



Published in final edited form as:

Mod Pathol. 2016 September ; 29(9): 1004–1011. doi:10.1038/modpathol.2016.85.

Region of Interest Identification and Diagnostic Agreement in Breast Pathology

Dilip B. Nagarkar¹, Ezgi Mercan², Donald L. Weaver, MD³, Tad T. Brunyé, PhD⁴, Patricia A. Carney, PhD⁵, Mara H. Rendi, MD, PhD⁶, Andrew H Beck, MD, PhD⁷, Paul Frederick, MPH, MBA¹, Linda G. Shapiro, PhD², and Joann G. Elmore, MD, MPH¹

¹Department of Medicine, University of Washington School of Medicine

²Department of Computer Science and Engineering, University of Washington

³Department of Pathology and UVM Cancer Center, University of VT, Burlington, VT

⁴Department of Psychology, Tufts University

⁵Department of Family Medicine, Oregon Health & Science University

⁶Department of Pathology, University of Washington School of Medicine

⁷Department of Pathology, Beth Israel Deaconess Medical Center, Harvard Medical School

Abstract

A pathologist's accurate interpretation relies on identifying relevant histopathological features. Little is known about the precise relationship between feature identification and diagnostic decision making. We hypothesized that greater overlap between a pathologist's selected diagnostic region of interest (ROI) and a consensus derived ROI is associated with higher diagnostic accuracy. We developed breast biopsy test cases which included atypical ductal hyperplasia (n=80); ductal carcinoma in situ (n=78); and invasive breast cancer (n=22). Benign cases were excluded due to absence of specific abnormalities. Three experienced breast pathologists conducted an independent review of the 180 digital whole slide images, established a reference consensus diagnosis and marked one or more diagnostic ROIs for each case. Forty-four participating pathologists independently diagnosed and marked ROIs on the images. Participant diagnoses, and ROI were compared to consensus reference diagnoses and ROI. Regression models tested whether percent overlap between participant ROI and consensus reference ROI predicted diagnostic accuracy. Each of the 44 participants interpreted 39–50 cases for a total of 1,972 individual diagnoses. Percent ROI overlap with the expert reference ROI was higher in pathologists who self-reported academic affiliation (69% vs 65%, p=0.002). Percent overlap between participants ROI and consensus reference ROI was then classified into ordinal categories: 0%, 1%–33%, 34%–66%, 66%–99%, and 100% overlap. For each incremental change in the ordinal percent ROI overlap, diagnostic agreement increased by 60% (OR 1.6, 95% CI [1.5–1.7],

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use:http://www.nature.com/authors/editorial_policies/license.html#terms

Correspondence to: Joann G. Elmore, MD, MPH, University of Washington, Mailbox 359780, 325 Ninth Avenue, Seattle, WA 98104, Telephone: (206) 744-3632, Fax: (206) 744-9917, jelmore@uw.edu.

Disclosure and conflict of interest: The authors do not have any financial conflicts of interest.

p<0.001) and the association remained significant even after adjustment for other covariates. The magnitude of the association between ROI overlap and diagnostic agreement increased with increasing diagnostic severity. The findings indicate that pathologists are more likely to converge with an expert reference diagnosis when they identify an overlapping diagnostic image region, suggesting that future computer-aided detection systems that highlight potential diagnostic regions could be a helpful tool to improve accuracy and education.

INTRODUCTION

Each year, millions of breast biopsies are performed, yet interpreting such specimens is considered to be one of the more challenging areas in pathology.(1–5) While evaluating a breast biopsy slide, it is critical that the pathologist identifies and then analyzes regions of potential diagnostic interest that might support criteria for diagnosing breast cancer or diagnosing risk-associated non-invasive breast lesions. Pathologists use a complex set of skills to establish a histopathological diagnosis when interpreting a biopsy slide. At least two of these skills may be amenable to computer-assisted image identification and analysis, including: 1) finding relevant diagnostic regions (e.g., salient visual features) and 2) interpreting contextual architectural and cytological features in epithelial proliferations.

The fields of pathology and radiology are similar in requiring interpretation of an image to arrive at a diagnosis. Research in interpreting radiology images suggests that searching and diagnosis are possibly separate skills.(6) By inference, insight may be gained by studying how pathologists search for and identify regions of interest and then how they diagnose these regions. Digital whole slide imaging in pathology may facilitate this research and lead to future educational and clinical support tools.

In this study, we explored the relationship between areas that pathologists indicated as diagnostic regions of interest (ROI) on whole slide digital images and their diagnostic accuracy. An expert consensus diagnosis on each case was used to define accuracy. We hypothesized that as the electronically-marked ROI exhibited increasing overlap with the expert consensus ROI, agreement with the consensus reference diagnosis would increase. This is an intuitive hypothesis for highly reproducible diagnoses such as invasive carcinoma; however, for more ambiguous diagnoses, a “correct” diagnosis could be based on irrelevant features. Additionally, the method to test this hypothesis is unique and demonstrates the advantages of digital whole slide imaging. Thus, an evaluation of pathologist indicated diagnostic ROI would be informative and novel.

MATERIALS AND METHODS

A cohort of cases from the Breast Pathology Study (B-Path) was used for this analysis. The methods for test case identification and the development and recruitment of pathologists have been previously described (3, 7). Briefly, single, representative diagnostic slides from excisional or core breast biopsies of 180 women were included in this pilot study. Each slide was digitally scanned (iScan Coreo, Ventana Medical Systems, Tucson, AZ), and a whole slide image (WSI) was created, allowing the digital virtual slide to be viewed, magnified, and annotated on a computer using a web-based viewer.

Case and Consensus Reference Data

Three experienced breast pathologists, who were involved in the original B-Path study (3, 6), independently interpreted the 180 cases in the digital WSI format using a standardized diagnosis reporting form and a web-based viewer. Each expert recorded an independent diagnosis and digitally marked an ROI that best exemplified the critical features on the slide, supporting the most severe diagnosis. The electronic ROI tool allowed the experts to select a rectangular region that could be as small or as large as the pathologist desired. Webinars and in-person meetings of these same three experts were then held, using a modified Delphi approach (6) to establish a final consensus reference diagnosis and one or more consensus ROIs that supported the diagnosis for each case. A consensus diagnosis was achieved for 176 cases using the categorical mapping scheme described in Appendix 1. The remaining four cases were not reviewed during in-person meetings due to time and logistical constraints; thus, the consensus diagnosis for these cases was defined by a majority agreement between the experts' independent diagnoses (two of three experts agreed).

Consensus on the ROI markings was achieved for all 180 cases. The final consensus ROI was a single rectangle, multiple overlapping rectangles, or two or three independent rectangles containing the regions that the expert panel agreed would support the final consensus diagnosis. Partially overlapping ROIs were merged. The consensus process could select a region of any size up to the entire slide, multiple locations, and multisided polygonal regions made up of overlapping rectangles.

The expert panel's categorical diagnoses included atypia (n=80), ductal carcinoma in situ (DCIS; n=78), and invasive breast cancer (n=22) per the classification scheme in Appendix 1. Additional benign test cases that were part of the larger B-Path study were excluded in this pilot study, as there was often no specific abnormality present on these slides to mark as an ROI. Standardized data for each case were available on the woman's age, the type of biopsy, and her breast tissue density as assessed on the mammography preceding the biopsy using the BI-RADS 4-grade scale.(8) In the four increasing BI-RADS breast density categories, from 1 through 4, there were 11, 75, 76, and 18 cases, respectively. For analyses, breast density was collapsed into two categories, low- (categories 1 and 2) and high- density (categories 3 and 4).

Participant Data

Pathologists from eight U.S. states (Alaska, Maine, Minnesota, New Hampshire, New Mexico, Oregon, Vermont, and Washington), who interpret breast biopsies in their clinical practices, were invited to participate in this study. Participants independently interpreted one of the test sets. While all test sets included 60 cases in the main B-Path study, after exclusion of benign without atypia cases and retention of cases with incidental examples of flat epithelial atypia and atypical lobular hyperplasia, as defined in **Appendix 1**, the analytic cases for this study included 41, 44, 45, and 50 cases from the four B-Path test sets, respectively. The participating pathologists independently identified and marked a single ROI and provided a diagnosis for each case using the same standardized diagnosis recording form and web-based viewer as the expert reference panel. The participant pathologists could select an ROI area limited to 8603 pixels in height and width, a single region on the slide,

and only a rectangular shaped region. These restrictions ensured that participants marked the features most supporting their diagnosis. Participants were informed of the woman's age and biopsy type for each case at the time of each interpretation.

Participants' interpretations were made independently without knowledge of either the consensus or each other's diagnosis. On each case, participants could select one or more assessment terms, from 14 possible terms. The cumulative number of unique individual assessment terms given to a case by the participants was tabulated for each case. Participants also rated the level of difficulty of each case and their confidence in their assessment using a six-point Likert scale from 1 ('very confident') to 6 ('not confident at all') that was simplified to binary categories for analyses (1, 2, 3 vs. 4, 5, 6). Participants also filled out a baseline survey, providing standardized data on demographics, experience with breast cancer pathology, and perceptions on breast cancer interpretation.

Analysis

The study pathologists' diagnoses and marked ROIs were compared to the consensus reference diagnoses and ROIs (Figure 1). Diagnostic agreement was defined such that participant diagnoses that agreed with the consensus were given a score of one and those that disagreed, a score of zero. The percent ROI overlap was calculated as the pixel area of the ROI selected by the participant that was within the consensus ROI:

$$\% \text{ ROI Overlap} = \frac{O}{P} \times 100$$

Where P is the number of pixels in the participant ROI selection and O is the number of pixels in the region of overlap (union) between the participant and consensus marked ROI selections (Figure 2). Because the consensus ROI could vary in shape, maximum size, number of areas selected, and the participant ROI had a defined maximum size, we defined percent ROI overlap as the proportion of the participant area selected that overlapped with the consensus ROI to avoid any penalty associated with the participant ROI size restriction.

Summary statistics in descriptive tables and regression estimates of percent ROI overlap were obtained using a repeated design model based on generalized estimating equations (GEE) and an independent working correlation structure to account for correlated responses within participants.

Because prominent bimodality was present at 0% and 100% ROI overlap, and to aid interpretation, the ROI overlap was transformed to a categorized ordinal variable (0%, 1%–33%, 33%–66%, 66%–99% and 100%). We identified covariates for inclusion into the multivariable analyses based on a criterion of $p < 0.05$. We fitted the outcome and percent ROI overlap model with diagnostic classification as a moderating variable and any covariate that remained significant in the adjusted model at $p < 0.15$. The categorized ordinal variable reported the change in log odds of agreement for each increment of change in ROI overlap.

To produce odds ratios from significant interactions, the ordinal main effect of ROI overlap was re-parameterized by creating a separate ROI overlap variable for each category within

each diagnostic class. Odds ratios and their corresponding Wald p-values and 95% confidence intervals were calculated to test the significance of the independent effect of ROI on agreement. A P-value of <0.05 (two-tailed) or a 95% confidence interval not including unity, was considered statistically significant. All statistical analyses were conducted using SAS software version 9.4 (SAS Institute, Cary, NC, USA).

RESULTS

Characteristics of the 44 participating pathologists and characteristics for the 180 test cases are shown in Tables 1 and 2 respectively, along with the percentage ROI overlap for each characteristic. A total of 1 972 individual diagnoses on cases with marked ROIs were available for analyses. The majority of pathologists were 50 years of age (68%), male (73%), and worked in a facility with fewer than 10 pathologists (66%) with no affiliation with an academic medical center (75%). The only characteristic associated with higher ROI overlap was academic affiliation ($p = 0.002$). Case characteristics associated with higher ROI overlap included biopsies from women with lower breast density ($p < 0.001$) and higher disease classification ($p < 0.001$). Cases with higher severity diagnoses (invasive and DCIS) had more ROI overlap than the atypia cases. Additionally, the cases with the lowest number of assessment terms applied by the participating pathologists to the case ($p < 0.001$), with assessments that were rated as 'very easy' or 'easy' to interpret ($p = 0.005$), and with no borderline call designation ($p < 0.001$) had significantly higher overlap of participants' ROI markings with the reference ROI markings.

The box plot representation of data (Figure 3) shows that the median percent ROI overlap for assessments with agreement of diagnoses was 89% (interquartile range: 52%–100%). This was significantly higher than the median percent overlap of 53% (interquartile range: 0%–89%) for assessments with disagreement in the diagnosis ($p < 0.001$ by Friedman test). A statistically significant positive trend in diagnostic agreement was noted in the aggregate data for all cases when percent ROI overlap was categorized into five incremental categories of increasing overlap (Figure 4), (p -trend < 0.001).

When stratified by diagnostic classification (atypia, DCIS, invasive cancer), the predominant upward trend exhibited in the aggregate data was less pronounced, particularly for invasive disease (eFigure 1). For example, the results for invasive breast cancer suggest a binary relationship—when there was no ROI overlap, the diagnostic agreement was lower—rather than an upward linear trend. In invasive cancer cases, there was 75% agreement for cases with no overlap and over 91% agreement for cases with any level of ROI overlap. However, the numbers were small and these findings only suggestive. In contrast to invasive cancer, the relationship between percent ROI overlap and diagnostic concordance for atypia cases suggests a linear relationship. The DCIS cases demonstrate an intermediate pattern between invasive and atypia. For DCIS cases, there was 51% agreement with the reference diagnosis when there was no ROI overlap, but agreement increased from 82% to 90%, as the percent ROI overlap increased.

The unadjusted odds of participant diagnostic agreement with the expert consensus diagnosis increased with each incremental increase in ROI overlap category (Figure 5). For

example, the unadjusted odds of agreement with the consensus diagnosis was 2.2 times greater (odds ratio [OR] 2.2, 95% confidence interval [CI] 1.5–3.0, $p < 0.001$) when the area of the ROI selected by the participant encompassed between 1% and 33% of the consensus ROI when compared to the reference of no ROI overlap. When the ROI overlap was 100%, the odds of agreement in the diagnosis were more than seven times greater (OR 7.7, 95% CI [5.2–11.3], $p < 0.001$) (reference 0% overlap in the ROI). The global p -value for this association was $p < 0.001$. The unadjusted ordinal effect of ROI overlap on agreement of the diagnoses increased 60% for each incremental increase in ROI overlap category (OR 1.6, 95% CI [1.5 – 1.7], $p < 0.001$) (data not shown).

In the multivariable model, the interaction term representing ROI overlap and consensus diagnosis was significant ($p < 0.001$), indicating that the association between ROI overlap and agreement with the reference diagnosis was modified by the consensus reference diagnosis category of the case. For each incremental increase in the ordinal predictor of percent ROI overlap, agreement with the consensus reference diagnosis increased 130% when the reference diagnosis was invasive disease (OR 2.3, 95% CI [1.7 – 3.1], $p < 0.001$). For atypia and DCIS, the magnitude of the effect of increasing ROI overlap on concordance with the reference diagnosis was less. The incremental effect of increasing ROI overlap on agreement among DCIS cases was 60% (OR 1.6, 95% CI [1.4 – 1.9], $p < 0.001$). When interpreting atypia cases, the effect of increasing ROI overlap resulted in a 20% increase in agreement (OR 1.2, 95% CI [1.0 – 1.3], $p = 0.01$) (data not shown).

DISCUSSION

Our study investigated whether diagnostic agreement increases as pathologists identify and mark the same region on a slide as annotated by an expert consensus reference standard. In the simplest terms, if two pathologist observers are looking at the same location and features on a slide image, they are more likely to arrive at the same diagnosis. The largest incremental increases in diagnostic concordance were observed between the 0% ROI overlap and the next lowest category (1–33% ROI overlap), especially for the invasive breast cancer cases; smaller incremental increases in diagnostic concordance were observed as percent ROI overlap increased incrementally.

Our findings support the concept that diagnostic accuracy in pathology is dependent, first, on a visual scanning and search process and locating potentially important diagnostic ROIs within medical images, and second, on discriminately focusing on particular diagnostic features within the ROI. These concepts are intuitive to teachers and trainees but are challenging to measure objectively. Our observation that the largest incremental increase in diagnostic concordance occurred between no observed overlap of the ROI annotations and a small amount of overlap supports the concept that searching for and identifying potential diagnostic features is a critical skill.

Eye tracking studies of trainee and practicing pathologists have further shown this scanning and targeted focusing behavior becomes increasingly pronounced as trainees gain advanced experience (9, 10). Digital WSI may provide opportunities to improve the way we train pathologists and how we evaluate histopathology skills, leading to improvements in overall

diagnostic abilities and medical care. The inclusion of automated algorithms to highlight regions of interest may focus a practicing pathologist's attention on potentially diagnostic regions, particularly when non-diagnostic distracting features are present. Automated algorithms used in radiology to accentuate relevant radiographic characteristics have proven effective in improving radiology trainee localization of critical diagnostic features(11). Similarly, masking and unmasking diagnostic ROIs in the field of pathology may help trainees become proficient at identifying diagnostic regions.

The data demonstrate a binary trend for DCIS and invasive cases where any level of ROI marking overlap achieves significantly higher levels of diagnostic agreement with a reference diagnosis. The cases of atypia alone do not demonstrate the same binary pattern. Previous studies have shown that breast atypia is consistently the most challenging diagnostic category for pathologists (3, 12–14). Thus, even if a pathologist reviews a critical image region, recognizing architectural and cytologic features of atypia and accurately assimilating these features into a diagnostic rubric can be challenging.

Increased breast density on mammography has been associated with lower sensitivity of the radiologists' interpretations.(15–17) As breast density increases, there may be significantly more background proliferative change for the pathologist to review on the slide, which could make it more difficult to screen for the most relevant ROI. We did note a trend that percent ROI overlap decreased with increasing breast density. However, our test cohort included only 18 cases with the highest density classification, and thus our analysis was restricted to a binary density classification rather than evaluating the four BI-RADS density categories independently. Thus our data suggests that interpreting the images of denser breast tissue is challenging for both radiologists and pathologists, and further research is indicated.

Technical challenges encountered during this pilot study may help to inform future research and experimental design. Diagnostic agreement of participants was still noted when there was no ROI overlap; values for diagnostic agreement ranged from 40% for atypia cases to 50% for DCIS cases to 75% for invasive carcinoma cases when percent ROI overlap was zero. A non-overlapping participant ROI could theoretically occur for three reasons: 1) the ROI area selected by participants did not meet the reference panel criteria for diagnosis, 2) the reference panel did not include all diagnostic regions within the consensus ROI, and 3) the participant annotated an ROI that did not support their final interpretation. The first two explanations are associated with standard issues associated with diagnostic accuracy: pathologists may have different opinions as to whether a feature set meets diagnostic criteria. The third explanation is a technical limitation. Although the viewer software required annotation of an ROI to complete a case, there was no quality control check to assure that the participant's ROI supported their final diagnosis rather than an intermediate conclusion or whether the ROI even contained breast tissue or epithelium rather than a nonrelevant part of the slide image.

Another technical challenge was that we allowed the reference ROI to be larger and include multiple shapes compared to the participant ROIs. Thus, a future educational tool might benefit from rank-ordering ROIs when multiple reference ROIs are present. The participant ROI had a maximum size limitation and the ROI was constrained to a rectangular shape.

Thus, when the reference ROI was smaller than the participant maximum allowed ROI, the percent ROI overlap could be low if the participant did not carefully draw the ROI around only the diagnostic features. This may have limited our ability to conclude that diagnostic accuracy increases as pathologists focus on the same diagnostic features. Since we observed that diagnostic accuracy generally increased as percent ROI overlap increased, we cannot exclude the conclusion that pathologists who were the most discriminating when drawing an ROI tightly bordering the diagnostic features were also the pathologists who most rigorously applied and interpreted diagnostic criteria. Finally, confluent areas of diagnostic invasive carcinoma were generally large, limiting the association between percent ROI overlap and diagnostic concordance for invasive cancer cases. We also did not evaluate the participants' full interpretive behaviors using computer image analysis or analyze the image content on each slide. Future research should include linking such captured diagnostic image features and linking these to patient outcomes.

Other non-technical limitations included the use of only one slide per case, which may not represent clinical practice. However, this could be beneficial in a study scenario, as it limits the amount of variance in interpretation and can help in better isolating specific tissue characteristics that lead to particular diagnoses. Another limitation is that this study was done only in breast tissue, and it is unknown if these findings would be relevant to other areas in pathology.

Limitations aside, this study is the first of its kind. Strengths of this study include the large number of cases and high number of participants, many of whom spent up to 20 hours participating in the larger study without compensation other than an opportunity to earn continuing medical education credits. Additionally, each test case had a carefully defined expert-based consensus reference diagnosis for comparison and defining accuracy.

In conclusion, this study used digital WSI in a novel manner and demonstrates a potential application of WSI for teaching and improving diagnostic skills of pathologists. Identifying an important region on a histopathology slide image is a significant predictor of diagnostic accuracy and thus may be an indirect indicator of search and screening skills. These findings suggest that computer-aided detection algorithms that highlight potential regions of diagnostic interest on pathology slide images may potentially improve diagnostic accuracy.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under award numbers R01 CA172343, R01 CA140560 and KO5 CA104699. 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. The authors wish to thank Ventana Medical Systems, Inc. (Tucson, AZ, USA), a member of the Roche Group, for use of iScan Coreo Au™ whole slide imaging system, and HD View SL for the source code used to build our digital viewer. For a full description of HD View SL please see <http://hdviewsl.codeplex.com/>.

REFERENCES

1. Silverstein M Where's the outrage? *J Am Coll Surg.* 2009;208(1):78–9. [PubMed: 19228507]
2. Silverstein M, Recht A, Lagios MD, Bleiweiss IJ, W. BP, Gizienski T, et al. Special report: Consensus conference III. Image-detected breast cancer: state-of-the-art diagnosis and treatment. *J Am Coll Surg.* 2009;209(4):504–20. [PubMed: 19801324]
3. Elmore JG, Longton G, Carney PA, Geller B, Onega T, Tosteson ANA, et al. Diagnostic Concordance Among Pathologists Interpreting Breast Biopsy Specimens. *JAMA.* 2015;313(11):1122–32. [PubMed: 25781441]
4. O'Malley FP, Pinder SE, Mulligan AM. *Breast pathology Philadelphia, PA: Elsevier/Saunders; 2011.*
5. Marrow M Increasing Access to Modern Multidisciplinary Breast Cancer Care. DTIC Document, 2002.
6. Oster NV, Carney PA, Allison KH, Weaver DL, Reisch LM, Longton G, 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. [PubMed: 23379630]
7. Beam CA, Conant EF, Sickles EA. Correlation of radiologist rank as a measure of skill in screening and diagnostic interpretation of mammograms. *Radiology.* 2006;238(2):446–53. [PubMed: 16436811]
8. Kopans DB, D'Orsi CJ, Adler DD, Bassett LW, Brenner RJ, Dodd GD. *American College of Radiology Breast Imaging Reporting and Data System (BI-RADS).* Jefferson (VA): American College of Radiology 1993.
9. Krupinski EA, Graham AR, Weinstein RS. Characterizing the development of visual search expertise in pathology residents viewing whole slide images. *Hum Pathol.* 2013;44(3):357–64. [PubMed: 22835956]
10. Krupinski EA, Tillack AA, Richter L, Henderson JT, Bhattacharyya AK, Scott KM, et al. Eye-movement study and human performance using telepathology virtual slides: implications for medical education and differences with experience. *Hum Pathol.* 2006;37(12):1543–56. [PubMed: 17129792]
11. Nodine CF, Liu H, Miller WT, Kundel HL. Observer performance in the localization of tubes and catheters on digital chest images: the role of expertise and image enhancement. *Academic radiology.* 1996;3(10):834–41. [PubMed: 8923902]
12. Schnitt SJ, Connolly JL, Tavassoli FA, Fechner RE, Kempson RL, Gelman R, et al. Interobserver reproducibility in the diagnosis of ductal proliferative breast lesions using standardized criteria. *Am J Surg Pathol.* 1992;16(12):1133–43. [PubMed: 1463092]
13. Rosai J Borderline epithelial lesions of the breast. *The American journal of surgical pathology.* 1991;15(3):209–21. [PubMed: 1847606]
14. Allison KH, Reisch LM, Carney PA, Weaver DL, Schnitt SJ, O'Malley FP, et al. Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel. *Histopathology.* 2014;65(2):240–51. [PubMed: 24511905]
15. Barlow WE, White E, Ballard-Barbash R, Vacek PM, Titus-Ernstoff L, Carney PA, et al. Prospective breast cancer risk prediction model for women undergoing screening mammography. *Journal of the National Cancer Institute.* 2006;98(17):1204–14. [PubMed: 16954473]
16. Tice JA, Cummings SR, Smith-Bindman R, Ichikawa L, Barlow WE, Kerlikowske K. Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. *Ann Intern Med.* 2008;148(5):337–47. [PubMed: 18316752]
17. Sprague BL, Gangnon RE, Burt V, Trentham-Dietz A, Hampton JM, Wellman RD, et al. Prevalence of mammographically dense breasts in the United States. *J Natl Cancer Inst.* 2014;106(10).

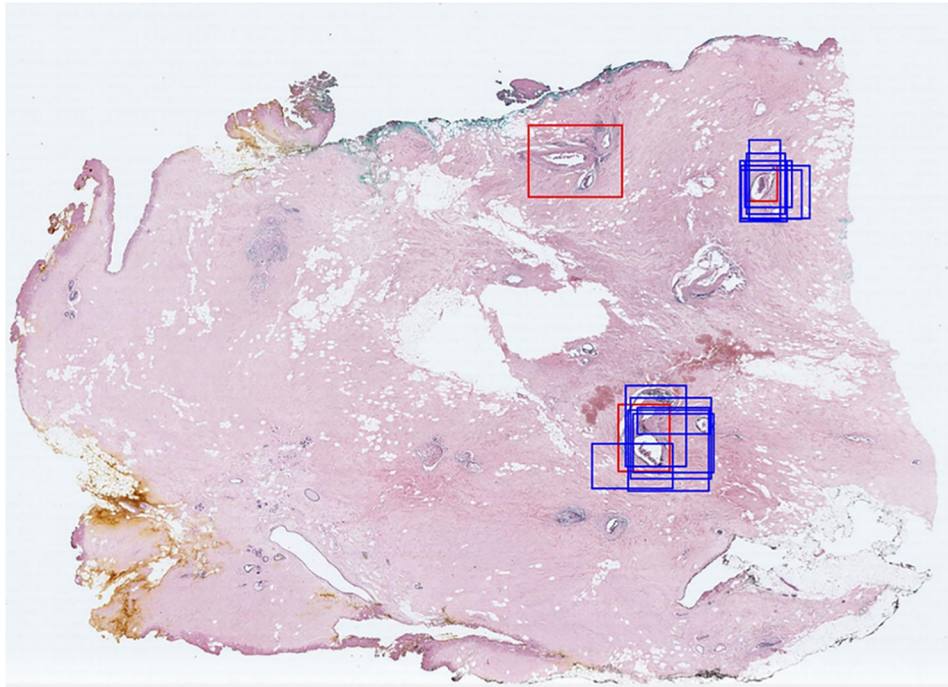


Figure 1. Example breast biopsy slide, Hematoxylin and eosin stain, demonstrating three different markings for the consensus reference regions of interest (ROI) (shown in red) and the ROI annotations for 12 participants (shown in blue). Participants were instructed to select a single ROI that supported their diagnosis.

$$\% ROI \text{ Overlap} = \frac{\text{Pixels in } O}{\text{Pixels in } P} \times 100$$

O = Overlap
 P = Participant ROI
 C = Consensus ROI

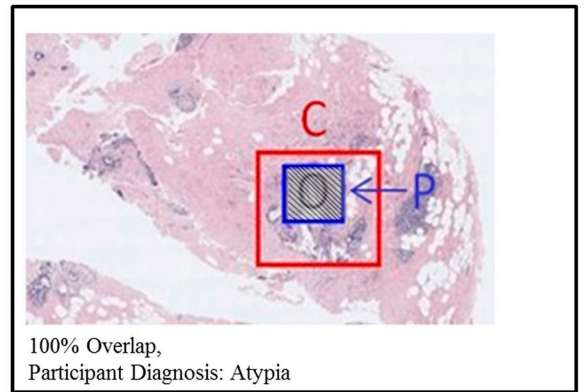
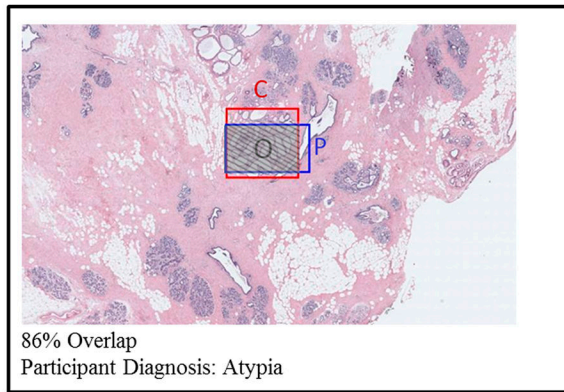
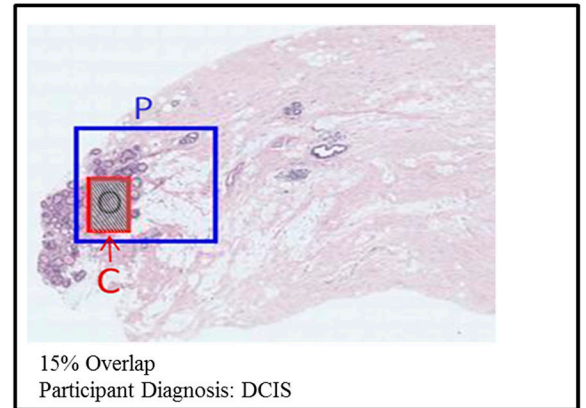
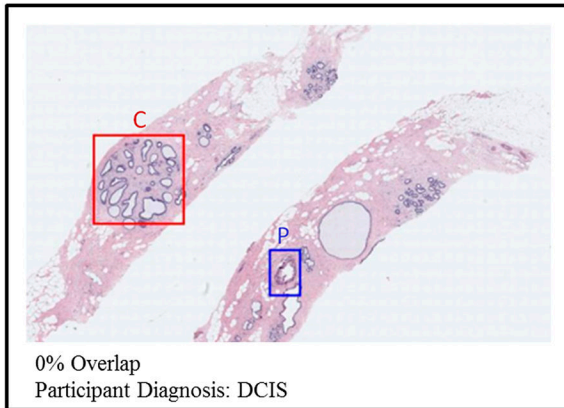


Figure 2. Example cases defined as atypia by the expert consensus with expert consensus ROI markings (in red), participant ROI markings (in blue), and example determination of percent ROI overlap for the case.

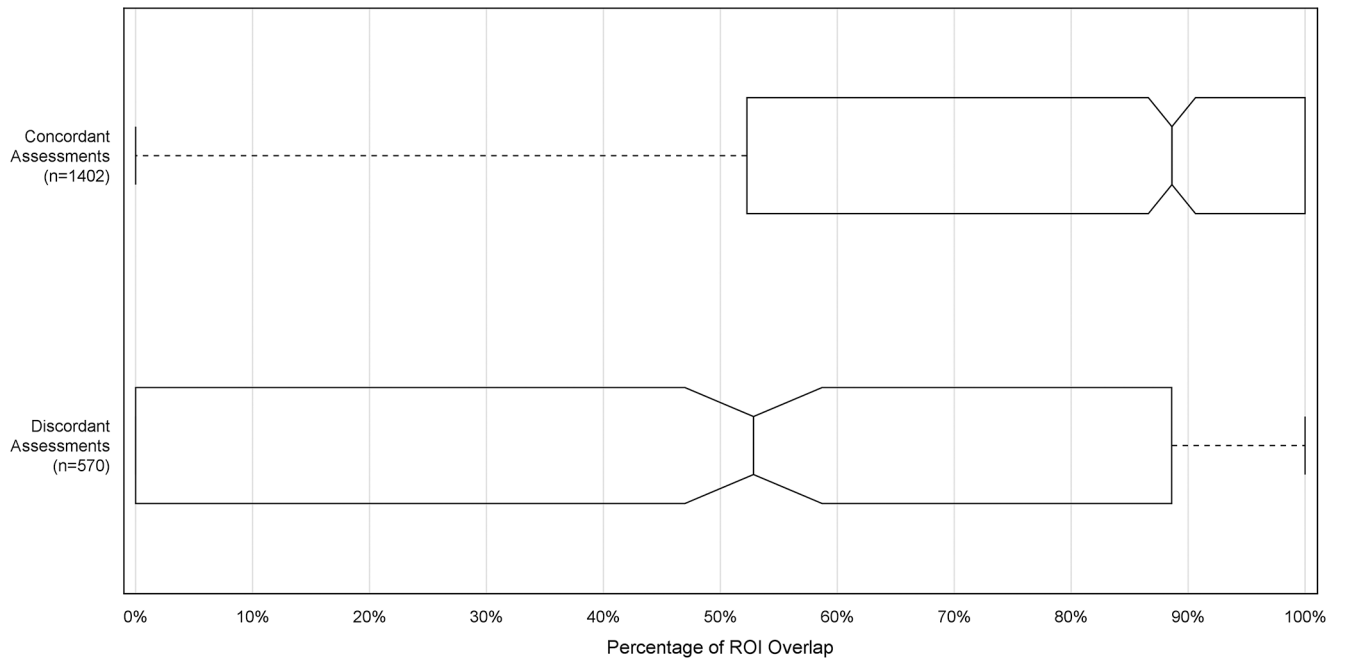


Figure 3. Box plot of percent region of interest (ROI) overlap of participating pathologists and the consensus reference ROI by diagnostic concordance or discordance with the consensus reference diagnosis.

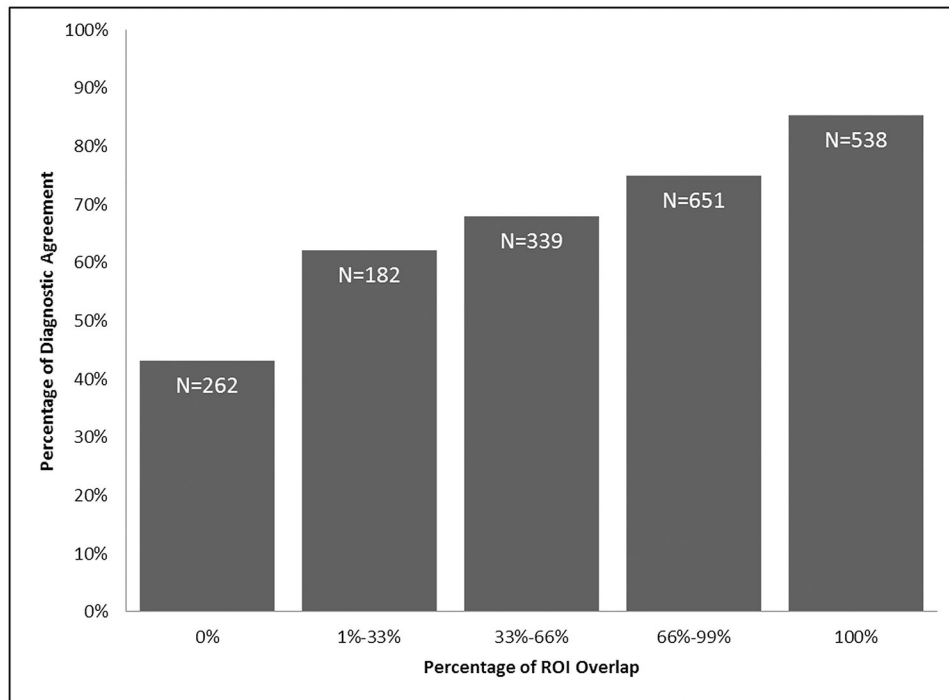


Figure 4. Average diagnostic agreement with the consensus reference diagnosis for individual assessments in five categories of percent ROI overlap with the consensus reference ROI (N=1 972 individual assessments). The numeric labels within the bars represent the number of individual assessments within each percent ROI overlap category.

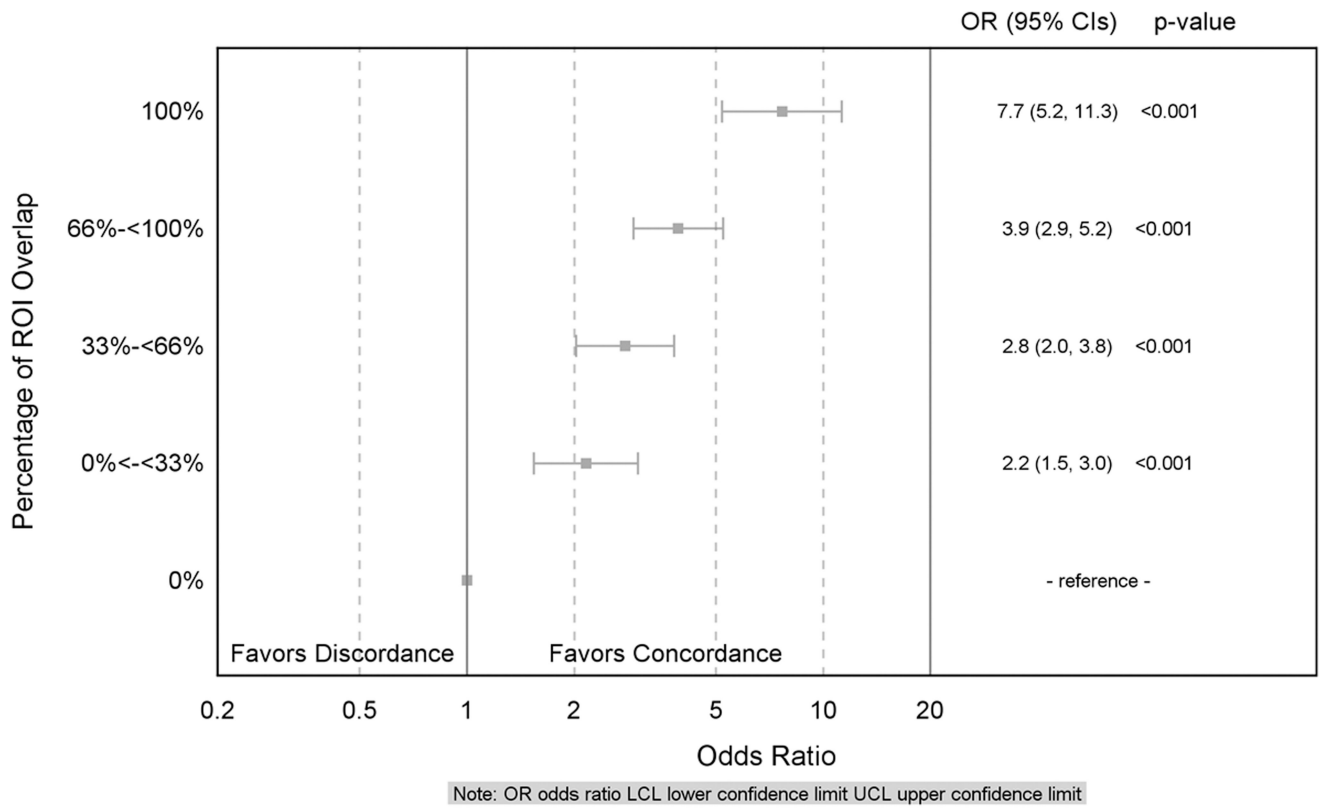


Figure 5. Unadjusted odds of diagnostic concordance by increasing ROI overlap category.

Note: OR odds ratio CI confidence interval

¹When the ROI overlap was entered as an ordinal independent variable, OR 1.6, 95% CI [1.5–1.7], p<0.001.

Table 1:

Characteristics of pathologists (N=44) and average percent overlap of the region of interest (ROI) marked by participating pathologists and the ROI marked by the expert consensus (N=1 972 total interpretations)

Pathologist Characteristics	# Pathologists n (%)	Participant % ROI Overlap compared with Expert Consensus ROI (%)	p-value
Total	44 (100.0)	66 (64 – 67)	--
Demographics			
Age at survey (years)			
30–39	6 (13.6)	65 (61 – 69)	0.92
40–49	8 (18.2)	67 (63 – 71)	
50–59	19 (43.2)	66 (63 – 68)	
60+	11 (25.0)	66 (62 – 69)	
Gender			
Male	32 (72.7)	66 (65 – 68)	0.38
Female	12 (27.3)	65 (61 – 68)	
Breast Pathology Expertise			
Facility size			
< 10 Pathologists	29 (65.9)	66 (65 – 68)	0.69
10 Pathologists	15 (34.1)	65 (62 – 69)	
Fellowship training in surgical or breast pathology			
No	19 (43.2)	67 (65 – 70)	0.067
Yes	25 (56.8)	65 (62 – 67)	
Affiliation with academic medical center			
No	33 (75.0)	65 (63 – 66)	0.002
Yes	11 (25.0)	69 (67 – 72)	
Do your colleagues consider you an expert in breast pathology?			
No	36 (81.8)	66 (64 – 68)	0.36
Yes	8 (18.2)	64 (61 – 68)	
Breast pathology experience (years)			
< 5	8 (18.2)	67 (63 – 70)	0.37
5–9	5 (11.4)	69 (65 – 74)	
10–19	17 (38.6)	65 (62 – 68)	
20	14 (31.8)	65 (63 – 68)	
No. breast cases (per week)			
< 5	12 (27.3)	65 (62 – 67)	0.63
5–9	19 (43.2)	66 (64 – 69)	
10+	13 (29.5)	66 (63 – 70)	

Table 2:

Breast biopsy case characteristics and average percent region of interest (ROI) overlap of pathologists participating in the breast pathology study and the expert consensus ROI (n=1972 independent interpretations by participating pathologists)

Patient and Case Characteristics	# Interpretations n (%)	ROI Overlap (%)	p-value
Total Interpretations	1972 (100.0)	66 (64 – 67)	--
Patient Characteristics			
Breast density			
Low density	941 (47.7)	69 (67 – 71)	<0.001
High density	1031 (52.3)	63 (61 – 65)	
Case Characteristics			
Biopsy type			
Core needle biopsy	1119 (56.7)	65 (64 – 67)	0.48
Excisional biopsy	853 (43.3)	66 (64 – 69)	
Expert Consensus Diagnosis			
Atypia	871 (44.2)	51 (48 – 54)	<0.001
DCIS	859 (43.6)	73 (70 – 76)	
Invasive	242 (12.3)	93 (90 – 95)	
Cumulative number of unique individual assessment terms given to a case by participants ¹			
< 4	393 (19.9)	85 (83 – 88)	<0.001
4–7	1161 (58.9)	64 (62 – 66)	
8	418 (21.2)	54 (50 – 58)	
Level of diagnostic difficulty of this case			
Very easy to easy	1378 (69.9)	68 (66 – 69)	0.005
Challenging to very challenging	594 (30.1)	62 (58 – 65)	
Confidence in assessment			
High confidence	1606 (81.4)	67 (65 – 68)	0.062
Low confidence	366 (18.6)	63 (59 – 66)	
Case considered borderline			
Yes	524 (26.6)	57 (54 – 61)	<0.001
No	1448 (73.4)	69 (67 – 71)	

¹Includes 14 terms: non-proliferative changes only, fibroadenoma, atypical lobular hyperplasia, lobular carcinoma in situ, intraductal papilloma without atypia, usual ductal hyperplasia, columnar cell hyperplasia/columnar cell change, sclerosing adenosis, radial scar/complex sclerosing lesion, flat epithelial atypia, atypical ductal hyperplasia, intraductal papilloma with atypia, ductal carcinoma in situ (DCIS), & invasive carcinoma