

Interobserver agreement and reproducibility in classification of invasive breast carcinoma: an NCI breast cancer family registry study

Teri A Longacre¹, Marguerite Ennis², Louise A Quenneville³, Anita L Bane^{4,5}, Ira J Bleiweiss⁶, Beverley A Carter⁷, Edison Catelano⁸, Michael R Hendrickson¹, Hanina Hibshoosh⁹, Lester J Layfield¹⁰, Lorenzo Memeo⁹, Hong Wu¹¹ and Frances P O'Malley^{4,5}

¹Department of Pathology, Stanford University, Stanford, CA, USA; ²Applied Statistician, Markham, Ontario, Canada; ³Department of Pathology, Jewish General Hospital, Montreal, Quebec, Canada; ⁴Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, Toronto, Canada; ⁵Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Canada; ⁶Department of Pathology, Mount Sinai Medical Center, New York, NY, USA; ⁷Department of Pathology, St Clare's Hospital, St Johns, Newfoundland, Canada; ⁸Department of Pathology, Cooper Health, Camden, NJ, USA; ⁹Department of Pathology, Columbia University, New York, NY, USA; ¹⁰Department of Pathology, University of Utah, Salt Lake City, UT, USA and ¹¹Department of Pathology, Fox Chase Cancer Center, Philadelphia, PA, USA

The United States National Cancer Institute Breast/Ovarian Cancer Family Registry is the largest international Registry of this type; over 37 724 individuals have been enrolled to date. One activity of this Registry is the semicentralized pathologic review of tumors from all probands. Given the semicentralized nature of the review, this study was undertaken to determine the reproducibility, source(s) of classification discrepancies and stratagems to circumvent discrepancies for histologic subtyping and grading of invasive breast cancer among the reviewing pathologists. A total of 13 pathologists reviewed 35 invasive breast cancers and classified them by primary and secondary histologic type, Nottingham grade and score. Lymph–vascular space invasion, circumscribed margins, syncytial growth and lymphocytic infiltrate were also evaluated. A training session using a separate set of slides was conducted prior to the study. General agreement, in terms of category-specific κ 's and percent agreement, and accuracy of classification relative to a reference standard were determined. Classification of histologic subtype was most consistent (and accurate) for mucinous carcinoma ($\kappa = 1.0$), followed by tubular ($\kappa = 0.8$) and lobular subtypes ($\kappa = 0.8$). Classification of medullary subtype was moderate ($\kappa = 0.4$), but additional evaluation of degree of lymphocytic infiltrate, syncytial growth and circumscribed margins identified most cases. Category-specific κ 's were moderate to good for Nottingham grade ($\kappa = 0.5–0.7$), with the greatest agreement obtained in categorizing grade I ($\kappa = 0.7$), and grade III tumors ($\kappa = 0.7$). A flexible classification strategy that employs individual and combined criteria provides good interobserver agreement for invasive breast cancers with uniform, unambiguous histology and compensates for classification discrepancies in the more histologically ambiguous or heterogeneous cancers.

Modern Pathology (2006) 19, 195–207. doi:10.1038/modpathol.3800496; published online 9 December 2005

Keywords: interobserver reproducibility; invasive breast cancer; familial breast cancer; breast/ovarian cancer family registry

The reproducibility of the classification and grading of invasive breast cancer and the cause(s) of interobserver disagreement among pathologists have

not been adequately evaluated. Prior studies evaluating interobserver concordance in categorizing breast lesions have documented improved diagnostic agreement when the pathologists involved used agreed-upon criteria,¹ but other potential sources of poor interobserver agreement, such as the difficulties in the application of the individual histologic criteria, the individual pathologist's variation in use of these criteria, and most importantly, the ambiguous or borderline and heterogeneous nature of the

Correspondence: Dr FP O'Malley, MB, FRCPC, Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, 600 University Ave., Toronto, Canada M5G1X5.

E-mail: fomalley@mtsinai.on.ca

Received 20 June 2005; revised 12 August 2005; accepted 13 August 2005; published online 9 December 2005

cases themselves have received less attention. Since it is unlikely that any significant pathologic differences between patient subgroups can be detected without accounting for the presence of tumor heterogeneity (the latter of which may well play a key role in determining variability in clinical behavior), it is important that the collection of pathologic data for cases accessioned into a breast cancer registry database incorporate this variability of tumor classification and grading into the classification scheme.

The Breast/Ovarian Cancer Family Registry is an international consortium that was initiated in 1995 and is supported through the United States National Cancer Institute (NCI). This group was established to provide a comprehensive infrastructure for interdisciplinary research studies of hereditary breast cancer. The participating sites include an Informatics Support Center (Irvine, CA, USA) and six Registry sites. There are three population-based sites: The Ontario Cancer Genetics Network, Cancer Care Ontario, Canada; Northern California Cancer Center, San Francisco and the University of Melbourne, Melbourne, Australia, and three clinical-based sites: Fox Chase Cancer Center, Philadelphia; the Huntsman Cancer Center, Salt Lake City; Columbia University, New York. Each site has at least one study pathologist and some sites also have pathology fellows involved in the review of cases. Currently, funded activities of the registry include establishment of common databases for family history, epidemiology, biospecimens and pathology. As of early 2005, 12 507 families and 37 724 individuals have been enrolled in this Registry.

The Pathology Working Group of the Breast/Ovarian Cancer Family Registry developed a pathology data collection and retrieval system for registry cases that would afford optimal diagnostic concordance among the participating registry sites without sacrificing potentially important information that could be obtained from cases with heterogeneous or borderline histologic features. An abbreviated version of the data collection system was then used to evaluate the registry group pathologists' diagnostic accuracy and reproducibility of invasive breast cancer classifications using an initial group training session with a standard set of slides, followed by individual assessment of a separate set of slides, using the agreed upon criteria. One of the registry pathologists assessed the study set of slides twice in order to establish a reference standard and ascertain the degree of intraobserver agreement of the chosen reference standard. Since one of the aims of the registry was to obtain an accurate assessment of familial breast cancer subtype(s) and to determine whether there are specific phenotypes of hereditary cancer, we also evaluated the flexibility of the pathology data collection system, given the presence of interobserver disagreement, in identifying all potential examples of these phenotypes.

Materials and methods

Study Design

To identify all areas in which potential diagnostic inaccuracy and/or poor interobserver reproducibility would engender significant misclassification of individual breast lesions subsequently enrolled in the registry, 'problematic' breast cancer slides were circulated and discussed at an initial meeting of the Pathology Working Group of the Breast/Ovarian Cancer Family Registry. Based on review and discussion of the problematic slides, a Registry Pathology Review form was specifically designed to capture all relevant pathologic findings in such a manner that 'borderline' or 'ambiguous' lesions could be identified and retrieved from the registry database without a laborious rereview of all the registry slides. For example, using agreed upon criteria, the form was designed to capture all potential medullary carcinomas of the breast entered into the registry database by a search for 'medullary carcinoma' and 'atypical medullary carcinoma', as well as 'ductal carcinoma, not otherwise specified', and restricting the latter to only those cases with marked lymphocytic infiltrate, presence of syncytial growth pattern or circumscribed margins. Similarly, all potential infiltrating lobular carcinomas could be retrieved by a global search for 'typical lobular carcinoma', 'pleomorphic lobular carcinoma' or 'mixed ductal and lobular cancer'. Since the form called for assignment of a primary and if present, a secondary pattern, even those cases in which only a small component of a particular histologic type was identified could be retrieved for future clinicopathologic, epidemiologic or basic research investigations. Following a trial use of the form by the members of the group, modifications were made and the form was standardized (Figure 1). Criteria used for scoring the individual components of the data form were based on published criteria.²

A study set of slides was selected to test the accuracy and reproducibility of invasive breast cancer diagnoses using the agreed upon classification system and data entry form. To assess the utility of the data entry form, 35 cases of primary invasive breast cancer were selected by the study group chair (FOM) from routinely processed archival cases accessioned during the same period as the cases enrolled in the registry. Cases were selected to highlight problem areas identified in the initial intergroup meeting. For the purposes of the study, the 'gold standard' or reference diagnosis was that rendered by the study chair. For each study case, a single set of 5- μ m-thick hematoxylin and eosin-stained sections was prepared, all by the same laboratory. Since in many instances, it is the submitting pathologist and not the actual registry pathologist, who selects the actual registry slide, it was concluded that a single representative slide for each of the cases most optimally simulated actual

PATHOLOGY REVIEW (NIH Breast Cancer Family Registry)

INVASIVE CARCINOMA - I: (Type of Invasive Cancer)

	Primary pattern	Secondary pattern
NST (NOS)	<input type="checkbox"/> (1)	NST (NOS) <input type="checkbox"/> (1)
Tubular	<input type="checkbox"/> (2)	Tubular <input type="checkbox"/> (2)
Cribriform	<input type="checkbox"/> (3)	Cribriform <input type="checkbox"/> (3)
Micropapillary	<input type="checkbox"/> (4)	Micropapillary <input type="checkbox"/> (4)
Mucinous	<input type="checkbox"/> (5)	Mucinous <input type="checkbox"/> (5)
Medullary		Medullary
Classic	<input type="checkbox"/> (7.0)	Classic <input type="checkbox"/> (7.0)
Atypical	<input type="checkbox"/> (7.1)	Atypical <input type="checkbox"/> (7.1)
Metaplastic	<input type="checkbox"/> (9)	Metaplastic <input type="checkbox"/> (9)
Other	<input type="checkbox"/> _____	Other <input type="checkbox"/> _____
Lobular	<input type="checkbox"/> (10.0), <input type="checkbox"/> (10.5)	Lobular <input type="checkbox"/> (10.0) classic, <input type="checkbox"/> 10.5: pleomorphic

INVASIVE CARCINOMA - II AND III: (Grade of Invasive Cancer + Micro Staging)

Nottingham Grade	I <input type="checkbox"/>	II <input type="checkbox"/>	III <input type="checkbox"/>	Score	/9
tubule formation	1 <input type="checkbox"/>	2 <input type="checkbox"/>	3 <input type="checkbox"/>		
nuclear pleomorphism	1 <input type="checkbox"/>	2 <input type="checkbox"/>	3 <input type="checkbox"/>		
mitotic score	1 <input type="checkbox"/>	2 <input type="checkbox"/>	3 <input type="checkbox"/>		
mitoses	/10 HPF		Field diam. (40X)	_____	
Lymphatic invasion	present <input type="checkbox"/>	absent <input type="checkbox"/>	N/A <input type="checkbox"/>		
Blood vessel invasion	present <input type="checkbox"/>	absent <input type="checkbox"/>	N/A <input type="checkbox"/>		
Invasive tumour necrosis	present <input type="checkbox"/>	absent <input type="checkbox"/>	N/A <input type="checkbox"/>		
Lymphocytic infiltration	marked <input type="checkbox"/>	moderate <input type="checkbox"/>	mild <input type="checkbox"/>	absent <input type="checkbox"/>	N/A <input type="checkbox"/>
Circumscribed margins	present <input type="checkbox"/>	absent <input type="checkbox"/>	N/A <input type="checkbox"/>		
Syncytial growth	present <input type="checkbox"/>	absent <input type="checkbox"/>	N/A <input type="checkbox"/>		
DCIS	present <input type="checkbox"/>	absent <input type="checkbox"/>	N/A <input type="checkbox"/>		
DCIS nuclear grade	1 <input type="checkbox"/>	2 <input type="checkbox"/>	3 <input type="checkbox"/>		
DCIS necrosis	present <input type="checkbox"/>	absent <input type="checkbox"/>	N/A <input type="checkbox"/>		
Extensive intraductal comp.	positive <input type="checkbox"/>	negative <input type="checkbox"/>	N/A <input type="checkbox"/>		
Margin positive - in situ	present <input type="checkbox"/>	absent <input type="checkbox"/>	N/A <input type="checkbox"/>		
Margin positive - invasive	present <input type="checkbox"/>	absent <input type="checkbox"/>	N/A <input type="checkbox"/>		
Other (tick present if comment is added)	present <input type="checkbox"/>	absent <input type="checkbox"/>	N/A <input type="checkbox"/>		
Comment:	_____				

NON-INVASIVE PROLIFERATIONS - I (Associated benign/preneoplastic lesions)

None

Non-proliferative FCC <input type="checkbox"/> 12.0	Proliferative disease without atypia (PDWA) <input type="checkbox"/> 11.0	ADH <input type="checkbox"/> 5.0
Columnar change <input type="checkbox"/> 12.2	Sclerosing adenosis <input type="checkbox"/> 11.2	ALH <input type="checkbox"/> 6.0
PAC <input type="checkbox"/> 12.3	Florid epithelial hyperplasia <input type="checkbox"/> 11.4	DIALH <input type="checkbox"/> 6.4
		LCIS <input type="checkbox"/> 1.0

NON-INVASIVE PROLIFERATIONS - II: (PAPILLARY NON-INVASIVE SPECTRUM)

Papilloma(s)	<input type="checkbox"/>	1.0
Fibroadenoma(s)	<input type="checkbox"/>	1.0

Comment: _____

Representative Blocks

Review performed by: _____
 Date: _____

Amendment(s) made by: _____
 Date: _____

Figure 1 Data form used to evaluate breast carcinomas.

data recovery. All patient and hospital identifiers were removed and a study number was assigned to each slide. The complete set was evaluated by each of the participating pathologists following a brief training session using a separate set of 15 slides, each selected to depict potential problem areas in invasive breast cancer diagnosis. Since geographic limitations prevented a single training session with all of the members of the working group, two separate training sessions were conducted, each by the same individual and with the same set of training slides. Each participant was asked to evaluate 35 slides with the pathology form (Figure 1). In addition, one participant (FOM) assessed the set

of 35 slides twice, the second time after an interval of more than a year in order to set the reference diagnosis and to ascertain reproducibility of the assessments for the reference diagnosis. Other than the single page data entry form, no other instructions or teaching sets were supplied. Each participant evaluated the same slide set. The forms were returned to the study coordinator and entered in the database by preassigned codes, thus masking the identity of the pathologist.

The individual characteristics of the cases in the study set with representative comparisons to the registry database are shown in Table 1. The histologic type of invasive breast cancer is

Table 1 Distribution of histologic features of study cases^a

<i>Histologic feature</i>	<i>Number</i>	<i>%</i>	<i>Registry database (%)</i>
<i>Nottingham grade</i>			
I	4	11.4	23
II	12	34.3	36
III	19	54.3	41
<i>Nottingham score</i>			
<7	10	28.6	84
≥7	25	71.4	16
<i>Primary histologic pattern</i>			
NST ^b	26	74.3	—
Other	9	25.7	—
<i>Primary or secondary pattern</i>			
Any cribriform	0	0	0
Any tubular	2	5.7	1.8
Any micropapillary	2	5.7	0.4
Any mucinous	4	11.4	1.4
Any lobular	5	14.3	8.5
Any metaplastic	1	2.9	0.4
Any medullary	4	11.4	2.8
<i>Lymphatic invasion</i>			
Present	6	17.1	—
Absent	29	82.9	—
<i>Blood vessel invasion</i>			
Present	0	0	—
Absent	35	100	—
<i>Lymphocytic infiltration</i>			
Marked	3	8.6	—
Moderate	4	11.4	—
Mild	17	48.6	—
Absent	11	31.4	—
<i>Syncytial growth</i>			
Present	3	8.6	—
Absent	32	91.4	—
<i>Circumscribed margins</i>			
Present	7	20	—
Absent	28	80	—

^aBased on reference standard assessment (FOM).

^bNST, no special type.

summarized in two ways: (1) the primary pattern, collapsed into two categories (no special type vs all others), and (2) the presence or absence of any individual pattern, either primary or secondary. ‘Any medullary feature’ includes both classical and atypical types. Similarly, ‘any lobular feature’ includes both classical and pleomorphic types. The histologic grade of invasive breast cancer is also summarized in two ways: (1) the overall Nottingham grade; I, II or III, and (2) Nottingham score summarized as below seven or seven and above. Cribriform architecture and blood vessel invasion were not present in any of the slides selected and were not further assessed.

The distribution of the histologic characteristics as well as the association among the individual

characteristics of the 35 cases based on the reference standard were tabulated. Cramer’s V statistic was used as a measure of association since it is suitable for categorical data and takes on values between -1 and 1 , similar to the usual Pearson’s correlation coefficient.³ Data were analyzed using two approaches. In the first approach, category-specific multirater percent agreement and κ statistics were used to characterize the agreement among the 13 pathologists.⁴ This approach does not assume to know the ‘truth’ or gold standard diagnosis for the given slide, but simply assesses the extent to which the pathologists agree among themselves. κ statistics were calculated to evaluate levels of agreement adjusted for agreement expected to occur by chance alone. Since κ is influenced by prevalence of the characteristic being measured, agreement was measured by category-specific κ and percent agreements to accommodate uncommon or low prevalent features.⁵ In general, κ statistics less than 0.4 are associated with relatively poor agreement, values of 0.4 – 0.6 moderate agreement, values of 0.6 – 0.8 substantial (good) agreement and values greater than 0.8 are associated with excellent (almost perfect) agreement.⁶

In the second approach, the reproducibility (or accuracy) of the study pathologists’ diagnoses was assessed relative to the reference standard. This latter analysis can be extrapolated as a reflection of completely centralized pathology review vs the semicentralized review actually implemented. High accuracy in this approach was defined as a high probability of an individual pathologist detecting a feature given that it is detected by the reference standard. In order to assess the intraobserver agreement of the standard, category-specific κ ’s and percentages of agreement were calculated using the first and second assessments by the reference standard. The other pathologists’ assessments were then compared to the reference to determine what percentage of the slides assigned to a category by the reference was also assigned to that category by the reviewing pathologists.

Occasionally, reviewing pathologists did not score individual items on the assessment sheets, but since the number of nonscored items was low (on average, 2.5% were missing), the nonscored items were accommodated by adjusting the denominator number of slides appropriately in the calculations.

Results

The study group consisted of six pathologists from population-based sites and seven from clinical-based sites. Three have a special interest in breast pathology, three were surgical pathology fellows during the study period and all remaining participants either practiced general pathology or have a special interest in other areas of surgical pathology.

Since cases were selected to (1) highlight problem areas previously identified in the initial intergroup

meeting and (2) represent cases accessioned into the registry database, the study set tended to over-represent unusual histologic subtypes and higher grade tumors. The distribution of cases in the study set and in the registry database is presented

in Table 1. Representative dot plots of the individual pathologists' scores for primary and secondary histologic patterns, Nottingham grade and lymphocytic infiltrate for each of the 35 slides is presented in Figures 2–5.

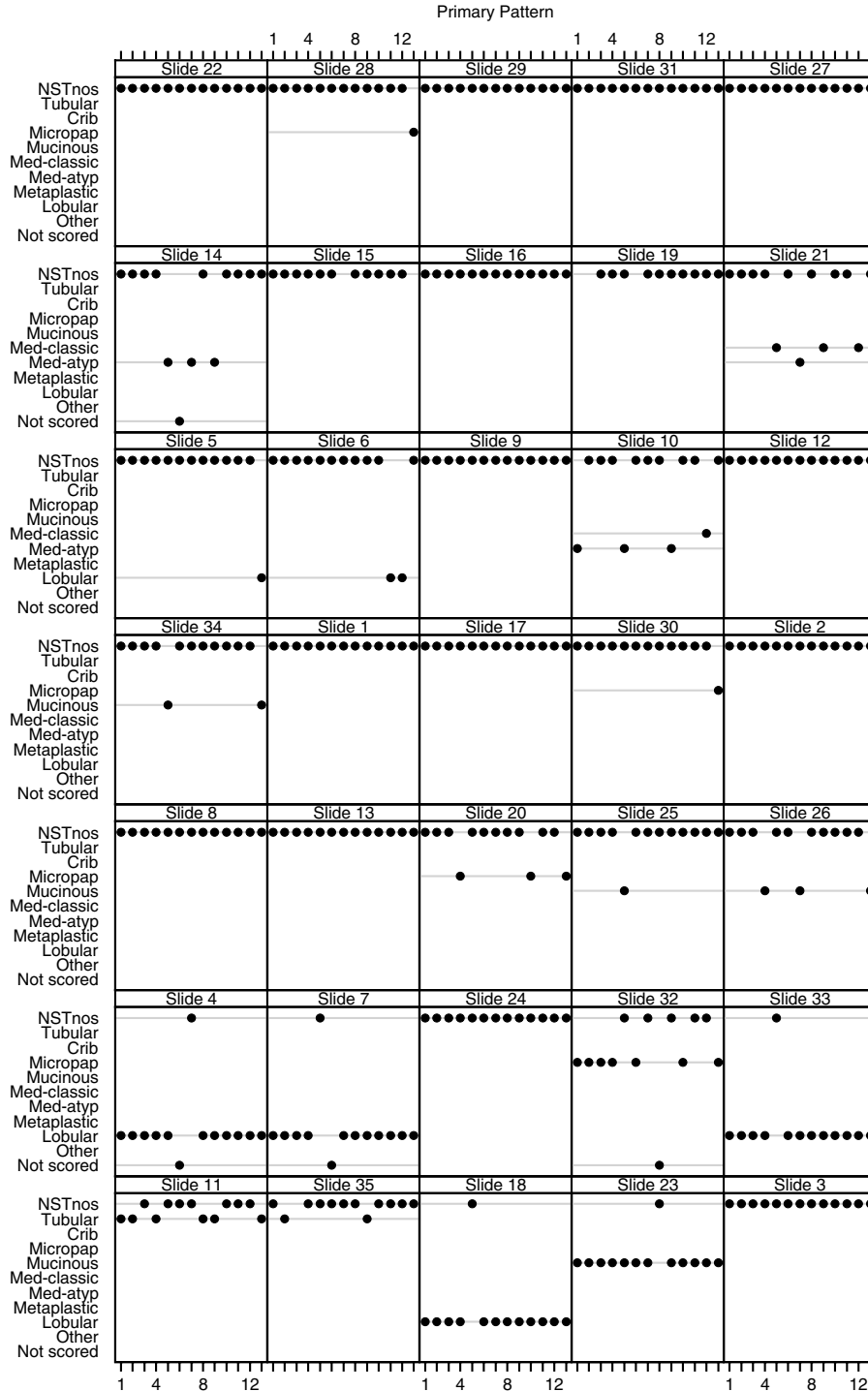


Figure 2 Dot plot depicting distribution of scores for primary histologic pattern, (specific patterns are provided at left), of 35 invasive breast cancers evaluated by the 13 pathologists. The slide number of the individual breast cancer is provided above each boxed entry. Grids at top and bottom represent individual pathologists, with the reference standard at 1. If a specific histologic pattern is not identified by an individual pathologist as the primary pattern (eg lobular, pathologist 7 case 33), it is often identified as the secondary pattern (see Figure 3).

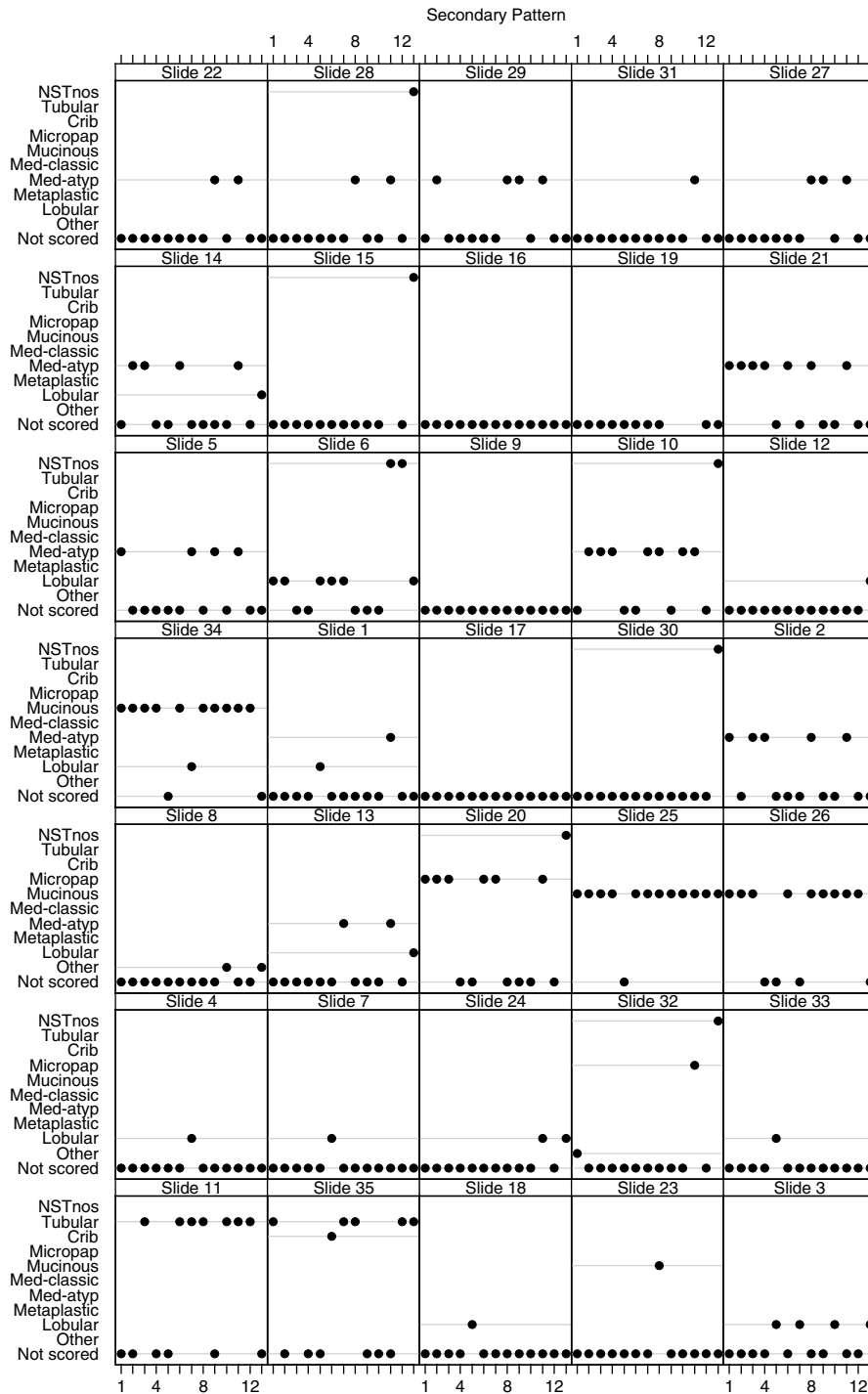


Figure 3 Dot plot depicting distribution of scores for secondary histologic pattern, (specific patterns are provided at left), of 35 invasive breast cancers evaluated by the 13 pathologists. The slide number of the individual breast cancer is provided above each boxed entry. Grids at top and bottom represent individual pathologists, with the reference standard at 1.

Classification of the specific subtype of breast cancer by primary pattern showed generally high agreement (Figure 2). Causes of discrepant diagnoses were most commonly attributed to a ‘no special type’ classification by one reviewer and a ‘lobular’, ‘atypical medullary’, or ‘mucinous’ type by

other reviewers. In many of these cases, the discrepant diagnoses were ultimately captured in the secondary pattern; that is, although ‘no special type’ was assigned to the primary pattern by a reviewer, ‘lobular’ or ‘medullary’ or ‘mucinous’ was assigned to the secondary pattern (and vice versa for

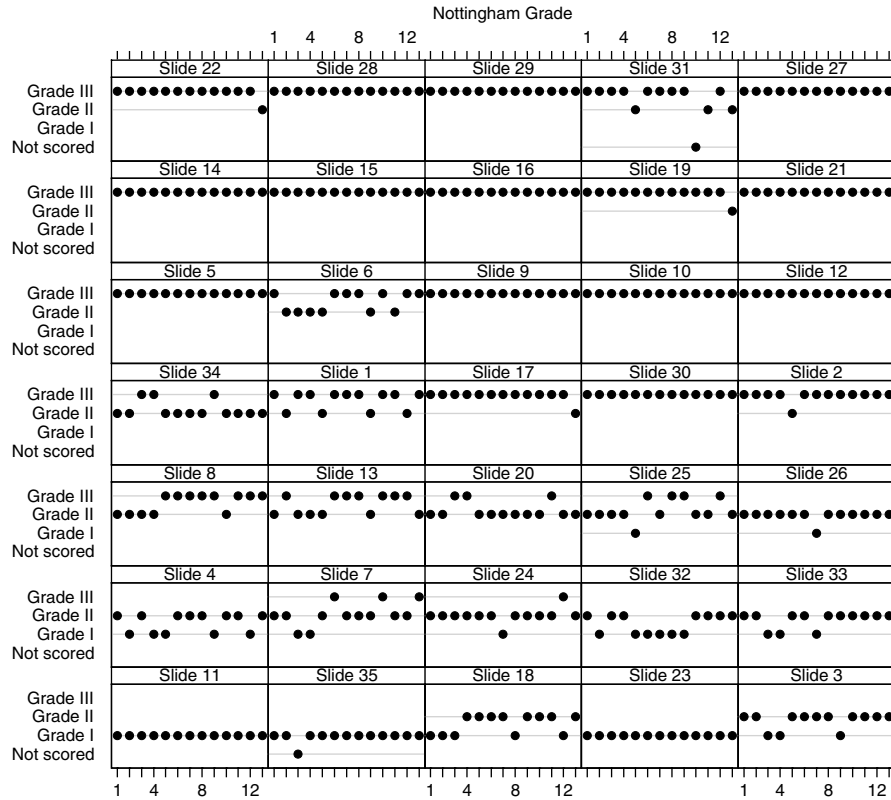


Figure 4 Dot plot depicting distribution of scores for Nottingham grade of 35 invasive breast cancers evaluated by the 13 pathologists. The slide number of the individual breast cancer is provided above each boxed entry. Grids at top and bottom represent individual pathologists, with the reference standard at 1. Interobserver reproducibility for Nottingham grade is good overall, but significantly better for grade III cancers than grade II cancers based on the reference standard.

the other reviewers) (Figures 2 and 3). In these instances, when the primary and secondary patterns were grouped, discrepancies in classification of histologic type were significantly decreased, but not completely eradicated (Table 2).

The interobserver percent agreement for classification of the specific histologic type of the 35 invasive breast cancers by primary or secondary pattern ranged from 35 to 99.5%; this corresponded to a category-specific κ range of 0.3–1.0 (Table 2). Despite the relatively large range in category-specific κ for the entire group, most of the poor reproducibility of classification of the invasive cancers could be attributed to the classification/misclassification of specific uncommon subtypes of invasive breast cancer. This was true for the 13 pathologists' interobserver agreement as well as for the reference pathologist's intraobserver agreement. Persistent causes for disagreement involved classification of the primary pattern as 'micropapillary', 'medullary' or 'metaplastic' carcinoma by one reviewer and classification of the primary pattern as 'no special type' by another reviewer without assignment of a specific secondary pattern. However, interobserver percent agreement was quite high when each of these diagnoses was considered to be absent. The category-specific agreement was

highest for tubular (78.7%), mucinous (96.0%) and lobular (78.0%) subtypes. Not surprisingly, there was significant interobserver disagreement in the classification of cancers as 'medullary'. Despite the wide range in κ for the classification of histologic type, the accuracy for assignment of histologic subtype (defined as the degree to which the reviewing pathologist identified the same feature as the reference pathologist) was quite high (Table 3), especially for ductal carcinoma, no special type (mean, 92%), any mucinous carcinoma (mean, 95.8%) and any lobular carcinoma (mean, 90%). The accuracy for classification of metaplastic carcinoma was quite low, in part due to the presence of only one such case in the study set. The case included was particularly difficult to interpret as the vast majority of the lesion was comprised of high-grade epithelial cells, with a small focus of chondroid change present at the edge of the section.

Category-specific κ values for the Nottingham grade of the 35 invasive breast cancers ranged from 0.5 to 0.7, with a corresponding percent agreement of 61.4–87.8%. κ values for Nottingham score ≥ 7 or < 7 were slightly better ($\kappa = 0.7$). The intraobserver percent agreement for Nottingham grade (87–100%) and score (94.7–98%) were markedly better than

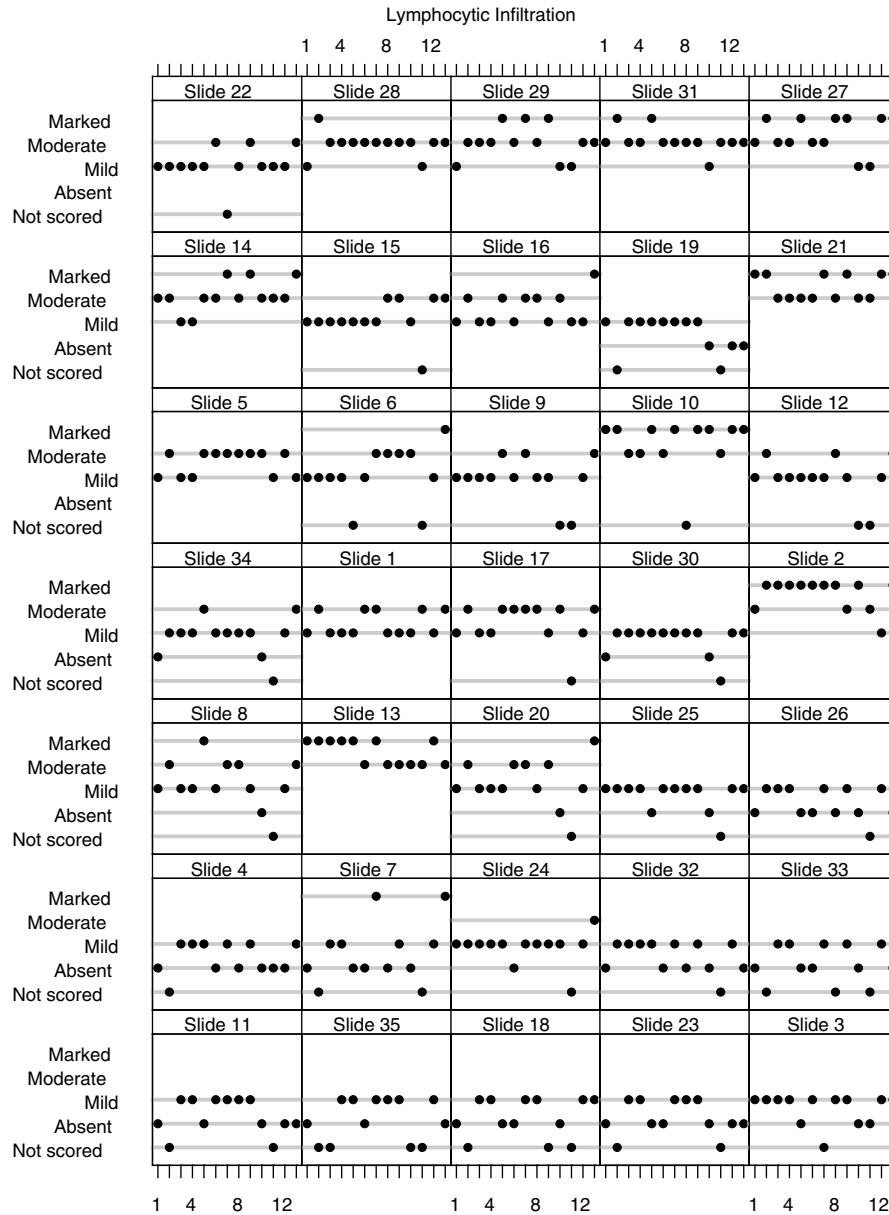


Figure 5 Dot plot depicting distribution of scores for lymphocytic infiltration in 35 invasive breast cancers evaluated by the 13 pathologists. The slide number of the individual breast cancer is provided above each boxed entry. Grids at top and bottom represent individual pathologists, with the reference standard at 1. Although agreement appears low, when collapsed into a binary score (absent or mild vs moderate or marked) the accuracy of classification with respect to the reference standard was quite high (mean, 73.7–92.9%).

the interobserver agreement (the value of this level of intraobserver agreement is admittedly limited in that only one pathologist was utilized for this calculation, but it provides a sense of reproducibility for the reference standard used to assess the accuracy of classification in this study). Disagreement on grading was usually attributed to differences in classification of the grade II carcinomas. Even though there was a relatively wide range in interobserver agreement for Nottingham grade, the accuracy for classification of the grade was quite high, ranging from 75 to 100% (mean, 83.3%) for grade I; 50 to 83.3% (mean,

64.6%) for grade II and 79 to 100% (mean, 92.3%) for grade III tumors.

The category-specific percent agreement among the 13 pathologists for the presence of lymphatic space invasion was 55% in this study, whereas the category-specific agreement for the absence of lymphatic space invasion was 91%. The comparative intraobserver agreement for the presence or absence of this histologic feature was 90.9 and 98.3%, respectively. The accuracy for the individual pathologists ranged from 33.3 to 100% (mean, 65.6%) for the determination of the presence of lymphatic space invasion, while the accuracy for the

Table 2 Category-specific κ and percent interobserver and intraobserver agreement

Histologic feature	Category	Interobserver agreement ^a		Intraobserver agreement ^b	
		Category-specific κ	Category specific % agreement	Category-specific κ	Category-specific % agreement
NST	Yes	0.5	88.9	0.9	96.2
	No	0.5	61.7	0.9	88.9
Any tubular	Yes	0.8	78.7	1.00	100
	No	0.8	99.1	1.00	100
Any micropapillary	Yes	0.6	58.2	1.00	100
	No	0.6	98.2	1.00	100
Any mucinous	Yes	1.0	96	1.00	100
	No	1.0	99.5	1.00	100
Any medullary	Yes	0.4	50	0.7	75.0
	No	0.4	93.3	0.7	96.8
Any metaplastic	Yes	0.3	35	0.7	66.7
	No	0.3	98.5	0.7	98.5
Any lobular	Yes	0.8	78	1.00	100
	No	0.8	96	1.00	100
Nottingham grade	I	0.7	68.8	1.00	100
	II	0.4	61.4	0.8	87.0
	III	0.7	87.8	0.8	92.3
Overall score	≥ 7	0.7	91.4	0.9	98.0
	< 7	0.7	78.9	0.9	94.7
Lymphatic invasion	Present	0.5	55	0.9	90.9
	Absent	0.5	91	0.9	98.3
Lymphocytic infiltration ^c	High	0.6	73.8	0.5	66.7
	Low	0.6	80	0.5	85.1
Circumscribed margins	Present	0.4	50.9	0.7	75.0
	Absent	0.4	83.3	0.7	92.6
Syncytial growth	Present	0.5	61.2	0.6	66.7
	Absent	0.5	91.4	0.6	96.9

^aThirteen pathologists.

^bReference pathologist.

^cMarked or moderate vs mild or absent.

determination of the *absence* of lymphatic space invasion ranged from 48.2 to 100% (mean, 92.9%). Two-thirds or more of the cases were accurately classified as positive for lymphatic space invasion by 50% of the reviewing pathologists while all but one reviewing pathologist accurately classified two-thirds or more of the cases as negative for lymphatic space invasion.

The interobserver percent agreement for lymphocytic infiltration ranged from 31.4 to 58.3% (category-specific $\kappa = 0.2-0.4$) when the infiltrate was evaluated using a four-tier score (absent, mild, moderate, marked), but improved to 73.8–80% (category-specific $\kappa = 0.6$) when collapsed into a binary score (absent or mild vs moderate or marked). The level of intraobserver agreement for this feature was not markedly better than the interobserver agreement and was only modestly improved by the use of the binary score. However, the accuracy of classification with respect to the reference standard was quite high with the binary scheme (mean, 73.7–92.9%) (Table 3). The interobserver and intraobserver percent agreement for the presence of syncytial growth and circumscribed margins was 61.2 and 66.7%, and 50.9 and 75%, respectively;

however, the level of agreement was markedly higher for the *absence* of these two features (91.4 vs 96.9% and 83.3 vs 92.6%, respectively). Average accuracy of identifying the presence of syncytial growth or circumscribed margins was 97.2 and 67.9%, respectively (Table 3).

Since one of the primary aims of the cancer registry involved correlation of specific histologic subtypes of invasive cancer with BRCA1 and BRCA2 mutation status, as well as with other potential molecular, epidemiologic, and therapeutic outcomes, the ability to search and identify specific histologic types within the registry was of interest to the Pathology Working Group. Although the category-specific interobserver agreement for the individual histologic criteria for medullary carcinoma (lymphocytic infiltrate, syncytial growth, circumscribed margin) varied widely, on average 94% of the primary and secondary pattern medullary cases identified by the reference standard could be identified in the data sheet for each reviewer by either primary or secondary histologic pattern, marked lymphocytic infiltrate, presence of circumscribed margins or presence of a syncytial growth pattern. By widening the net in this way, the

Table 3 Accuracy of classification of histologic features in invasive breast cancer relative to reference standard^a

Feature	Category	n ^b	Percent cases identified by all reviewers	
			Mean	Range
NST	Yes	26	92.0	73.1–100
Any tubular	Yes	2	75.0	0–100
Any micropapillary	Yes	2	62.5	0–100
Any mucinous	Yes	4	95.8	75–100
Any medullary	Yes	4	56.3	0–100
Any metaplastic	Yes	1	41.7	0–100
Any lobular	Yes	5	90.0	80–100
Grade	I	4	83.3	75–100
	II	12	64.6	50–83.3
	III	19	92.5	79–100
Overall score	≥7	25	93.3	84–100
	<7	10	86.7	50–100
Lymphatic invasion	Present	6	65.6	33.3–100
	Absent	29	92.9	48.2–100
Lymphocytic infiltration ^c	High	7	92.9	71.4–100
	Low	28	73.7	46.4–92.9
Circumscribed margins	Present	7	67.9	28.6–100
Syncytial growth	Present	3	97.2	66.7–100

^aAccuracy is defined as the percent of the reference standard classifications classified or identified similarly by the reviewing pathologists.

^bNumber of cases according to the reference standard.

^cMarked or moderate vs mild or absent.

number of potential medullary cases was increased on average three-fold compared to primary/secondary patterns alone. Similarly, of the primary and secondary patterns identified as tubular or lobular by the reference standard, on average 96 and 88%, respectively could be retrieved from the data sheets for each reviewer by also including cases with Nottingham grade I.

Discussion

Recent studies have indicated that interobserver agreement in breast cancer grading and typing can be optimized by the use of well-defined, agreed upon criteria and terminology.^{7,8} Despite these advances in our understanding of breast cancer diagnosis and histopathologic classification, an unspecified degree of interobserver and intraobserver disagreement is inescapable in the assignment of any classification, grade or overall score to a pathologic process that is (1) based on subjective distinctions along a histologic continuum and (2) requires evaluation for a variety of pathologic characteristics, some of which may be relatively uncommon. In implementing the pathology review and data collection for the breast cancer registry, our goals were to optimize interobserver agreement by establishing uniform, well-specified and agreed upon criteria for classification, to identify potential sources of persistent interobserver disagreement and to design a data entry form that could accommodate this level of disagreement.

Given the constraints imposed by the overall registry goals, the level of agreement obtained by the registry pathologists in grading these 35 invasive carcinomas was quite good. Category-specific κ scores were moderate to good for Nottingham grade ($\kappa = 0.5–0.7$), good for histologic score <7 vs ≥ 7 ($\kappa = 0.7$) and, although not directly comparable somewhat higher than that which has been reported previously. Although Frierson *et al*¹⁰ reported moderate to substantial κ values for interobserver agreement for histologic grade, Delides *et al*⁹ found low interobserver agreement and overall κ values for the European Working Group and for the Japan National Surgical Adjuvant Study were moderate at best.^{8–11} Interobserver and intraobserver agreement, as well as accuracy of classification relative to the reference standard were higher for the grade I and grade III tumors than for the grade II tumors. These results are similar to those of Dalton *et al*,¹² who showed that excellent agreement for histologic grade was more likely to occur for extremely low-grade and extremely high-grade cancers. In our study, it appeared that differences in assessment of the degree of nuclear pleomorphism were most commonly responsible for differences in assessment of overall histologic score, followed by mitotic index and tubule formation (data not shown).

The significant degree of interobserver disagreement that occurs in the allocation of nuclear grade in breast cancer has been noted in previous studies.^{9,10,12–14} In at least one series, it was suggested that pathologists who are not specialists in breast disease tend to underscore, possibly due to a

preconception that invasive breast cancer sorts equally into each of the three grades.¹⁵ It has also been argued that reproducibility of nuclear pleomorphism is difficult because of the nonquantitative nature of the scoring method,¹⁶ but in our opinion, the intermediate nature of some breast cancers and the heterogeneity in nuclear pleomorphism that can occur in these malignancies is underappreciated and probably contributes to the relatively high and variable degree of interobserver disagreement, especially with respect to the intermediate grade tumors. In comparison to nuclear pleomorphism, the criteria for scoring tubule formation and mitotic score are relatively robust. Concordance in mitotic counts is highest when counts are determined in the same area, using established counting methods and established criteria.^{16–19} Mitotic counts also depend on the quality of the tissue processing and the size of the ocular lens.^{16,18,19} Since the registry relies in part upon the ability of the participating pathologists to select the optimum area for mitotic counts, there was no attempt to guide the reviewers to any single designated area on the study slides and this is likely responsible for some of the interobserver disagreement. Nevertheless, it is likely that the category-specific κ obtained in our study overestimates the level of agreement that would occur during the actual performance of the registry data collection, since the area of determination of mitotic counts is occasionally selected from among several sections of tumor by the registry pathologist during the actual review procedure. The level of degradation in interobserver agreement would depend on the numbers of cases in which several sections were examined, the number of sections examined and the relative contribution of moderately differentiated tumors, all of which would likely vary depending on the registry center and the submitting hospital.

Only one previous study has evaluated the ability of a group of pathologists to assign a histologic subtype to a range of invasive carcinomas. In the study conducted by the European Working Group, it was found that subtyping was most consistent for mucinous carcinoma, followed by lobular carcinoma and least consistent for medullary carcinoma.¹¹ Our results are similar with the additional finding of a relatively high degree of consistency for subtyping tubular carcinoma. The poor reproducibility for the diagnosis of invasive lobular carcinoma relative to ductal and mucinous tumors has been noted by others and appears to be due to (1) a tendency for overdiagnosis of lobular cancer; (2) confusion regarding diagnostic criteria for the pleomorphic subtype and (3) suboptimal histology.²⁰ The moderate degree of interobserver agreement for medullary carcinoma has been the subject of prior studies.^{21–23} Utilizing the criteria of Ridolfi *et al*,²⁴ consensus diagnoses were achieved in 56.3% of cases in the current study. These results are similar to that

achieved by others.^{21,22} In our study, the range in interobserver agreement was greatest for lymphocytic infiltrate, followed by margin status and syncytial growth pattern. Moreover, when the co-association of individual histologic features was analyzed, the medullary subtype was most highly associated with circumscribed margins, followed by syncytial growth and lymphocytic infiltration (data not shown). These findings are similar to those of Gaffey *et al*²³ and contrast with Pedersen *et al*,²² who found that interobserver agreement was lowest for circumscription, although the latter authors used a three-tiered scoring system for circumscription and lymphocytic infiltrate.

It was anticipated that cases could be ambiguous either due to the histologic features of the individual tumor or due to difficulties in the application of the individual criteria. Therefore, the pathology data sheet was designed to accommodate cases that appeared to be ambiguous, mixed or borderline to the reviewer, regardless of whether it was due to the borderline nature of the tumor itself or due to underspecified or poorly specified criteria. As expected, interobserver disagreement in classification of histologic type was largely due to differences in classification of lobular and medullary carcinoma, but differences in classification of the latter diagnosis were markedly decreased by assigning a primary and secondary pattern to each of the cases. Thus, a case that was scored as 'lobular' for the primary pattern by most reviewers was scored as 'lobular' for the secondary pattern by the remaining reviewers in two cases (100%) and all but one reviewer in the other two cases (92%) (Figure 3). Classification of cancers exhibiting a medullary pattern was also improved by incorporating primary and secondary pattern, but not to the same degree, in part due to the absence of classical medullary carcinomas in the study set and in part due to the inherent poor reproducibility for this diagnosis. However, even though reproducibility for the diagnosis of medullary carcinoma is quite poor, the four cancers scored as 'any medullary' by the reference standard in this study could be identified by most pathologists by a combination of medullary, marked lymphocytic infiltrate, circumscribed margins or syncytial pattern. The ability to identify the majority of cases falling within the diagnostic range for these particular subtypes is important, given the predilection for lobular and medullary cancer in familial and hereditary breast cancer families.^{25–31}

Central review continues to be a necessary component to any large cooperative study involving pathologic materials. However, we have shown that a well-designed data entry sheet for pathology review obviates the need for a single central review and permits the review process to occur on a more localized basis, provided the data entry form is designed to facilitate the identification and retrieval of the histologically ambiguous and unambiguous

cases. This approach promotes a shift from the dualistic paradigms of lobular/ductal or medullary/nonmedullary to one that embraces a histologic continuum and recognizes tumor heterogeneity. In our opinion, this latter approach, in conjunction with epidemiologic, therapeutic and molecular developments, is the approach that is most likely to advance our understanding of carcinogenesis and ultimately, our therapeutic decisions.

Acknowledgements

This work is supported in part by the National Cancer Institute, National Institutes of Health under RFA #CA-95-011. The content of this paper does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the CFR, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government or the CFR. Additional pathologists who participated in the interobserver slide evaluation include Deon J Venter, MD, and Jane Armes MD, University of Melbourne, Melbourne, Australia. We acknowledge Thelma Santa Maria and Caroline Tudor, Stanford University and Marie Maguire, Mount Sinai Hospital, Toronto, for assistance in the paper and figure preparation.

References

- Schnitt SJ, Connolly JL, Tavassoli FA, *et al*. Interobserver reproducibility in the diagnosis of ductal proliferative breast lesions using standardized criteria [see comments]. *Am J Surg Pathol* 1992;16:1133–1143.
- Elston CW, Ellis IO. *The Breast*, 3rd edn. Churchill Livingstone: Edinburgh, 1998.
- Everitt BS. *The Analysis of Contingency Tables*. Chapman & Hall: London, 1977.
- Chmura Kraemer H, Periyakoil VS, Noda A. Kappa coefficients in medical research. *Stat Med* 2002;21:2109–2129.
- Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol* 1990;43:551–558.
- Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–174.
- Page DL, Ellis IO, Elston CW. Histologic grading of breast cancer. let's do it [editorial]. *Am J Clin Pathol* 1995;103:123–124.
- Tsuda H, Akiyama F, Kurosumi M, *et al*. The efficacy and limitations of repeated slide conferences for improving interobserver agreement when judging nuclear atypia of breast cancer. The Japan National Surgical Adjuvant Study of Breast Cancer (NSAS-BC) Pathology Section. *Jpn J Clin Oncol* 1999;29:68–73.
- Delides GS, Garas G, Georgouli G, *et al*. Intralaboratory variations in the grading of breast carcinoma. *Arch Pathol Lab Med* 1982;106:126–128.
- Frierson Jr HF, Wolber RA, Berean KW, *et al*. Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma. *Am J Clin Pathol* 1995;103:195–198.
- Sloane JP, Amendoeira I, Apostolikas N, *et al*. Consistency achieved by 23 European pathologists from 12 countries in diagnosing breast disease and reporting prognostic features of carcinomas. European Commission Working Group on Breast Screening Pathology. *Virchows Arch* 1999;434:3–10.
- Dalton LW, Page DL, Dupont WD. Histologic grading of breast carcinoma. A reproducibility study. *Cancer* 1994;73:2765–2770.
- Harvey JM, de Klerk NH, Sterrett GF. Histological grading in breast cancer: interobserver agreement, and relation to other prognostic factors including ploidy. *Pathology* 1992;24:63–68.
- Jacquemier J, Charpin C. Reproducibility of histoprognostic grades of invasive breast cancer. *Ann Pathol* 1998;18:385–390.
- Dunne B, Going JJ. Scoring nuclear pleomorphism in breast cancer. *Histopathology* 2001;39:259–265.
- Tsuda H, Akiyama F, Kurosumi M, *et al*. Evaluation of the interobserver agreement in the number of mitotic figures of breast carcinoma as simulation of quality monitoring in the Japan National Surgical Adjuvant Study of Breast Cancer (NSAS-BC) protocol. *Jpn J Cancer Res* 2000;91:451–457.
- Hilsenbeck SG, Allred DC. Improved methods of estimating mitotic activity in solid tumors. *Hum Pathol* 1992;23:601–602.
- Simpson JF, Dutt PL, Page DL. Expression of mitoses per thousand cells and cell density in breast carcinomas: a proposal. *Hum Pathol* 1992;23:608–611.
- van Diest PJ, Baak JP, Matze-Cok P, *et al*. Reproducibility of mitosis counting in 2469 breast cancer specimens: results from the Multicenter Morphometric Mammary Carcinoma Project. *Hum Pathol* 1992;23:603–607.
- Cserni G. Reproducibility of a diagnosis of invasive lobular carcinoma. *J Surg Oncol* 1999;70:217–221.
- Rigaud C, Theobald S, Noel P, *et al*. Medullary carcinoma of the breast. A multicenter study of its diagnostic consistency. *Arch Pathol Lab Med* 1993;117:1005–1008.
- Pedersen L, Holck S, Schiodt T, *et al*. Inter- and intraobserver variability in the histopathological diagnosis of medullary carcinoma of the breast, and its prognostic implications. *Breast Cancer Res Treat* 1989;14:91–99.
- Gaffey MJ, Mills SE, Frierson Jr HF, *et al*. Medullary carcinoma of the breast: interobserver variability in histopathologic diagnosis. *Mod Pathol* 1995;8:31–38.
- Ridolfi RL, Rosen PP, Port A, *et al*. Medullary carcinoma of the breast: a clinicopathologic study with 10 year follow-up. *Cancer* 1977;40:1365–1385.
- Pathology of familial breast cancer: differences between breast cancers in carriers of BRCA1 or BRCA2 mutations and sporadic cases. Breast Cancer Linkage Consortium. *Lancet* 1997;349:1505–1510.
- Armes JE, Egan AJ, Southey MC, *et al*. The histologic phenotypes of breast carcinoma occurring before age 40 years in women with and without BRCA1 or BRCA2 germline mutations: a population-based study. *Cancer* 1998;83:2335–2345.
- Claus EB, Risch N, Thompson WD, *et al*. Relationship between breast histopathology and family history of breast cancer. *Cancer* 1993;71:147–153.

- 28 Eisinger F, Jacquemier J, Charpin C, *et al*. Mutations at BRCA1: the medullary breast carcinoma revisited. *Cancer Res* 1998;58:1588–1592.
- 29 Lakhani SR, Gusterson BA, Jacquemier J, *et al*. The pathology of familial breast cancer: histological features of cancers in families not attributable to mutations in BRCA1 or BRCA2. *Clin Cancer Res* 2000;6:782–789.
- 30 Marcus JN, Watson P, Page DL, *et al*. Hereditary breast cancer: pathobiology, prognosis, and BRCA1 and BRCA2 gene linkage. *Cancer* 1996;77:697–709.
- 31 Quenneville L, Phillips KA, Ozelik H, *et al*. HER2/neu status and tumor morphology of invasive breast carcinomas in Ashkenazi women with known BRCA-1 mutation status in the OFBCR. *Cancer* 2002;95:2068–2075.