incorporates a correction for the amount of agreement to be expected by chance. Kappa statistics can be calculated both for comparison between two and between many observers; for more than two categories an overall value of kappa is defined as a weighted average of the kappa values for individual categories.[7] Standard errors have been calculated according to the formulae derived by Fleiss *et al.*[8] It has been suggested that a kappa value of 0·75 or above should be taken as an indication of excellent agreement, 0·4 to 0·75 as fair to good, and a value of less than 0·4 as poor agreement.[9]

Table 2 *Individual diagnoses of benign or malignant compared with majority diagnosis for first circulation of consecutive series (figures in parentheses are numbers per cent)*

Individual category	*Majority category*	
	Benign A to C	*Malignant D to F*
Benign a to c	207 (92)	2 (1)
Malignant d to f	18 (8)	133 (99)
Total	225 (100)	135 (100)
Total No of Slides	25	15

Results

CONSECUTIVE SERIES

Table 1 shows the diagnoses reported by individual pathologists compared with the majority diagnosis for the first circulation. Overall, the percentage agreement was 82% (the sum of agreement on the diagonal of the Table), and the kappa value was 0·62 ($p < 0·001$), indicating good agreement. Very good agreement was obtained for the diagnosis of invasive carcinoma (category F) as 99% of individual reports agreed with this majority diagnosis; and for recognised non-invasive carcinomas (category E) as there was 89% agreement. Similarly, at the "benign" end of the scale 82% of individual reports agreed with the majority for category A and 70% for category B. The greatest variability was in the reporting of the borderline state of cellular atypia (category C); 33% of individual reports overrated these cases as non-invasive carcinoma and a further 11% as invasive carcinoma. To assess the

Table 3 *Individual diagnoses compared with majority diagnosis for second circulation of consecutive series (figures in parentheses are numbers per cent)*

Individual category of diagnosis	*A*	*B*	*C*	*D*	*E*	*F*	*Total*
A	62 (86)	11 (9)	0	1 (6)	0	0	74 (21)
B	10 (14)	87 (69)	0	3 (17)	1 (6)	1 (1)	102 (28)
C	0	18 (14)	0	5 (28)	0	2 (2)	25 (7)
D	0	2 (2)	0	8 (44)	2 (11)	0	12 (3)
E	0	8 (6)	0	1 (6)	15 (83)	5 (4)	29 (8)
F	0	0	0	0	0	118 (94)	118 (33)
Total (= 100%)	72	126	0	18	18	126	360
Total No of slides	8	14	0	2	2	14	40

importance of these results in terms of patient management categories A–C were grouped as "benign" and categories D–F as "malignant" (Table 2). An overall percentage agreement of 94% with a kappa value of 0·82 ($p < 0·001$) showed that excellent agreement was found.

Tables 3 and 4 give the same results for the second circulation of the consecutive series. The distribution of the majority diagnosis had shifted somewhat towards the more malignant end of the scale, but the levels of agreement for individual and grouped categories were very similar to those in the first circulation (81% and 94%, respectively), with kappa values of 0·62 and 0·81 (Table 3). The increased reporting of category D, lobular carcinoma in situ, may well have been due to discussion by the panel of this category, which took place between the two circulations.

Table 4 *Individual diagnoses of benign or malignant compared with majority diagnosis for second circulation of consecutive series (figures in parentheses are numbers per cent)*

Individual category	*Majority category*	
	Benign A to C	*Malignant D to F*
Benign A to C	188 (95)	13 (8)
Malignant D to F	10 (5)	149 (92)
Total	198 (100)	162 (100)
No of Slides	22	18

Table 5 shows the consistency of each pathologist in reaching the same diagnosis on rereading the same sections in the two circulations of the consecutive series. One pathologist left the panel during the interval between the two circulations, so that this table refers to only eight readings of 40 slides. The overall agreement was 78%, and here again it was clear that the greatest inconsistency occurred in categories C and D.

BORDERLINE SERIES 1

Not surprisingly, among this series of slides deliberately selected for their closeness to the borderline between benign and malignant, there was less agreement among pathologists. Table 6 shows the comparison of individual categories and majority categories (overall agreement 70%, kappa value 0·39). Table 7 shows the resulting summarised classifications into benign or malignant (overall agreement 86%, kappa value 0·56).

BORDERLINE SERIES 2

Tables 8 and 9 give the same results for the second borderline series of cases. The overall agreement of individual and malignant categories was 77% (kappa value 0·48), with 91% agreement on the summarised classifications of benign and malignant (kappa value 0·72).

Table 5 *Repeat diagnoses by same reader in two circulations of the consecutive series (figures in parentheses are numbers per cent)*

Category of diagnosis in second circulation	*Category of diagnosis, first circulation*						
	A	*B*	*C*	*D*	*E*	*F*	*Total*
A	55 (81)	10 (11)	2 (11)	0	0	0	67 (21)
B	13 (19)	67 (72)	8 (44)	4 (33)	3 (13)	0	95 (30)
C	0	9 (10)	7 (9)	2 (17)	3 (13)	1 (1)	22 (7)
D	0	1 (1)	1 (6)	4 (33)	0	0	6 (2)
E	0	6 (6)	0	2 (17)	15 (63)	3 (3)	26 (8)
F	0	0		0	0	3 (13)	101 (96)104
(33)							
Total (= 100%)	68	93		18	12	24	105320

(Centre 5 omitted due to change of pathologists).

Table 6 *Individual diagnoses compared with majority diagnosis for borderline series I (figures in parentheses are numbers per cent)*

Individual category	*Majority category*					
	B	*C*	*D*	*E*	*F*	*Total*
A	4 (6)	1 (6)			3 (17)	8
B	48 (76)	3 (17)	6 (17)	1 (1)	3 (17)	61
C	7 (11)	10 (56)	1 (3)	6 (8)		24
D	1 (2)	4 (22)	23 (64)		1 (6)	29
E	1 (2)		4 (11)	57 (79)	5 (28)	67
F	2 (3)		2 (6)	8 (11)	6 (33)	18
Total (= 100%)	63	18	36	72	18	207
Total No of slides	7	2	4	8	2	23

Table 7 *Individual diagnoses of benign or malignant compared with majority diagnosis for borderline series 1 (figures in parentheses are numbers per cent)*

Individual category	*Majority category*	
	Benign (B and C)	*Malignant (D to F)*
Benign (A to C)	73 (90)	20 (16)
Malignant (D to F)	8 (10)	106 (84)
Total = 100%	81	126
Total No of slides	9	14

PERFORMANCE OF INDIVIDUAL PATHOLOGISTS

Tables 10 and 11 show how far each of the individual pathologists agreed with the majority diagnosis, firstly for the consecutive series (second circulation, Table 10) and secondly, for the two sets of borderline series combined (Table 11). In most cases the same pathologists tended to underrate (indicated by a negative mean deviation) or overrate (positive mean deviation) in both series. The variances in the borderline series were generally higher than those in the consecutive series, indicating a greater extent of deviation from the majority.

AGREEMENT ON INDIVIDUAL CATEGORIES

Table 12 shows the kappa values for individual categories for observations from the consecutive series (second circulation) and the two borderline series combined. All the values differed significantly from zero ($p < 0.001$), but as suggested by the earlier analyses, agreement on categories C and D was poor. Agreement on categories A, B, and E was good and that on category F excellent. The kappa value for agreement between a diagnosis of benign and malignant from these data was 0·73.

PREDICTIVE VALUE OF INDIVIDUAL DIAGNOSES

From these data it was possible to calculate the predictive value of an individual diagnosis—the probability that it agreed with the majority verdict. From the two circulations of the consecutive series (Tables 1 and 3) 352 of 359 individual diagnoses in categories A and B concurred with the majority, giving an individual A or B diagnosis a predictive value of 98%. The predictive value of a category C diagnosis was calculated from the two borderline series (Tables 6 and 8); 49 of 62 individual diagnoses in category C agreed with the majority, giving a predictive value of 79%. Similarly, for malignant diagnoses, 250 of 271 individual diagnoses in categories D and E concurred with the majority (predictive value 92%) (Tables 6 and 8), 234 of 236 individual diagnoses in category F concurred with the majority (predictive value 99%) (Tables 1 and 3).

These predictive values can be applied to the average distribution of categories in the two circulations of the consecutive series to indicate the implications of these findings in normal practice. If one assumes that the majority verdict is correct, Table 13 shows that the predictive value of an individual benign diagnosis being correct is 97·2% and that of an individual malignant diagnosis 97·9%.

These calculations refer to the reporting performance of members of the panel and the exact predictive values cannot be compared with those of other pathologists; it would, however, be expected that the general pattern would be universally applicable.

Discussion

Several studies have looked at the variability of histopathological and cytological reporting of neoplasia and dysplasia in different sites.[10-13] From these it is

Table 9 *Individual diagnoses of benign and malignant compared with majority diagnosis for borderline series 2 (figures in parentheses are numbers per cent)*

Individual category	*Majority category*	
	Benign (B and C)	*Malignant (D to F)*
Benign (A to C)	109 (87)	12 (6)
Malignant (D to F)	17 (13)	186 (94)
Total = 100%	126	198
Total No of slides	14	22

Table 8 *Individual diagnoses compared with majority diagnosis for borderline series 2 (figures in parentheses are numbers per cent)*

Individual category	*Majority category*					
	B	*C*	*D*	*E*	*F*	*Total*
A	6 (6)			3 (2)		9
B	66 (67)	5 (19)		3 (2)		74
C	20 (20)	12 (44)		6 (4)		38
D	2 (2)	1 (4)	4 (44)			7
E	4 (4)	8 (30)	3 (33)	152 (89)	1 (6)	168
F	1 (1)	1 (4)	2 (22)	7 (4)	17 (94)	28
Total (= 100%)	99	27	9	171	18	324
Total No of slides	11	3	1	19	2	36

clear that in normal practice there is bound to be some disagreement between individual pathologists, and even by one pathologist on different occasions, in the diagnostic interpretation of particular histological appearances. In histological reporting it is often difficult to decide on a yardstick against which to judge the validity of differing interpretations; therefore, the original diagnosis is sometimes used. The pathologist who originally reported on a surgical specimen dissected it and examined multiple histological sections; thus the originating pathologist had access to more material than the reviewing pathologists who were required to base their opinion on what was shown in the single section that was circulated. On the other hand, use of the majority diagnosis gives a more democratic verdict. The use of kappa statistics, which correct for chance agreement and do not rely on assumptions of a correct diagnosis, allows com-

Table 10 *Measures of agreement of individual pathologists with the majority diagnosis in second circulation of the consecutive series*

Pathologist	*Agreement (%) with majority (six categories)*	*Mean deviation*	*Variance*	*Agreement (%) with majority (benign/malignant)*
1	95	0·000	0·41	95
2	75	− 0·200	0·63	93
3	85	− 0·075	0·53	98
4	83	− 0·050	0·25	95
5	70	+ 0·250	0·65	90
6	90	+ 0·150	0·29	98
7	85	+ 0·200	1·09	88
8	90	+ 0·025	0·69	93
9	78	+ 0·000	0·31	95

Table 11 *Measures of agreement of individual pathologists with the majority diagnosis for borderline series 1 and 2 combined*

Pathologist	*Agreement (%) with agreement (six categories)*	*Mean deviation*	*Variance*	*Agreement (%) with majority (benign/malignant)*
1	78	− 0·153	0·75	92
2	76	− 0·136	1·05	92
3	66	− 0·000	0·79	85
4	70	− 0·390	1·28	85
5	78	+ 0·136	0·36	95
6	85	+ 0·169	0·32	97
7	71	+ 0·339	1·40	83
8	71	+ 0·119	1·56	88
9	75	− 0·119	1·21	88

Table 12 *Kappa statistics for six categories, consecutive series, and two borderline series combined*

	Category						*All categories*
	A	*B*	*C*	*D*	*E*	*F*	
Kappa	0·60	0·48	0·16	0·39	0·67	0·76	0·57

($p < 0·001$ for all values).

Table 13 *Predictive values of individual diagnoses*

Category	*No of slides*	*Predictive value*	*% Correct*	*% Incorrect*
Benign:				
A	8	98%		
B	15		97·2	2·8
C	1	79%		
Malignant:				
D	1	92%		
E	2		97·9	2·1
F	14	99%		

parisons to be made between the overall agreement in different series and in individual categories, although the statistics do not give any indication of the direction of disagreement, or which direction might be correct. Another method that has been used in other studies is to compare the original diagnosis with the diagnosis made by an expert group.[14]

The artificial conditions of consistency surveys may accentuate disagreement; they force participants to put each slide into a single category without any qualifying report on their confidence in this diagnosis. It is therefore reassuring to find that the pathologists comprising the panel were so consistently in accord in their decisions as to whether a lesion was benign or malignant, and in particular that there was excellent agreement on the diagnosis of invasive cancer. It is the borderline categories that cause most difficulty as indicated both by the lower kappa values for these lesions and by individual pathologists' greater variance from the majority in the borderline series. The higher level of agreement on borderline in the second year of the trial, however, may indicate that the panel's discussion of these difficult cases, (taking place regularly among all members of the panel with the aid of a projecting microscope) may lead to greater consistency.

Borderline cases form only a minority of breast lesions coming to biopsy in the trial but from the patients' viewpoint the distinction between categories C and D or E—the cut off point between being labelled as either a woman with an unimportant excised lesion of her breast or a patient with breast cancer and all its attendant threats—is of prime importance. In one large American screening study re-examination of the histology of borderline cases by an expert group of histopathologists led to the conclusion that 9% of women treated for non-infiltrating carcinoma did not, in fact, have a malignant lesion—that is, gave false positive results.[14]

Taking the predictive values of individual diagnoses in our study and assuming that the distribution of histology in the trial as a whole was similar to that in the consecutive series, around 2% of individual diagnoses of malignancy would be classified as benign by the panel's majority view (false positives) and around 3% of individual benign diagnoses classified as malignant by the majority (false negatives). This level of inconsistency with the majority view of the panel was small. Moreover, there are other reasons such as sampling errors in excision or sectioning[15] which may also contribute to diagnostic misclassification.

The estimates of possibly erroneous diagnoses are based on the distribution of different histological categories among the 40 typical slides submitted by participating laboratories for the consecutive series. These may not be representative of the cross section of cases in each individual centre, as the screening districts and, to a lesser extent, the breast self examination districts tend to have higher rates of borderline lesions than the comparison districts. These differences between centres will be the subject of future publications, which will also report on the consistency of reporting all the future borderline cases that are routinely flagged.

A fundamental problem with this study, as with similar studies of observer variability in histological reporting, was the lack of a yardstick against which individual reports could be compared. The majority diagnosis is not necessarily more correct than that of the individual. Moreover, the reasons why the majority may reach a different conclusion from the original or from other individual pathologists are by no means limited to different histological interpretations. There may be errors in choosing the most representative slide from the block, differences in the technical quality of slide preparation, clerical errors in labelling or circulation of slides, errors in completing standardised report forms, and errors in the data processing of these forms. Difficulties such as these were found in the American study in which of 66 cases originally reported to be "false positive" malignant verdicts, 16 were subsequently judged to have been correct when the case was examined in more detail.[14] The panel is now investigating the extent of these problems, focusing in the first instance on outlying cases in the surveys reported here. It is hoped that this will be of value in more clearly defining future studies of histology, including the panel's continuing review of borderline cases. Identifying and eliminating extraneous sources of disagreement, as well as the ongoing educative effect of panel meetings, will probably show that the eventual level of agreement is even greater than that reported here.

We thank Drs Joan Lamb and Robertson, who participated in reading slides from the trial and Miss Chandra Nagarajah for help with computing. We thank the various pathology departments that provided facilities for meetings of the panel, including University College Hospital, London, and the Royal Marsden Hospital. The panel is supported by the Medical Research Council and the Trial of Early Detection of Breast Cancer by the Department of Health and Social Security and by the Scottish Home and Health Department.

Membership of the panel: Professor J Swanson Beck (chairman), Dr TJ Anderson, Dr L Bobrow (until 1981), Dr P Burton, Dr JD Davies, Dr CW Elston, Professor NM Gibbs, Dr T Marshall, Dr C Wells (from 1982), Dr DP Wijesinghe, Dr J Chamberlain, Dr R Ellman (epidemiologists), Ms S Moss (statistician).

References

[1] Shapiro S, Venet W, Strax P, Venet L, Roeser R. Ten to fourteen year effects of breast cancer screening on mortality. *Journal of the National Cancer Institute* 1982;**69**:349–55.

[2] Tabar L, Gad A, Holmberg LH, *et al.* Reduction in breast cancer mortality by mass screening with mammography: First results of a randomised trial in two Swedish counties. *Lancet* 1985;i:829–32.

[3] Collette HJA, Day NE, Rombach JJ, de Waard F. Evaluation of screening for breast cancer in a non-randomised study (the DOM-Project) by means of a case-control study. *Lancet* 1984;i:1224–6.

[4] United Kingdom Trial of Early Detection of Breast Cancer Group. Trial of early detection of breast cancer: description of method. *Br J Cancer* 1981;**44**:618–27.

[5] Anderson JA, Fechner RE, Lattes R, Rosen PP, Toker C. *Lobular carcinoma in-situ (lobular neoplasia) of the breast.* In: Sommers SC, ed. New York: Appleton Century Crofts, 1980:193–223.

[6] Cohen J. A coefficient of agreement for nominal scales. *Educational Psychological Measurements* 1960;**20**:37–46.

[7] Fleiss JL. *Statistical methods for rates and proportions.* 2nd ed. Chichester: Wiley and Sons, 1981.

[8] Fleiss JL, Nee JCM, Landis JR. The large sample variance of kappa in the case of different sets of raters. *Psychol Bull* 1979;**86**:974–7.

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

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

[11] Husain OAN, Butler BE, Woodford FP. Combined external quality assessment of cytology and histology opinions: a pilot scheme for a cluster of five laboratories. *J Clin Pathol* 1984;**37**:993–1001.

[12] Stenkvist B, Bengstsson E, Eriksson O, Jarkraus T, Nardin B, Westerman-Naeser S. Histopathological systems of breast cancer classification: reproducibility and clinical significance. *J Clin Pathol* 1983;**36**:392–8.

[13] Riddell RH, Goldman H, Ransohoff DF, *et al.* Dysplasia in inflammatory bowel disease: standardized classification with provisional clinical applications. *Hum Pathol* 1983;**14**:931–68.

[14] Beahrs OH, Shapiro S, Smart C. Report of the working group to review the National Cancer Institute—American Cancer Society breast cancer detection demonstration projects. *Journal of the National Cancer Institute* 1979;**62**:640–709.

[15] Patchefsky AS, Potok J, Hoch WS, Libshitz HI. Increased detection of occult breast carcinoma after more thorough histological examination of breast biopsies. *Am J Clin Pathol* 1973;**60**:799–804.

Requests for reprints to: Dr J Chamberlain, Trial of Early Detection of Breast Cancer, Epidemiology Section, Institute of Cancer Research, Block D, Clifton Avenue, Sutton, Surrey SM2 5PX, England.