Original Investigation

Diagnostic Concordance Among Pathologists Interpreting Breast Biopsy Specimens

Joann G. Elmore, MD, MPH; Gary M. Longton, MS; Patricia A. Carney, PhD; Berta M. Geller, EdD; Tracy Onega, PhD; Anna N. A. Tosteson, ScD; Heidi D. Nelson, MD, MPH; Margaret S. Pepe, PhD; Kimberly H. Allison, MD; Stuart J. Schnitt, MD; Frances P. O'Malley, MB; Donald L. Weaver, MD

IMPORTANCE A breast pathology diagnosis provides the basis for clinical treatment and management decisions; however, its accuracy is inadequately understood.

OBJECTIVES To quantify the magnitude of diagnostic disagreement among pathologists compared with a consensus panel reference diagnosis and to evaluate associated patient and pathologist characteristics.

DESIGN, SETTING, AND PARTICIPANTS Study of pathologists who interpret breast biopsies in clinical practices in 8 US states.

EXPOSURES Participants independently interpreted slides between November 2011 and May 2014 from test sets of 60 breast biopsies (240 total cases, 1 slide per case), including 23 cases of invasive breast cancer, 73 ductal carcinoma in situ (DCIS), 72 with atypical hyperplasia (atypia), and 72 benign cases without atypia. Participants were blinded to the interpretations of other study pathologists and consensus panel members. Among the 3 consensus panel members, unanimous agreement of their independent diagnoses was 75%, and concordance with the consensus-derived reference diagnoses was 90.3%.

MAIN OUTCOMES AND MEASURES The proportions of diagnoses overinterpreted and underinterpreted relative to the consensus-derived reference diagnoses were assessed.

RESULTS Sixty-five percent of invited, responding pathologists were eligible and consented to participate. Of these, 91% (N = 115) completed the study, providing 6900 individual case diagnoses. Compared with the consensus-derived reference diagnosis, the overall concordance rate of diagnostic interpretations of participating pathologists was 75.3% (95% CI, 73.4%-77.0%; 5194 of 6900 interpretations).

	Pathologist Interpretation vs Consensus-Derived Reference Diagnosis, % (95% CI)			
Consensus Reference Diagnosis	No. of Interpretations	Overall Concordance Rate	Overinterpretation Rate	Underinterpretation Rate
Benign without atypia	2070	87 (85-89)	13 (11-15)	
Atypia	2070	48 (44-52)	17 (15-21)	35 (31-39)
DCIS	2097	84 (82-86)	3 (2-4)	13 (12-15)
Invasive carcinoma	663	96 (94-97)		4 (3-6)

Disagreement with the reference diagnosis was statistically significantly higher among biopsies from women with higher (n = 122) vs lower (n = 118) breast density on prior mammograms (overall concordance rate, 73% [95% CI, 71%-75%] for higher vs 77% [95% CI, 75%-80%] for lower, $P < .001$), and among pathologists who interpreted lower weekly case volumes ($P < .001$) or worked in smaller practices ($P = .034$) or nonacademic settings ($P = .007$).

CONCLUSIONS AND RELEVANCE In this study of pathologists, in which diagnostic interpretation was based on a single breast biopsy slide, overall agreement between the individual pathologists' interpretations and the expert consensus–derived reference diagnoses was 75.3%, with the highest level of concordance for invasive carcinoma and lower levels of concordance for DCIS and atypia. Further research is needed to understand the relationship of these findings with patient management.

JAMA. 2015;313(11):1122-1132. doi:10.1001/jama.2015.1405

Editorial page 1109

JAMA Report Video and **Author Video Interview** at jama.com

Supplemental content at jama.com

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Joann G. Elmore, MD, MPH, Department of Medicine, University of Washington, 325 Ninth Ave, PO Box 359780, Seattle, WA 98104 (jelmore@u.washington.edu).

Approximately 1.6 million women in the United States have breast biopsies each year.[1,2] The accuracy of pathologists' diagnoses is an important and inadequately studied area. Although nearly one-quarter of biopsies demonstrate invasive breast cancer,[3] the majority are categorized by pathologists according to a diagnostic spectrum ranging from benign to preinvasive disease. Breast lesions with atypia or ductal carcinoma in situ (DCIS) are associated with significantly higher risks of subsequent invasive carcinoma, and women with these findings may require additional surveillance, prevention, or treatment to reduce their risks.[4] The incidence of atypical ductal hyperplasia (atypia) and DCIS breast lesions has increased over the past 3 decades as a result of widespread mammography screening.[5,6] Misclassification of breast lesions may contribute to either overtreatment or undertreatment of lesions identified during breast screening.

The pathological diagnosis of a breast biopsy is usually considered the gold standard for patient management and research outcomes. However, a continuum of histologic features exists from benign to atypical to malignant on which diagnostic boundaries are imposed. Although criteria for these diagnostic categories are established,[7,8] whether they are uniformly applied is unclear. Nonetheless, patients and their clinicians need a specific diagnostic classification of biopsy specimens to understand whether increased risk for breast cancer exists and how best to manage identified lesions. Although studies from the 1990s demonstrated challenges encountered by pathologists in agreeing on the diagnoses of atypia and DCIS,[9-12] the extent to which these challenges persist is unclear. These issues are particularly important in the 21st century because millions of breast biopsies are performed annually.

For these reasons, we investigated the magnitude of overinterpretation and underinterpretation of breast biopsies among a national sample of practicing US pathologists in the Breast Pathology (B-Path) study. We also evaluated whether patient and pathologist characteristics were associated with a higher prevalence of inaccurate interpretations.

Methods

Human Research Participants Protection

The institutional review boards at Dartmouth College, Fred Hutchinson Cancer Research Center, Providence Health and Services Oregon, University of Vermont, and University of Washington approved all study activities. Informed consent was obtained electronically from pathologists. Informed consent was not required of the women whose biopsy specimens were included.

Test Set Development

Study methods and test set development have been described.[13-15] Briefly, 240 breast biopsy specimens (excisional or core needle) were randomly identified from a cohort of 19 498 cases obtained from pathology registries in New Hampshire and Vermont that are affiliated with the Breast Cancer Surveillance Consortium.[16] Random, stratified sampling was used to select cases based on the original pathologists' diagnoses. Data on women's age, breast density, and biopsy type were available for each case. One or 2 new slides from candidate cases were prepared in a single laboratory for consistency. A single slide for each case best representing the reference diagnosis in the opinion of the panel members was selected during the consensus review meetings.[13]

We oversampled cases with atypia and DCIS to gain statistical precision in estimates of interpretive concordance for these diagnoses. We also oversampled cases from women aged 40 to 49 years and women with mammographically dense breast tissue because age and breast density are important risk factors for both benign breast disease and breast cancer.[17] We hypothesized that discordance would be higher for these biopsy cases and that discordance would be higher when pathologists reported cases as "borderline" between 2 diagnostic categories.

A panel of 3 experienced pathologists, internationally recognized for research and continuing medical education on diagnostic breast pathology, independently reviewed all 240 cases and recorded their rating of case difficulty and diagnoses using a Breast Pathology Assessment Tool and Hierarchy for Diagnosis form, which was designed and rigorously tested for this study (eFigure 1 in the Supplement).[15] Panel members were blinded to previous interpretations of each specimen and to each other's interpretations. Cases without unanimous independent agreement were resolved with consensus discussion. Four full-day in-person meetings were held following the panel members' independent reviews to establish a consensus reference diagnosis for each case using a modified Delphi approach,[18] to create case teaching points, and to discuss study design.

The 14 assessment terms were grouped into 4 diagnostic categories (eTable 1 in the Supplement). The categories and corresponding target distribution for the final sample of 240 cases were benign without atypia (30%, including 10% nonproliferative and 20% proliferative without atypia), atypia (30%), DCIS (30%), and invasive carcinoma (10%). The nonproliferative and proliferative without atypia cases were merged into 1 category (benign without atypia) because clinical management usually does not differ between the 2 categories. When pathologists noted multiple diagnoses on a case, the most severe diagnostic category was assigned.

The 3 reference pathologists agreed unanimously on the diagnosis for 75% (180 of 240) of the cases after the initial independent evaluation. Compared with the final consensus-derived reference diagnoses, overall concordance of the initial independent diagnoses of the expert panel members was 90.3% (650 of 720 interpretations; **Figure 1**). Concordance and rates of overinterpretation and underinterpretation of initial diagnoses by the panel members compared with consensus-derived reference diagnoses are presented in **Table 1**.

The 240 cases were randomly assigned to 1 of 4 test sets each including 60 cases with randomization stratified on the woman's age, breast density, reference diagnosis, and the experts' difficulty rating of the case.

Figure 1. Comparison of the 3 Reference Panel Members' Independent Preconsensus Diagnoses vs the Consensus-Derived Reference Diagnosis for 240 Breast Biopsy Cases[a]

		Reference Panel Members' Individual Diagnoses (Preconsensus)				
		Benign without atypia	Atypia	DCIS	Invasive carcinoma	Total
Consensus Reference Diagnosis	Benign without atypia	197	15	3	1[b]	216
	Atypia	18	173	25	0	216
	DCIS	2	2	213	2[c]	219
	Invasive carcinoma	0	0	2[d]	67	69
	Total	217	190	243	70	720

DCIS indicates ductal carcinoma in situ.

[a] Concordance noted for 650 of 720 diagnoses or 90.3%.

[b] The differential diagnosis was radial scar vs focal invasion (with a consensus-derived reference diagnosis of radial scar).

[c] The differential diagnosis was focal microinvasion vs DCIS (with a consensus-derived reference diagnosis of DCIS).

[d] The differential diagnosis was DCIS vs focal microinvasion (with a consensus-derived reference diagnosis of microinvasion).

Table 1. Rates of Overinterpretation, Underinterpretation, and Concordance for the Reference Pathologists' Independent Preconsensus Interpretations vs the Consensus-Derived Reference Diagnosis[a]

		Rate, % (Range)[b]		
		Rate of Overinterpretation or Underinterpretation vs Consensus Diagnosis		Overall Concordance Rate vs Consensus Diagnosis
Consensus Reference Diagnosis	Total, No.	Overinterpretation	Underinterpretation	Concordance
Benign without atypia	72	9 (3-13)		91 (87-97)
Atypia	72	12 (7-17)	8 (1-15)	80 (75-87)
DCIS	73	1 (0-1)	2 (0-4)	97 (95-100)
Invasive carcinoma	23		3 (0-4)	97 (96-100)

Abbreviation: DCIS, ductal carcinoma in situ.

[a] Three reference pathologists, 240 breast biopsy cases.

[b] Range values shown are the minimum and maximum of pathologist level rates for the 3 consensus panel reference pathologists.

Pathologist Identification, Recruitment, and Baseline Characteristics

We used publicly available information from 8 US states (Alaska, Maine, Minnesota, New Hampshire, New Mexico, Oregon, Vermont, and Washington) to invite pathologists to participate in this study (**Figure 2**). Pathologists interpreting breast specimens for at least 1 year with plans to continue for at least 1 additional year were eligible. Residents and fellows were ineligible.

Selected pathologists were sent an email invitation and, if needed, contacted with 2 follow-up emails, mailed invitations, and telephone follow-up. Participants completed a web-based questionnaire that assessed their demographic and clinical practice characteristics, and attitudes about breast pathology interpretation (eFigure 2 in the Supplement). The questionnaire was developed and pilot tested using cognitive interviewing techniques.[19] To compare clinical and demographic characteristics between participants and nonparticipants, information was obtained on the entire population of invited pathologists from Direct Medical Data.[20]

Test Set Implementation

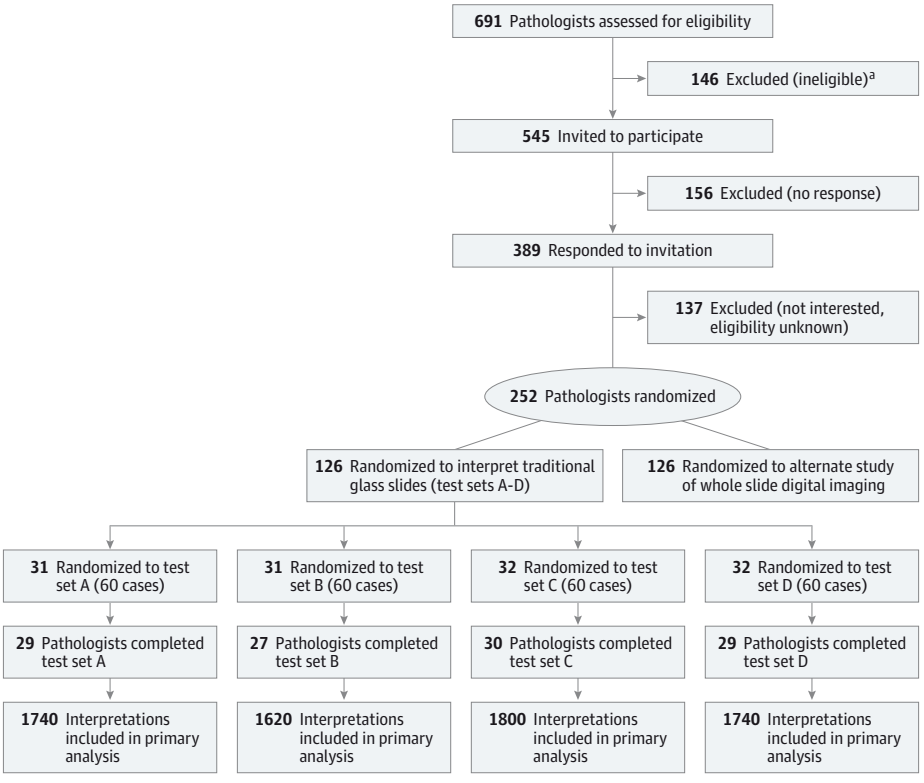
Participants interpreted the same slides as the reference panel members. Participants were randomized with stratification on clinical expertise to ensure equal distribution among the 4 test sets. Clinical expertise was defined as breast pathology fellowship completion, self-assessed perception that peers considered them a breast pathology expert, or both. Participants independently reviewed the 60-case test set in random order. No standardized diagnostic definitions were provided. Participants were asked to interpret the cases as they would in their own clinical practice and complete the diagnostic assessment form online for each case (eFigure 1 in the Supplement).

Participants were provided 1 hematoxylin and eosin-stained slide per case and told the woman's age and type of biopsy. They were not limited by interpretation time. As compensation for their effort, pathologists were offered free category 1 continuing medical education (CME) credits for the slide reviews and an educational program that compared their interpretations with both the consensus-derived reference diagnosis and the other participants' diagnoses. At the completion of the CME, participants were asked questions regarding how the test cases compared with cases they typically see in their practice.

Statistical Analysis

Primary outcome measures included rates of overinterpretation, underinterpretation, and overall concordance. Over-

Figure 2. Pathologist Recruitment and Randomization into Test Sets



[a] Reasons for ineligibility not known.

interpretation was defined as cases classified by the participants at a higher diagnostic category relative to the consensus-derived reference diagnosis; underinterpretation was defined as cases classified lower than the consensus-derived reference diagnosis; concordant cases were those in which the diagnostic category of participants and reference panel were in agreement. Confidence intervals accounted for both within- and between-participant variability by employing variance estimates of the form $\{\text{var}(\text{rate}_p) + [\text{avg}(\text{rate}_p) \times (1 - \text{avg}(\text{rate}_p))] / n_c\} / n_p$, for which $\text{avg}(\text{rate}_p)$ is the average rate among pathologists, $\text{var}(\text{rate}_p)$ is the sample variance of rates among pathologists, n_c is the number of cases interpreted by each pathologist, and n_p is the number of pathologists. We also investigated variability across participants and cases by examining distributions of participant and case-specific rates.

We investigated the extent to which experience of the participant and specific patient characteristics (age, breast density, and biopsy type) were associated with concordance. Logistic regression models of participant misclassification that simultaneously incorporated several pathologist characteristics (academic affiliation, breast-specific caseload, clinical expertise, and practice size) were modeled, and coefficients were tested with a bootstrap technique that resampled participant data.

Sensitivity analyses were performed to determine if the results were altered by use of a different diagnostic mapping scheme or by use of an alternate reference standard diagnosis instead of the expert-derived standard. First, we reanalyzed the data using an alternative diagnostic mapping strategy shown in eTable 1 in the Supplement. Second, we identified cases for which the 3 reference panel members' independent assessments did not unanimously agree and for which the consensus-derived reference diagnosis was different from the most frequent diagnosis recorded by the participants (17 of 240 cases). We reanalyzed the data by substituting the most frequent participant diagnosis as the reference diagnosis for the 17 cases, or by excluding the 17 cases. Testing was 2-sided using a *P* value of less than .05 for significance. Stata statistical software (StataCorp), version 13, was used.

Results

Test Set Cases

Nearly half of the 240 cases were from women aged 40 to 49 years (49%); the remainder were from women aged 50 to 59 years (28%), 60 to 69 years (12%), and 70 years or older (11%). Breast density categories assessed on previous mammography included almost entirely fat (5.4%), scattered fibroglandular densities (43.8%), heterogeneously dense (40.4%), and extremely dense (10.4%) categories. Cases were from both core needle (57.5%) and excisional (42.5%) biopsies. Among the final sample of 240 cases, 72 (30%) were benign without atypia, 72 (30%) were atypia, 73 (30%) were DCIS, and 23 (10%) were invasive carcinoma.

Table 2. Characteristics of Participating Pathologists (N = 115)

Demographics	No. (%)
Age at survey, y	
33-39	16 (13.9)
40-49	41 (35.7)
50-59	42 (36.5)
≥60	16 (13.9)
Sex	
Men	69 (60.0)
Women	46 (40.0)
State of clinical practice	
Alaska	4 (3.5)
Maine	11 (9.6)
Minnesota	19 (16.5)
New Hampshire	4 (3.5)
New Mexico	4 (3.5)
Oregon	15 (13.0)
Vermont	9 (7.8)
Washington	49 (42.6)
Clinical Practice and Breast Pathology Expertise	
Laboratory group practice size	
<10 pathologists	68 (59.1)
≥10 pathologists	47 (40.9)
Fellowship training in breast pathology	
No	109 (94.8)
Yes	6 (5.2)
Affiliated with an academic medical center	
No	87 (75.7)
Yes, adjunct/affiliated clinical faculty	17 (14.8)
Yes, primary appointment	11 (9.6)
Considered an expert in breast pathology by colleagues	
No	90 (78.3)
Yes	25 (21.7)
Years interpreting breast pathology cases (not including residency/fellowship training)	
0-4	22 (19.1)
5-9	23 (20.0)
10-19	34 (29.6)
≥20	36 (31.3)
Percentage of breast specimen interpretation in caseload	
0-9	59 (51.3)
10-24	45 (39.1)
25-49	8 (7.0)
50-74	2 (1.7)
≥75	1 (0.9)
No. of breast cases interpreted per week	
0-4	31 (27.0)
5-9	44 (38.3)
10-19	31 (27.0)
20-29	4 (3.5)
30-39	3 (2.6)
40-49	1 (0.9)
≥50	1 (0.9)
Impressions About Breast Pathology	
Confidence in assessments of breast cases	
1 (Very confident)	14 (12.2)
2	66 (57.4)
3	27 (23.5)
4	8 (7.0)
5	0
6 (Not confident at all)	0
Challenge of interpreting breast cases	
1 (Very easy)	1 (0.9)
2	13 (11.3)
3	43 (37.4)
4	44 (38.3)
5	14 (12.2)
6 (Very challenging)	0
More nervous interpreting breast pathology than other types of pathology	
1 (Strongly disagree)	13 (11.3)
2	35 (30.4)
3	16 (13.9)
4	28 (24.3)
5	20 (17.4)
6 (Strongly agree)	3 (2.6)
Enjoys interpreting breast pathology	
1 (Strongly disagree)	0
2	9 (7.8)
3	13 (11.3)
4	27 (23.5)
5	46 (40.0)
6 (Strongly agree)	20 (17.4)

Pathologist Participation and Characteristics

Rates of pathologist recruitment, which began November 2011, are shown in Figure 2. Among 691 pathologists invited, 146 were ineligible (21.1%). We were unable to contact or verify eligibility for 156 pathologists (22.6%), despite multiple email, postal mail, and telephone contact attempts. Among the remaining 389 pathologists, 137 (35%) declined and 252 (65%) agreed to participate. There were no statistically significant differences in mean age, sex, level of direct medical care, or proportion working in a population of 250 000 or more between the participants and those who declined or those we were unable to contact. Among the 252 participants, 126 participants were randomized to the current study and 91% (115 of 126 participants) completed independent interpretation of all 60 cases and full participation in the study by May 2014. The remaining 126 participants were offered participation in a related future study.

Participants' characteristics and clinical experience are shown in **Table 2**. Although most (93.1%) reported confidence interpreting breast pathology, 50.5% reported that breast pathology is challenging and 44.3% reported that breast pathology makes them more nervous than other types of pathology. The mean CME credits awarded for self-reported time

Figure 3. Comparison of 115 Participating Pathologists' Interpretations vs the Consensus-Derived Reference Diagnosis for 6900 Total Case Interpretations[a]

		Participating Pathologists' Interpretation				
		Benign without atypia	Atypia	DCIS	Invasive carcinoma	Total
Consensus Reference Diagnosis[b]	Benign without atypia	1803	200	46	21	2070
	Atypia	719	990	353	8	2070
	DCIS	133	146	1764	54	2097
	Invasive carcinoma	3	0	23	637	663
	Total	2658	1336	2186	720	6900

DCIS indicates ductal carcinoma in situ.

[a] Concordance noted in 5194 of 6900 case interpretations or 75.3%.

[b] Reference diagnosis was obtained from consensus of 3 experienced breast pathologists.

spent on this activity was 16 (95% CI, 15-17); 43 participants were awarded the maximum 20 hours.

Pathologists' Diagnoses Compared With Consensus-Derived Reference Diagnoses

The 115 participants each interpreted 60 cases, providing 6900 total individual interpretations for comparison with the consensus-derived reference diagnoses (**Figure 3**). Participants agreed with the consensus-derived reference diagnosis for 75.3% of the interpretations (95% CI, 73.4%-77.0%). Participants (n = 94) who completed the CME activity reported that the test cases were similar to the entire spectrum of breast pathology seen in their own practice (23% reported that they always saw cases like the study test cases, 51% often saw cases like these, 22% sometimes saw cases like these, no participants marked never, and 3% did not respond to this question).

In general, overinterpretation and underinterpretation of breast biopsy cases was not limited to a few cases or a few practicing pathologists but was widely distributed among pathologists (N = 115) and cases (N = 240) (eFigure 3A and 3B in the Supplement, respectively). The overall concordance rate for the invasive breast cancer cases was high, at 96% (95% CI, 94%-97%; **Table 3**), although 1 of the invasive test cases contained predominately DCIS with a focus of microinvasion. This focus was initially missed by 2 reference panelists, but was confirmed to be invasive during a consensus meeting.

The participants agreed with the consensus-derived reference diagnosis on less than half of the atypia cases, with a concordance rate of 48% (95% CI, 44%-52%; Figure 3, **Figure 4**, **Figure 5;** and eFigure 4 in the Supplement). Although overinterpretation of DCIS as invasive carcinoma occurred in only 3% (95% CI, 2%-4%), overinterpretation of atypia was noted in 17% (95% CI, 15%-21%) and overinterpretation of benign without atypia was noted in 13% (95% CI, 11%-15%). Underinterpretation of invasive breast cancer was noted in 4% (95% CI, 3%-6%), whereas underinterpretation of DCIS was noted in 13% (95% CI, 12%-15%) and underinterpretation of atypia was noted in 35% (95% CI, 31%-39%).

Diagnostic agreement did not change substantially when we used an alternate diagnostic mapping schema or an alternative participant-based method of defining the reference diagnosis (eTable 1 in the Supplement).

Patient and Pathologist Characteristics Associated With Overinterpretation and Underinterpretation

The association of breast density with overall pathologists' concordance (as well as both overinterpretation and underinterpretation rates) was statistically significant, as shown in Table 3 when comparing mammographic density grouped into 2 categories (low density vs high density). The overall concordance estimates also decreased consistently with increasing breast density across all 4 Breast Imaging-Reporting and Data System (BI-RADS) density categories: BI-RADS A, 81% (95% CI, 75%-86%); BI-RADS B, 77% (95% CI, 75%-79%); BI-RADS C, 74% (95% CI, 72%-76%); and BI-RADS D, 70% (95% CI, 64%-74%); $P < .001$, trend test. Overinterpretation rates were also significantly higher for breast biopsies from women in their 40s (vs ≥50 years), although underinterpretation rates were lower for women in their 40s (vs ≥50 years) (Table 3). The magnitude of the overall density association did not change when covariates for patient age and diagnosis (eg, benign, atypia, DCIS, and invasive) were included in a multivariable model.

Pathologists from outside of academic settings, those who interpret lower weekly volumes of breast cases and those from small-sized practices were statistically significantly less likely to agree with the consensus-derived reference diagnosis. Each of these pathologist variables remained statistically significant in a multivariable logistic model that accounted for the simultaneous contribution of all 3 (eTable 3 in the Supplement). Although the differences noted for pathologist characteristics and patient age and breast density are statistically significant, the absolute effects are small.

Discordance was higher when the pathologists indicated a case was difficult, borderline, they desired a second opinion, or when they reported low confidence in their assessment (Table 3 and eFigure 5 in the Supplement).

Discussion

In this study of US pathologists in which diagnostic interpretation was based on a single breast biopsy slide for each case, we found an overall diagnostic concordance rate of 75.3%, with a high level of agreement between the pathologists' and the consensus-derived reference diagnosis for invasive breast can-

cer, and a substantially lower level of agreement for DCIS and atypia. Disagreement with the consensus-derived reference diagnosis was statistically significantly more frequent when breast biopsies were interpreted by pathologists with lower weekly case volume, from nonacademic practices, or smaller practices; and from women with dense breast tissue on mam-

Table 3. Patient, Pathologist, and Case Characteristics and Rates of Overinterpretation, Underinterpretation, and Concordance for the Participating Pathologists' Interpretations vs the Consensus-Derived Reference Diagnosis

Characteristics	No. of Cases	No. of Interpretations	% (95%CI) Rate of Overinterpretation or Underinterpretation vs Reference Diagnosis				% (95%CI) Overall Concordance Rate vs Reference Diagnosis	
			Overinterpretation	*P* Value	Underinterpretation	*P* Value	Concordance	*P* Value
Test Case Patient Characteristics (N = 240 Test Cases)								
Consensus Reference Diagnosis[a]								
Benign without atypia	72	2070	13 (11-15)	<.001		<.001	87 (85-89)	<.001
Atypia	72	2070	17 (15-21)		35 (31-39)		48 (44-52)	
DCIS	73	2097	3 (2-4)		13 (12-15)		84 (82-86)	
Invasive Breast Cancer	23	663			4 (3-6)		96 (94-97)	
Age at time of biopsy, y								
40-49	118	3391	11 (9-13)	.009	14 (12-16)	<.001	76 (73-78)	.45
≥50	122	3509	9 (8-11)		16 (14-18)		75 (73-77)	
Breast density								
Low	118	3391	8 (7-10)	<.001	14 (12-16)	.03	77 (75-80)	<.001[b]
High	122	3509	11 (10-13)		16 (14-18)		73 (71-75)	
Pathologist Characteristics (N = 115 Participants)								
Academic affiliation								
None	87	5220	11 (9-12)	.06	15 (14-17)	.19	74 (72-76)	.007[c]
Adjunct affiliation	17	1020	8 (5-12)		14 (10-19)		78 (74-82)	
Primary academic	11	660	7 (5-11)		12-(8-16)		81 (76-85)	
Estimated No. of breast cases interpreted/week								
<5	31	1860	11 (8-14)	.17	17 (15-21)	.006	72 (68-75)	.001[d]
5-9	44	2640	10 (8-13)		15 (12-18)		75 (72-78)	
10-19	31	1860	9 (6-11)		13 (11-16)		78 (75-81)	
≥20	9	540	9 (5-15)		12 (7-18)		80 (70-87)	
Practice size[e]								
1-9 pathologists	68	4080	10 (8-12)	.81	16 (14-19)	.029	74 (71-76)	.034
≥10 pathologists	47	2820	9 (8-12)		13 (11-15)		78 (75-80)	
Expertise in breast pathology[f]								
Nonexpert	88	5280	10 (9-12)	.41	16 (14-17)	.14	74 (72-76)	.055
Expert	27	1620	9 (7-12)		12 (9-16)		79 (75-82)	
Case Characteristics (N = 6900 Interpretations)								
Difficulty rating								
Low difficulty (1-3)		4829	6 (5-7)	<.001	13 (11-15)	<.001	81 (79-83)	<.001
High difficulty (4-6)		2071	19 (17-22)		19 (16-22)		62 (59-64)	
Second opinion desired								
No		4449	6 (5-7)	<.001	12 (11-14)	<.001	82 (80-84)	<.001
Yes		2451	17 (15-20)		20 (17-23)		63 (60-66)	
Confidence in assessment								
Low (1-3)		5640	8 (7-9)	<.001	13 (12-15)	<.001	79 (77-80)	<.001
High (4-6)		1260	19 (15-24)		21 (17-26)		60 (55-65)	

[a] Values obtained using mapping scheme 1 described in eTable 1 in the Supplement.

[b] A test for trend based on a logistic regression model, which includes a single 4-category ordinal variable for Breast Imaging-Reporting and Data System density yields a *P* value of less than .001.

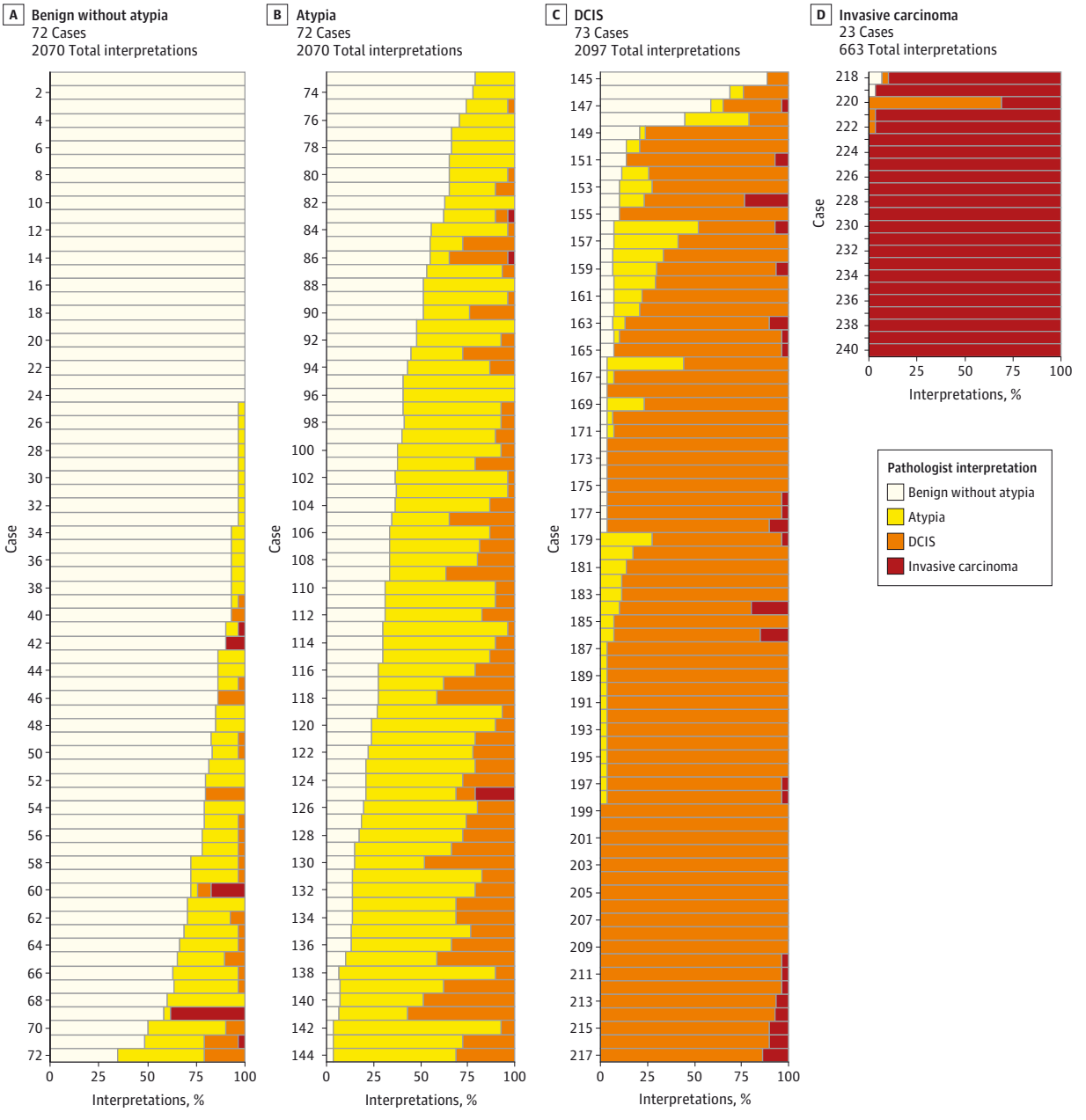
[c] *P* value comparing none vs any academic affiliation (adjunct or primary).

[d] A test for trend based on a logistic regression model, which included a single 4-category ordinal variable for number of cases interpreted per week.

[e] Fewer than 10 pathologists vs 10 or more other pathologists in the same laboratory who also interpret breast tissue.

[f] Clinical expertise defined as self-reported completion of a fellowship in breast pathology or their peers considering them an expert in breast pathology.

Figure 4. Participating Pathologists' Interpretations of Each of the 240 Breast Biopsy Test Cases

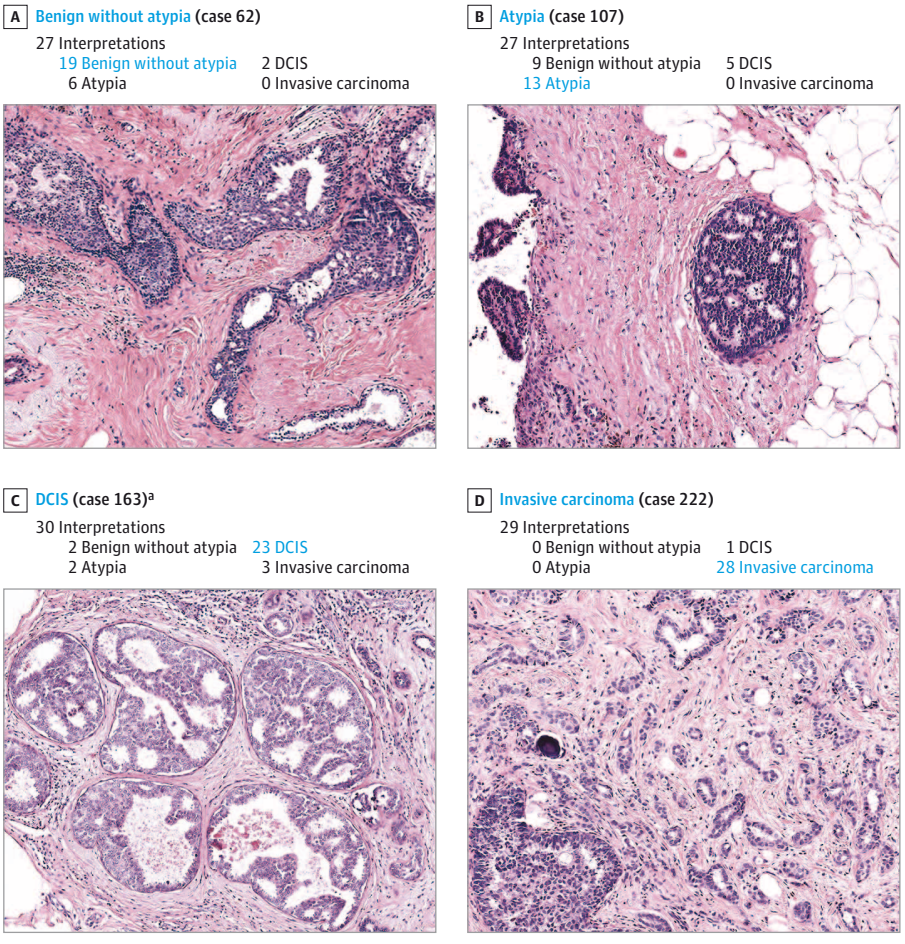


DCIS indicates ductal carcinoma in situ.

mography (vs low density), although the absolute differences in rates according to these factors were generally small.

Most of the 1.6 million breast biopsies performed each year in the United States have benign diagnoses. Our results suggest that overinterpretation of benign without atypia breast biopsies (13% among the 2070 interpretations for 72 benign without atypia cases in this study) may be occurring more often than underinterpretation of invasive breast cancer (4% among 663 interpretations for 23 cases in this study). In addition, although the prevalence of atypia is small (4%-10% of breast biopsies),[3,21] the large number of breast biopsies each year translates into approximately 64 000 to 160 000 women diagnosed with atypia annually. Our results show that atypia is a diagnostic classification with considerable variation among practicing pathologists, with an overall concordance rate of 48% compared with the consensus-derived reference diagnosis. Moreover, among the reference panel members, agreement of their independent preconsensus diagnosis of cases with the final consensus-derived reference diagnosis of atypia was 80%, suggesting that these cases may have the highest possibility of disagreement among pathologists.

Figure 5. Slide Example for Each Diagnostic Category



DCIS indicates ductal carcinoma in situ. Blue indicates concordant interpretations. Each slide is a hematoxylin-eosin stain (original magnification ×100).

[a] Sclerosing adenosis was present elsewhere in this slide.

The variability of pathology interpretations is relevant to concerns about overdiagnosis of atypia and DCIS.[5,22,23] When a biopsy is overinterpreted (eg, interpreted as DCIS by a pathologist when the consensus-derived reference diagnosis is atypia), a woman may undergo unnecessary surgery, radiation, or hormonal therapy.[9,10,24-26] In addition, overinterpretation of atypia in a biopsy with otherwise benign findings can result in unnecessary heightened surveillance, clinical intervention, costs, and anxiety.[27-30] It has been recently suggested that women with a diagnosis of atypia on a breast biopsy consider annual screening magnetic resonance imaging examinations and chemoprevention.[31] Given our findings, clinicians and patients may want to obtain a formal second opinion for breast atypia prior to initiating more intensive surveillance or risk reduction using chemoprevention or surgery.

The rates of overinterpretation and underinterpretation we observed for assessments of atypia and DCIS highlight important issues in breast pathology. However, diagnostic variability is not confined to this specialty, as reports of observer variability have been noted in other areas of clinical medicine.[32,33] For example, extensive variability among radiologists has been noted in the interpretation of mammograms.[34] In addition, results of this study document disagreements even among experienced and expert pathologists.

A unique aspect of our study is that we identified patient and pathologist characteristics associated with greater discordance to explore possible approaches to reducing discordance. In this study, biopsies from women with dense breast tissue on mammography compared with biopsies from women with less-dense breast tissue were more likely to have discordant pathology diagnoses (concordance rate, 73% for dense vs 77% for less-dense). Mammographic density is primarily attributable to increased fibrous tissue in the breast, and it is unlikely that this would contribute to diagnostic discordance. However, microenvironmental factors in dense breast tissue may also be associated with epithelial hyperplasia that may increase diagnostic discordance. Recently there have been efforts to educate women about the association of breast density with screening mammography accuracy and efforts to identify better methods to screen women with dense breast tissue.[35] Although our findings related to breast density and the accuracy of pathologists' interpretations were statistically significant, the absolute differences were small and their clinical significance should be further investigated.

Pathologists with higher clinical volumes of breast pathology and who work within larger group practices had less discordance. Experience and informal learning obtained within larger group practices may contribute to improving and maintaining interpretive performance. We also noted that, to some extent, pathologists could perceive when their interpretation may deviate from the reference diagnosis. For example, participants' diagnoses were more likely to disagree with the reference diagnosis when they indicated the diagnosis was unclear, or when they were less confident in their interpretation. In clinical practice, these factors may prompt a second consultative opinion, an option not allowed in our study environment. Understanding how second opinions may improve diagnostic accuracy is an area requiring further investigation.

Although diagnostic disagreement among breast pathologists has been noted in the past, most previous studies were published in the 1990s; had small numbers of test cases; employed cases that were not randomly selected; and included a smaller number of participants who were specialists in breast pathology.[9-11,24,36-39] In contrast, our study used standardized data on 240 randomly selected cases using a stratified sampling scheme that oversampled cases of atypia and DCIS to improve confidence in agreement estimates. We also enrolled 115 practicing pathologists from diverse geographic locations and clinical settings in 8 US states, providing 6900 individual case assessments. The high participation rate and commitment of the practicing pathologists participating in our study, with most investing 15 hours to 17 hours, is likely related to a desire to improve their diagnostic skills in a challenging clinical area.

Our study findings should be interpreted considering several important limitations. First, it is unclear how the use of test sets, weighted with more cases of difficult and problematic lesions, may have influenced interpretive performance. However, it is not feasible to add such a high number of test cases into a practicing pathologist's daily routine in a blinded fashion. Second, we used only a single slide per case to enhance participation. In clinical practice, pathologists typically review multiple slides per case and can request additional levels or ancillary immunohistochemical stains prior to arriving at a final diagnosis. Third, although no perfect gold standard exists for defining accuracy in pathology diagnosis, we used a carefully defined reference diagnosis based on a consensus of experienced breast pathologists. Among the 3 consensus panel members, unanimous agreement of their independent diagnoses was noted for 75% of cases. Moreover, there is no evidence that the classifications of the consensus panel members are more accurate with respect to predicting clinical outcomes than the classifications of the participating pathologists. However, we noted little change in results after considering alternative methods of defining the reference diagnosis. Fourth, diagnoses rendered in this study setting may not reflect those rendered in actual clinical practice due to subtle variations in the application of criteria or to different emphasis placed on the influence of clinical management. No attempt was made to standardize diagnostic criteria among participants through either written instructions or training slide sets. Fifth, no specific instructions were provided to participants regarding whether their diagnoses should be made purely on morphologic features, or whether biopsy type or clinical management should be considered. We have previously described some of the possible reasons for observer variability in the interpretation of research breast biopsies.[15]

Conclusions

In this study of pathologists, in which diagnostic interpretation was based on a single breast biopsy slide, overall agreement between the individual pathologists' interpretations and the expert consensus-derived reference diagnoses was 75.3%, with the highest level of concordance for invasive carcinoma and lower levels of concordance for DCIS and atypia. Further research is needed to understand the relationship of these findings with patient management.

ARTICLE INFORMATION

Author Affiliations: Department of Medicine, University of Washington School of Medicine, Seattle (Elmore); Program in Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center, Seattle, Washington (Longton, Pepe); Department of Family Medicine, Oregon Health and Science University, Portland (Carney); Department of Family Medicine, University of Vermont, Vineyard Haven, Massachusetts (Geller); Department of Community and Family Medicine, The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine at Dartmouth, Norris Cotton Cancer Center, Lebanon, New Hampshire (Onega, Tosteson); Department of Medicine, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire (Tosteson); Providence Cancer Center, Providence Health and Services Oregon, Portland (Nelson); Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland (Nelson); Department of Clinical Epidemiology and Medicine, Oregon Health and Science University, Portland (Nelson); Department of Pathology, Stanford University School of Medicine, Stanford, California (Allison); Department of Pathology, Beth Israel Deaconess Medical Center, Boston, Massachusetts (Schnitt); Harvard Medical School, Boston, Massachusetts (Schnitt); Department of Laboratory Medicine and the Keenan Research Centre of the Li Ka Shing Knowledge Institute, Toronto, Ontario, Canada (O'Malley); St Michael's Hospital and the University of Toronto, Ontario, Canada (O'Malley); Department of Pathology and University of Vermont Cancer Center, University of Vermont, Burlington (Weaver).

Author Contributions: Drs Elmore and Pepe and Mr Longton had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: Elmore, Onega, Tosteson, Nelson, Pepe, Allison, Weaver.
Acquisition, analysis, or interpretation of data: All authors.
Drafting of the manuscript: Elmore, Carney, Geller, Nelson, Pepe, Schnitt, Weaver.
Critical revision of the manuscript for important intellectual content: All authors.
Statistical analysis: Longton, Onega, Pepe.
Obtained funding: Elmore, Carney, Geller, Onega, Tosteson, Nelson, Weaver.
Administrative, technical, or material support: Onega, Tosteson, O'Malley, Weaver.
Study supervision: Elmore, Allison, Weaver.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Dr Elmore reports serving as a medical editor for the nonprofit Informed Medical Decisions Foundation. Dr Allison reports personal fees from Genentech. No other authors had potential conflicts of interest to report.

Funding/Support: This work was supported by the National Cancer Institute (R01CA140560, R01CA172343, and K05CA104699) and by the National Cancer Institute-funded Breast Cancer Surveillance Consortium (U01CA70013 and HHSN261201100031C). The collection of cancer

and vital status data used in this study was supported in part by several state public health departments and cancer registries throughout the United States. For a full description of sources, visit http://www.breastscreening.cancer.gov/work/acknowledgement.html. The American Medical Association (AMA) is the source for the raw physician data; statistics, tables, or tabulations were prepared by the authors using AMA Physician Masterfile data. The AMA Physician Masterfile data on pathologist age, sex, level of direct medical care, and proportion working in a population of 250 000 or more were used in comparing characteristics of participants and nonparticipants.

Role of the Funders/Sponsors: The funding organization had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the views of the National Cancer Institute or the National Institutes of Health.

Additional Contributions: We thank Ventana Medical Systems, a member of the Roche Group, for use of iScan Coreo Au digital scanning equipment, and HD View SL for the source code used to build our digital viewer.

REFERENCES

1. Silverstein M. Where's the outrage? *J Am Coll Surg*. 2009;208(1):78-79.

2. Silverstein MJ, Recht A, Lagios MD, et al. Special report: Consensus conference III: image-detected breast cancer: state-of-the-art diagnosis and treatment [published correction appears in *J Am Coll Surg*. 2009 Dec;209(6):802]. *J Am Coll Surg*. 2009;209(4):504-520.

3. Weaver DL, Rosenberg RD, Barlow WE, et al. Pathologic findings from the Breast Cancer Surveillance Consortium: population-based outcomes in women undergoing biopsy after screening mammography. *Cancer*. 2006;106(4):732-742.

4. Harris JR, Lippman ME, Morrow M, Osborne CK. *Diseases of the Breast*. 5th ed. Philadelphia, PA: Wolters Kluwer Health; 2014.

5. Bleyer A, Welch HG. Effect of 3 decades of screening mammography on breast-cancer incidence. *N Engl J Med*. 2012;367(21):1998-2005.

6. Hall FM. Identification, biopsy, and treatment of poorly understood premalignant, in situ, and indolent low-grade cancers: are we becoming victims of our own success? *Radiology*. 2010;254(3):655-659.

7. O'Malley FP, Pinder SE, Mulligan AM. *Breast pathology*. Philadelphia, PA: Elsevier/Saunders; 2011.

8. Schnitt SJ, Collins LC. *Biopsy interpretation of the breast*. Philadelphia, PA: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2009.

9. Rosai J. Borderline epithelial lesions of the breast. *Am J Surg Pathol*. 1991;15(3):209-221.

10. 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(12):1133-1143.

11. Wells WA, Carney PA, Eliassen MS, Tosteson AN, Greenberg ER. Statewide study of diagnostic agreement in breast pathology. *J Natl Cancer Inst*. 1998;90(2):142-145.

12. Della Mea V, Puglisi F, Bonzanini M, et al. Fine-needle aspiration cytology of the breast: a preliminary report on telepathology through Internet multimedia electronic mail. *Mod Pathol*. 1997;10(6):636-641.

13. Oster NV, Carney PA, Allison KH, et al. Development of a diagnostic test set to assess agreement in breast pathology: practical application of the Guidelines for Reporting Reliability and Agreement Studies (GRRAS). *BMC Womens Health*. 2013;13(1):3.

14. Feng S, Weaver DL, Carney PA, et al. A framework for evaluating diagnostic discordance in pathology discovered during research studies. *Arch Pathol Lab Med*. 2014;138(7):955-961.

15. Allison KH, Reisch LM, Carney PA, et al. Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel. *Histopathology*. 2014;65(2):240-251.

16. National Cancer Institute. Breast cancer surveillance consortium. http://breastscreening.cancer.gov/. Accessed June 1, 2011.

17. Ginsburg OM, Martin LJ, Boyd NF. Mammographic density, lobular involution, and risk of breast cancer. *Br J Cancer*. 2008;99(9):1369-1374.

18. Helmer O. The systematic use of expert judgment in operations research. http://www.rand.org/pubs/papers/P2795.html. Accessed March 27, 2012.

19. Willis GB. *Cognitive Interviewing: A Tool For Improving Questionnaire Design*. Thousand Oaks, CA:Sage Publications; 2005.

20. American Medical Association. Physicians. http://www.dmddata.com/data_lists_physicians.asp. Accessed January 27, 2015.

21. Rubin E, Visscher DW, Alexander RW, Urist MM, Maddox WA. Proliferative disease and atypia in biopsies performed for nonpalpable lesions detected mammographically. *Cancer*. 1988;61(10):2077-2082.

22. Zahl PH, Jørgensen KJ, Gøtzsche PC. Overestimated lead times in cancer screening has led to substantial underestimation of overdiagnosis. *Br J Cancer*. 2013;109(7):2014-2019.

23. Gøtzsche PC, Jørgensen KJ. Screening for breast cancer with mammography. *Cochrane Database Syst Rev*. 2013;6(6):CD001877.

24. Collins LC, Connolly JL, Page DL, et al. Diagnostic agreement in the evaluation of image-guided breast core needle biopsies: results from a randomized clinical trial. *Am J Surg Pathol*. 2004;28(1):126-131.

25. Haas JS, Cook EF, Puopolo AL, Burstin HR, Brennan TA. Differences in the quality of care for women with an abnormal mammogram or breast complaint. *J Gen Intern Med*. 2000;15(5):321-328.

26. Saul S. Prone to error: earliest steps to find cancer. *New York Times*. July 19, 2010. http://www.nytimes.com/2010/07/20/health/20cancer.html?pagewanted=all&_r=0. Accessed February 16, 2015.

27. Rakovitch E, Mihai A, Pignol JP, et al. Is expert breast pathology assessment necessary for the management of ductal carcinoma in situ? *Breast Cancer Res Treat*. 2004;87(3):265-272.

28. Dupont WD, Page DL. Risk factors for breast cancer in women with proliferative breast disease. *N Engl J Med*. 1985;312(3):146-151.

29. Dupont WD, Parl FF, Hartmann WH, et al. Breast cancer risk associated with proliferative breast disease and atypical hyperplasia. *Cancer*. 1993;71(4):1258-1265.

30. London SJ, Connolly JL, Schnitt SJ, Colditz GA. A prospective study of benign breast disease and the risk of breast cancer. *JAMA*. 1992;267(7):941-944.

31. Hartmann LC, Degnim AC, Santen RJ, Dupont WD, Ghosh K. Atypical hyperplasia of the breast—risk assessment and management options. *N Engl J Med*. 2015;372(1):78-89.

32. Feinstein AR. A bibliography of publications on observer variability. *J Chronic Dis*. 1985;38(8):619-632.

33. Elmore JG, Feinstein AR. A bibliography of publications on observer variability (final installment). *J Clin Epidemiol*. 1992;45(6):567-580.

34. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. *N Engl J Med*. 1994;331(22):1493-1499.

35. Lee CI, Bassett LW, Lehman CD. Breast density legislation and opportunities for patient-centered outcomes research. *Radiology*. 2012;264(3):632-636.

36. Carney PA, Eliassen MS, Wells WA, Swartz WG. Can we improve breast pathology reporting practices? a community-based breast pathology quality improvement program in New Hampshire. *J Community Health*. 1998;23(2):85-98.

37. Trocchi P, Ursin G, Kuss O, et al. Mammographic density and inter-observer variability of pathologic evaluation of core biopsies among women with mammographic abnormalities. *BMC Cancer*. 2012;12:554.

38. Shaw EC, Hanby AM, Wheeler K, et al. Observer agreement comparing the use of virtual slides with glass slides in the pathology review component of the POSH breast cancer cohort study. *J Clin Pathol*. 2012;65(5):403-408.

39. Stang A, Trocchi P, Ruschke K, et al. Factors influencing the agreement on histopathological assessments of breast biopsies among pathologists. *Histopathology*. 2011;59(5):939-949.