Research Article

A Randomized Study Comparing Digital Imaging to Traditional Glass Slide Microscopy for Breast Biopsy and Cancer Diagnosis

Joann G. Elmore[1], Gary M. Longton[2], Margaret S. Pepe[2,3], Patricia A. Carney[4], Heidi D. Nelson[5,6], Kimberly H. Allison[7], Berta M. Geller[8], Tracy Onega[9], Anna N. A. Tosteson[10], Ezgi Mercan[11], Linda G. Shapiro[11], Tad T. Brunyé[12], Thomas R. Morgan[1], Donald L. Weaver [13]

[1]Department of Medicine, University of Washington School of Medicine, Seattle, WA 98104, [2]Fred Hutchinson Cancer Research Center, Seattle, WA 98109, [3]Department of Biostatistics, University of Washington School of Public Health, Seattle, WA 98104, [4]Department of Family Medicine, Oregon Health and Science University, Portland, OR 97239, [5]Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR 97239, [6]Providence Cancer Center, Providence Health and Services Oregon, Portland, OR 97213, [7]Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305, [8]Department of Family Medicine, University of Vermont, Burlington, VT 05405, [9]Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756, [10]The Dartmouth Institute for Health Policy and Clinical Practice, Norris Cotton Cancer Center, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756, [11]Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195, [12]Department of Psychology, Tufts University, Medford, MA 02155, [13]Department of Pathology, UVM Cancer Center, University of Vermont, Burlington, VT 05405, USA

Received: 27 October 2016 Accepted: 18 January 2017 Published: 10 March 2017

Abstract

Background: Digital whole slide imaging may be useful for obtaining second opinions and is used in many countries. However, the U.S. Food and Drug Administration requires verification studies. **Methods:** Pathologists were randomized to interpret one of four sets of breast biopsy cases during two phases, separated by ≥9 months, using glass slides or digital format (sixty cases per set, one slide per case, $n = 240$ cases). Accuracy was assessed by comparing interpretations to a consensus reference standard. Intraobserver reproducibility was assessed by comparing the agreement of interpretations on the same cases between two phases. Estimated probabilities of confirmation by a reference panel (i.e., predictive values) were obtained by incorporating data on the population prevalence of diagnoses. **Results:** Sixty-five percent of responding pathologists were eligible, and 252 consented to randomization; 208 completed Phase I (115 glass, 93 digital); and 172 completed Phase II (86 glass, 86 digital). Accuracy was slightly higher using glass compared to digital format and varied by category: invasive carcinoma, 96% versus 93% ($P = 0.04$); ductal carcinoma *in situ* (DCIS), 84% versus 79% ($P < 0.01$); atypia, 48% versus 43% ($P = 0.08$); and benign without atypia, 87% versus 82% ($P < 0.01$). There was a small decrease in intraobserver agreement when the format changed compared to when glass slides were used in both phases ($P = 0.08$). Predictive values for confirmation by a reference panel using glass versus digital were: invasive carcinoma, 98% and 97% (not significant [NS]); DCIS, 70% and 57% ($P = 0.007$); atypia, 38% and 28% ($P = 0.002$); and benign without atypia, 97% and 96% (NS). **Conclusions:** In this large randomized study, digital format interpretations were similar to glass slide interpretations of benign and invasive cancer cases. However, cases in the middle of the spectrum, where more inherent variability exists, may be more problematic in digital format. Future studies evaluating the effect these findings exert on clinical practice and patient outcomes are required.

Keywords: Breast cancer, diagnostic accuracy, digital whole-slide imaging, intraobserver reproducibility

Introduction

Cancer diagnoses rely on a pathological interpretation of biopsy tissue using traditional glass slide microscopy. The process frequently involves obtaining second opinions before initiating treatment. Numerous prior studies have shown that more than 10% of breast biopsy diagnoses are changed after obtaining a second review.[1-6] Digital whole-slide imaging (WSI) has the potential to transform the diagnostic process by creating high-resolution digital images of glass slides that are easily transported electronically and viewable on a computer monitor with pan and zoom features, which emulates screening a glass slide at varied magnification. The digital format has replaced

Access this article online

Quick Response Code:

Website:
www.jpathinformatics.org

DOI:
10.4103/2153-3539.201920

Address for correspondence: Dr. Joann G. Elmore,
University of Washington, 325 Ninth Avenue, Box 359780,Seattle,
WA 98104-2499.
E-mail: jelmore@uw.edu

How to cite this article: Elmore JG, Longton GM, Pepe MS, Carney PA, Nelson HD, Allison KH, *et al*. A randomized study comparing digital imaging to traditional glass slide microscopy for breast biopsy and cancer diagnosis. J Pathol Inform 2017;8:12.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2017/8/1/12/201920

the microscope in many medical schools, clinical conferences, and medical board tests[7-9] and is diffusing into clinical practices for telemedicine and archiving, including rapid retrieval.[10] Telepathology using digital WSI could accelerate pathology consultations and aid the field of oncology.

While the digital format is increasingly used internationally in Europe and Canada,[11-17] it is not approved by the Food and Drug Administration (FDA) for primary diagnostic interpretation in the U.S.[11] Although several studies report promising outcomes using digital WSI, often fewer than 12 pathologists participated in these studies, or participating pathologists were experts in their clinical field, and the spectrum of cases was often limited to just a few diagnostic categories or prototypical cases.[14-22] More robust studies will be required by the FDA to sufficiently validate digital WSI technology.

The digital format may be particularly useful for breast specimens given the high volume of biopsies[23] and challenges associated with interpreting breast pathology.[24]

In this prospective randomized study, we evaluate the results of 208 practicing U.S. pathologists randomly assigned to interpret breast biopsy specimens in either traditional glass slide or digital WSI format. We also evaluate the potential for improvement with experience using the digital format during their test set interpretation, and we calculate the predictive value of cases interpreted using digital WSI by estimating the likelihood of diagnostic confirmation by a reference consensus panel.

Methods

Institutional review boards

The Institutional Review Boards at Fred Hutchinson Cancer Research Center (#9249), the University of Vermont (#M13-269), and the University of Washington (#43717) approved all study activities. Pathologists provided informed consent. All activities were HIPAA compliant.

Test case development

Test set case development and study design are previously described.[24-27] Briefly, 240 breast biopsy specimens were randomly selected from pathology registries. Each case included standardized data on the woman's age at biopsy, breast density, and biopsy type. We oversampled cases with atypia (atypical ductal hyperplasia [ADH] and ADH in a papilloma) and ductal carcinoma *in situ* (DCIS), biopsies from women aged 40–49 years, and cases from women with dense breasts. Nearly half of the 240 cases were from women aged 40–49 years (n = 118); the remainder were from women aged 50–59 years (n = 67), 60–69 years (n = 29), and >70 years (n = 26). Breast Imaging-Reporting and Data System breast density categories assessed on the previous mammography included almost entirely fat (n = 13), scattered fibroglandular densities (n = 105), heterogeneously dense (n = 97), and extremely dense (n = 25).[28] Cases were from both core needle (n = 138) and excisional (n = 102) biopsies. The 240 cases were randomly assigned to one of four test sets, with stratification to achieve balance for these factors.

Each glass slide was scanned using an iScan Coreo Au® digital slide scanner in 40× high-resolution mode. A technician and an experienced breast pathologist reviewed each digital image, rescanning as needed to obtain the highest quality. A custom online digital slide viewer was built using HD View SL, Microsoft's open source Silverlight gigapixel image viewer. The viewer, like popular online mapping applications and industry-sponsored WSI viewers, allowed pathologists to pan the image and zoom (up to 40× actual scanned magnification with additional digital magnification for a final maximum magnification of 60×). Additional tools were available for measuring lesion size and counting mitotic figures.

Determination of reference standard

Three experienced breast pathologists developed a reference interpretation by consensus agreement for each case in glass format using standardized diagnostic categories.[24] The case distribution, defined by glass slide reference categories, was: benign without atypia (30%), atypia (30%), DCIS (30%), and invasive carcinoma (10%). We present all data in comparison to the glass slide reference diagnoses. Reference panel members independently interpreted all cases again in digital format approximately 19 months after glass slide interpretation and established a digital format reference diagnosis.

Pathologist recruitment, selection, and baseline data collection

The study pathologists were recruited from eight U.S. states (AK, ME, MN, NH, NM, OR, VT, and WA), had completed residency training, had interpreted breast specimens for ≥1 year, and intended to continue interpreting breast specimens for ≥1 year. Pathologists were invited to participate through E-mail(s), subsequent mail invitations, and telephone calls. After enrolling, pathologists completed a demographic and practice characteristic survey.

Test case interpretations

Pathologists were randomly assigned to a test set and interpretive format (glass slide vs. digital) for Phase I, stratified by clinical expertise (defined by self-reported expertise in breast pathology and/or completion of a breast pathology fellowship). All interpretations were performed by pathologists using their own microscopes and computers. After at least 9 months, the pathologists were invited to interpret cases in Phase II. The pathologists were again randomly assigned to interpretive format in Phase II, with stratification based on Phase I format and clinical expertise [Figure 1 and Appendix 1].

The pathologists interpreted the same cases in both phases; however, the cases were randomly ordered for each participant and also for each phase. Pathologists were not informed that the cases in Phase II were the same exact, reordered cases they had already interpreted in Phase I. Pathologists used a web-based form to document interpretations and indicate whether they

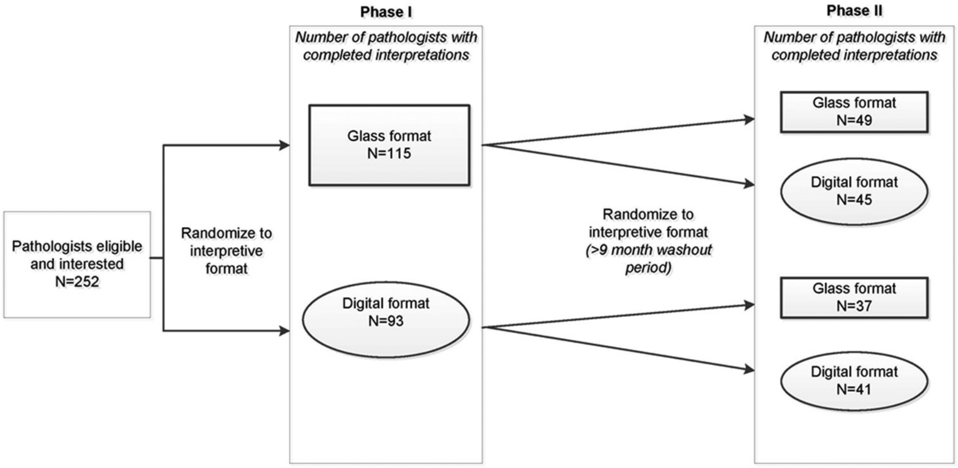

Phase I
Number of pathologists with completed interpretations
Glass format N=115
Digital format N=93
Pathologists eligible and interested N=252
Randomize to interpretive format
Randomize to interpretive format (>9 month washout period)
Phase II
Number of pathologists with completed interpretations
Glass format N=49
Digital format N=45
Glass format N=37
Digital format N=41

Figure 1: Flow diagram for pathologist randomization [see Appendix 1 for further details on recruitment and randomization]

desired a second opinion for each case.[24,27] Pathologists received up to 20 hours of Category 1 Continuing Medical Education (CME) credits after participating.

Statistical analyses

We calculated case agreement rates for Phase I with the reference diagnoses as a measure of accuracy for glass and digital format. *A priori*, we planned to use Phase I data only when comparing accuracy to avoid assumptions about carryover effects from Phase I to Phase II and because we had sufficient statistical power from Phase I data. Tests for agreement rates and confidence intervals (CIs) accounted for both within- and between-participant variability by employing variance estimates of the form (var [ratep] + [avg (ratep) × (1 – avg (ratep))]/nc)/np, where avg (ratep) is the average rate among pathologists, var (ratep) is the sample variance among pathologists, nc is the number of cases interpreted by each pathologist, and np is the number of pathologists. Effects of pathologist characteristics (e.g., expertise, digital experience) and case characteristics (e.g., patient age, biopsy type) on accuracy were examined. Results of the 6-point Likert scales for confidence and difficulty ratings were simplified to a binary variable of 1, 2, 3 versus 4, 5, 6.

When rate comparisons involved more than one factor or more than two levels for a single factor, we used logistic regression models of agreement rates with a robust variance estimator to account for the lack of independence between interpretations by the same pathologist.

We used logistic regression to examine if the effect on accuracy of glass versus digital format remained after adjusting for pathologist characteristics. Adjusting for case-level characteristics was unnecessary, as pathologists interpreted the same cases, eliminating the potential for case-level characteristics to confound the glass versus digital comparison.

We evaluated whether a learning curve existed as pathologists became more experienced using the digital format during this study. In this analysis, the average pathologist-level accuracy was estimated separately for each of the six consecutive subsets of ten cases in a pathologist's sequence of cases. We used logistic regression with an ordered covariate with values one to six indicating interpretive sequence (i.e., group of ten cases) to determine if there was an increasing trend.

To assess reproducibility, pathologists' interpretations in Phase II were compared with their interpretations of the same cases in Phase I. Agreement rates and CIs were based on logit models utilizing a robust estimator of the variance to account for correlation of case interpretations from the same pathologist. Differences in reproducibility (agreement rates) were calculated when using glass slides in both phases, when using digital format in both phases, and when the format changed between phases (e.g., using glass slides in one phase and digital in the other). Hypothesis tests were based on Wald tests of logit model coefficients distinguishing between interpretations made on different combinations of diagnostic formats.

We calculated the probability that an initial biopsy interpretation in clinical practice using the digital format would be confirmed by the reference diagnosis (i.e., the predictive value). We used previously described techniques[29] combining the Phase I data with the prevalence of diagnostic outcomes in U.S. women 50–59 years old who received breast biopsies after screening.[30]

Results

Characteristics of participating pathologists

Of responding pathologists, 252 (65%) were eligible and agreed to participate [Figure 1 and Appendix 1]. Between participating pathologists and those who declined or whom we were unable to contact, there were no statistically significant differences in mean pathologist age, sex, or the proportion working in a population of 250,000 or more.[24] Table 1 shows the characteristics and clinical experience of the 208 pathologists completing Phase I. Approximately half (48%) reported using the digital format in their professional work,

Table 1: Characteristics of the 208 participating pathologists shown aggregated and by Phase I random assignment to traditional glass or digital whole slide imaging interpretation

Characteristics	Pathologists, *n* (%)		
	Total	Phase I randomization[a]	
		Glass format	Digital format
Total	208 (100.0)	115 (55.3)	93 (44.7)
Demographics			
Age at survey (years)			
30-39	28 (13.5)	16 (13.9)	12 (12.9)
40-49	70 (33.7)	41 (35.7)	29 (31.2)
50-59	74 (35.6)	42 (36.5)	32 (34.4)
60+	36 (17.3)	16 (13.9)	20 (21.5)
Sex			
Male	132 (63.5)	69 (60.0)	63 (67.7)
Female	76 (36.5)	46 (40.0)	30 (32.3)
Clinical practice and breast pathology expertise			
Laboratory group practice size			
<10 pathologists	134 (64.4)	68 (59.1)	66 (71.0)
≥10 pathologists	74 (35.6)	47 (40.9)	27 (29.0)
Fellowship training in breast pathology or surgical pathology			
No	105 (50.5)	56 (48.7)	49 (52.7)
Yes	103 (49.5)	59 (51.3)	44 (47.3)
Affiliation with academic medical center			
No	153 (73.6)	87 (75.7)	66 (71.0)
Yes, adjunct/affiliated	35 (16.8)	17 (14.8)	18 (19.4)
Yes, primary appointment	20 (9.6)	11 (9.6)	9 (9.7)
Do your colleagues consider you an expert in breast pathology?			
No	164 (78.8)	90 (78.3)	74 (79.6)
Yes	44 (21.2)	25 (21.7)	19 (20.4)
Breast pathology experience (years)			
<5	39 (18.8)	22 (19.1)	17 (18.3)
5-9	34 (16.3)	23 (20.0)	11 (11.8)
10-19	74 (35.6)	34 (29.6)	40 (43.0)
≥20	61 (29.3)	36 (31.3)	25 (26.9)
Breast specimen case load (% of total clinical work)			
<10	104 (50.0)	59 (51.3)	45 (48.4)
10-24	87 (41.8)	45 (39.1)	42 (45.2)
25-49	13 (6.3)	8 (7.0)	5 (5.4)
≥50	4 (1.9)	3 (2.6)	1 (1.1)
Number of breast cases (per week)			
<5	47 (22.6)	31 (27.0)	16 (17.2)
5-9	91 (43.8)	44 (38.3)	47 (50.5)
10-19	53 (25.5)	31 (27.0)	22 (23.7)
20-29	9 (4.3)	4 (3.5)	5 (5.4)
≥30	8 (3.8)	5 (4.3)	3 (3.2)
Do you have any experience using digitized whole slides in your professional work?[b]			
No	109 (52.4)	63 (54.8)	46 (49.5)
Yes	99 (47.6)	52 (45.2)	47 (50.5)
Impressions about breast pathology			
How confident are you interpreting breast pathology?			
1 very confident	31 (14.9)	14 (12.2)	17 (18.3)
2	113 (54.3)	66 (57.4)	47 (50.5)
3	49 (23.6)	27 (23.5)	22 (23.7)
4	12 (5.8)	8 (7.0)	4 (4.3)
5	3 (1.4)	0 (0.0)	3 (3.2)
6 not confident at all	0 (0.0)	0 (0.0)	0 (0.0)

Contd...

Table 1: Contd...

Characteristics	Pathologists, *n* (%)		
	Total	Phase I randomization[a]	
		Glass format	Digital format
How challenging is breast pathology?			
1 very easy	2 (1.0)	1 (0.9)	1 (1.1)
2	21 (10.1)	13 (11.3)	8 (8.6)
3	71 (34.1)	43 (37.4)	28 (30.1)
4	85 (40.9)	44 (38.3)	41 (44.1)
5	27 (13.0)	14 (12.2)	13 (14.0)
6 very challenging	2 (1.0)	0 (0.0)	2 (2.2)
Breast pathology makes me more nervous than other types of pathology			
1 strongly disagree	24 (11.5)	13 (11.3)	11 (11.8)
2	64 (30.8)	35 (30.4)	29 (31.2)
3	28 (13.5)	16 (13.9)	12 (12.9)
4	51 (24.5)	28 (24.3)	23 (24.7)
5	36 (17.3)	20 (17.4)	16 (17.2)
6 strongly agree	5 (2.4)	3 (2.6)	2 (2.2)

[a]No statistically significant differences were noted in any of the characteristics between pathologists randomized to glass format versus digital format. The *P* values correspond to a Pearson Chi-square test for a difference in pathologist factor distribution between those reading glass and digital formats where there were two or three categories per factor. A *t*-test for continuous pathologist age was used. A Wilcoxon rank-sum test was used for all other factors with four or more ordered categories, [b]Pathologists were asked, "In what ways do you use digitized whole slides in your professional work?" Pathologists were deemed to have experience in digital pathology if they reported any answer other than "not at all." The full list of possible answers included: Primary pathology diagnosis, tumor board/clinical conference, consultative diagnosis, CME/board exams/teaching in general, archival purposes, research, other (text box provided), not at all. CME: Continuing Medical Education

mostly during conferences and teaching. While most (93%) pathologists reported confidence when interpreting breast pathology, 55% reported that breast pathology is challenging, and 44% reported that breast pathology makes them more nervous than other pathology types.

Pathologists' confidence by interpretive format

Phase I results include 6,900 interpretations in glass slide format and 5,580 in digital format. When comparing glass slide versus digital format, pathologists reported similar rates of confidence (81.7% vs. 78.6%, $P = 0.22$) and percentage of interpretations marked as borderline between two diagnoses (26.1% vs. 24.6%, $P = 0.35$). However, glass slide interpretations were less likely than digital interpretations to be rated as challenging cases (30.0% vs. 38.5%, $P = 0.003$), and pathologists were less likely to desire a second opinion on glass than on digital interpretations (35.5% vs. 42.5%, $P = 0.03$).

Accuracy by format

Pathologists' accuracy within each diagnostic category was 3–5% higher for pathologists interpreting glass slides compared to those assigned to digital format: benign without atypia (glass: 87%, digital: 82%; $P < 0.01$); atypia (glass: 48%, digital: 43%; $P = 0.08$); DCIS (glass: 84%, digital: 79%; $P < 0.01$); and invasive carcinoma (glass: 96%, digital: 93%; $P = 0.04$) [Table 2 and Figure 2]. Similar trends occurred when compared to the reference standard established by experts using the digital format, though the differences were slightly smaller, ranging from 2% to 3% [Appendix 2].

The pathologist and case characteristics associated with accuracy using digital format (and lack thereof) were consistent with those previously observed in the interpretation of glass format [Appendix 3]. For example, pathologists reporting higher breast interpretation case volume had higher accuracy in both interpretive formats, and accuracy was not influenced by patient age or breast biopsy type. Biopsy interpretations from women with dense breast tissue on prior mammography also had lower accuracy in the digital format compared to low-density breast tissue, similar to findings in traditional glass.

Reproducibility (intraobserver agreement between Phase I and Phase II)

Pathologists (n = 172) who completed interpretations in both phases on the same cases provided a total of 20,640 individual case assessments. Intraobserver agreement between interpretations of the same case (Phase I vs. Phase II) by diagnostic category and interpretive format is shown in Figure 3 and Appendix 4. The overall intraobserver agreement was highest when glass format was used in both phases at 79% (95%CI: 77%–81%). When the interpretive format changed between phases, the intraobserver agreement was slightly lower at 77% (95%CI: 75%–78%) but not statistically significantly different from the findings noted when the glass format was used in both phases ($P = 0.08$). A statistically significant difference, however, was noted when the glass format was used in both phases versus when the digital format was used in both phases, where the overall intraobserver agreement was 73% (95%CI: 71%–76%; $P < 0.001$). While pathologists' reproducibility was high for cases of invasive breast carcinoma, regardless of which format was used in the two phases or whether the format changed (93%–97%), it was low for cases in the middle categories such as atypia (56%–62%), regardless of interpretive format.

Evaluation for a learning curve among pathologists in the digital format

No learning curve was observed over the sixty cases interpreted digitally in Phase I ($P = 0.85$). There was also no difference in the accuracy between Phase II and Phase I among pathologists randomized to the digital format in both phases ($P = 0.90$). This was also true for pathologists randomized to the glass slide format in both phases ($P = 0.35$).

Predictive values of digital format compared with glass slide interpretations

The estimated numbers of cases under- and over-interpreted in the U.S. (i.e., that would be reclassified to a different diagnostic category by the reference consensus panel review) is shown in Figure 4 by interpretive format and diagnostic category of the initial interpretation [Appendix 5]. The predictive values for cases initially interpreted as invasive breast carcinoma are

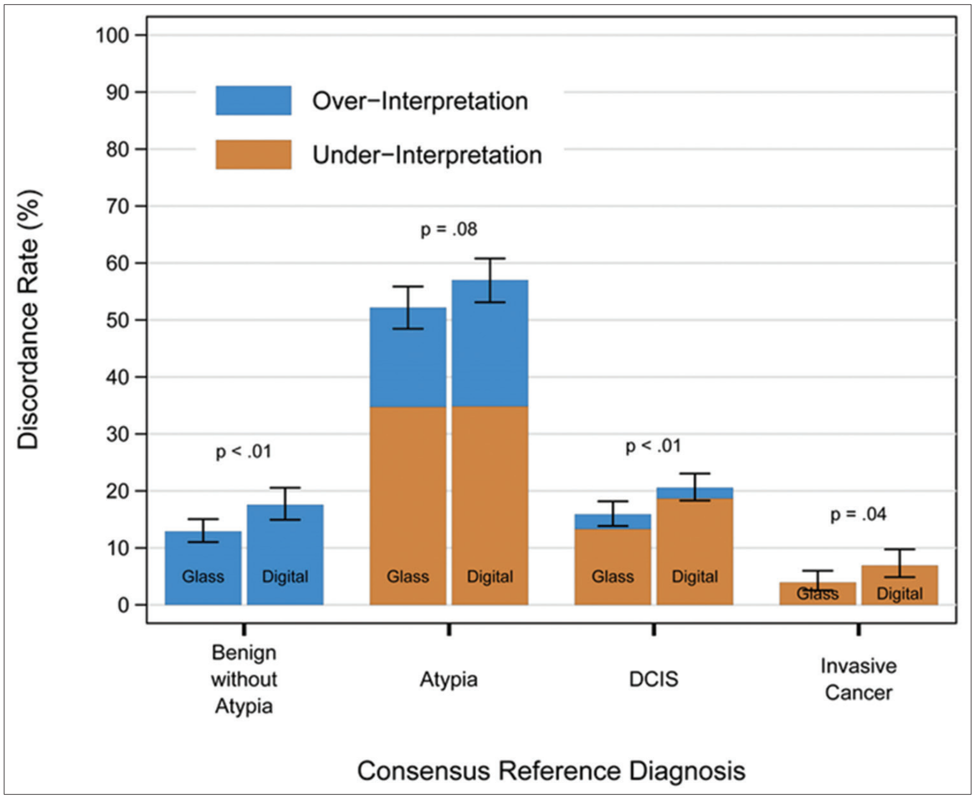


Figure 2: Percent of Phase I under- and over-interpretations compared with the consensus reference diagnosis by pathologist interpretive format (glass slide or digital whole-slide imaging format)

Table 2: Pathologists' accuracy by interpretive format (Phase I interpretations compared with the consensus panel reference interpretations)[a]

Consensus reference interpretation	Benign without atypia	Atypia	DCIS	Invasive	Total number of interpretations	Percentage agreement of pathologists with consensus reference (95% CI)
Glass pathologists; interpretation on glass slides (6900 interpretations)						
Benign without atypia	1803	200	46	21	2070	87 (85-89)
Atypia	719	990	353	8	2070	48 (44-52)
DCIS	133	146	1764	54	2097	84 (82-86)
Invasive breast cancer	3	0	23	637	663	96 (94-97)
Digital pathologists' interpretation on digital WSI Images (5580 interpretations)						
Benign without atypia	1380	216	62	16	1674	82 (79-85)
Atypia	583	720	356	15	1674	43 (39-47)
DCIS	170	147	1348	32	1697	79 (77-82)
Invasive breast cancer	14	1	22	498	535	93 (90-95)

[a] Expert consensus reference diagnosis obtained using the glass slide format. DCIS: Ductal carcinoma *in situ*, WSI: Whole-slide imaging

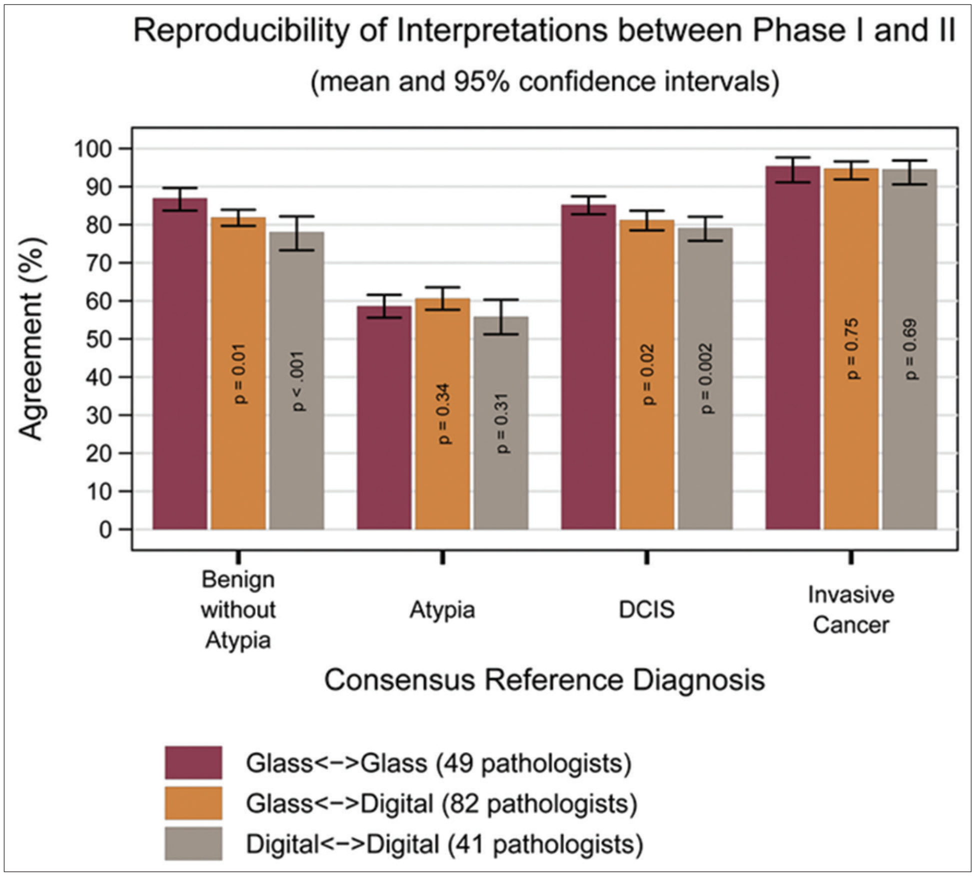


Figure 3: Reproducibility of interpretations: Intraobserver agreement of participants' interpretations of the same case in Phase I and Phase II by diagnostic format used by the participant for interpretation in both phases. Data shown by the reference diagnosis of the case (*n* = 172 pathologists with a total of 20,640 individual case assessments) *P*-values correspond to comparrisons with intraobserver aggrement of pathologists who read glass slides in both phases

similar regardless of interpretive format. For example, a slide interpreted digitally as invasive carcinoma was 97.2% (95%CI: 95.6%–98.6%) likely to be confirmed as invasive carcinoma by our expert reference panel using the original glass slide. This is comparable to the previously reported predictive value when the initial interpretation was obtained by glass slide of 97.7% (95%CI: 96.5%–98.7%).[29] Similarly, interpretations of benign without atypia were highly likely to be confirmed by the reference panel regardless of format (95.7% digital vs. 97.1% glass).

Of note, the estimated predictive values were significantly lower for atypia and DCIS in the digital interpretation format compared with glass interpretations (Wald test: atypia *P* = 0.002; DCIS *P* = 0.007). While these predictive values were statistically significantly lower for interpretations obtained in the digital format, the predictive values of these challenging cases as previously reported are also low in the glass format.[29] For example, the predictive values for an initial atypia interpretation in the U.S. being in agreement with a reference review were 27.8% in the digital format versus 37.8% glass format, and for DCIS cases, the values were 57.1% digital versus 69.6% glass.

Interpretation time

Pathologists using the digital format spent more time interpreting than pathologists using glass slides, as measured by total requested CME hours. The percentage of pathologists who reported spending 20 hours participating in the study (the maximum allowed) was higher among those interpreting in the digital format in both phases versus those interpreting in the glass format in both phases (76% digital versus 51% glass, respectively; *P* = 0.01; Wilcoxon rank-sum test for difference).

Conclusions

To date, our study of 240 biopsy cases interpreted by >200 pathologists from across the U.S. is the largest randomized study comparing traditional glass microscopy and digital WSI. Our study highlights the many challenges we face as we move into the digital era in the design and analyses of quality assessment studies. In our study, predictive value estimates were nearly identical regardless of interpretive format at the extremes of the diagnostic spectrum (e.g., invasive cancer and benign tissue), suggesting digital WSI could be employed for the primary diagnosis for these extreme categories. However, the more challenging (and less common) atypia and DCIS

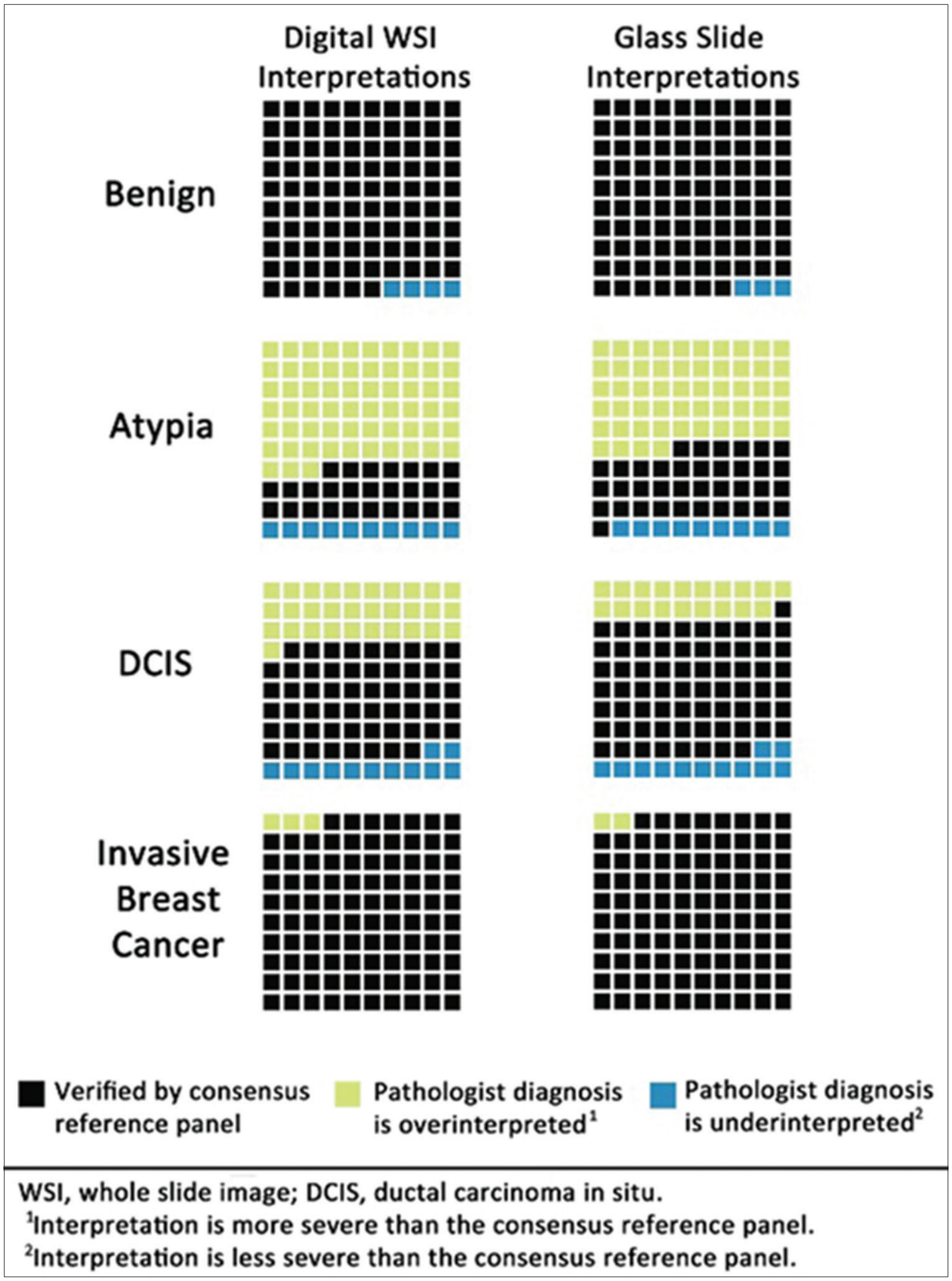


Figure 4: Estimated numbers of breast biopsy cases that are under- and over-interpreted in the U.S. Results are shown for the number of cases that would be reclassified to a more (blue) or less (green) severe diagnostic category by the reference consensus panel diagnosis. Results pertain to women aged 50–59 years with recent screening mammograms in the U.S. and assume their biopsies were interpreted by pathologists using either a glass slide or a digitized image (one slide per case and without second opinions)

diagnostic categories in the middle of the spectrum have lower reproducibility and accuracy in the digital interpretive format. It should be noted that reproducibility and accuracy are also lower for atypia and DCIS when using glass slides, but the effect is amplified using digital WSI. As the field of digital pathology moves forward, attention to inclusion of the full spectrum of cases in validation studies will be important.

While our study followed the digital imaging validation guidelines recommended by the College of American Pathologists,[31] our design also exceeded their recommendations in a few notable ways. Our study design included randomly allocating pathologists to interpretive format, using a random selection process for identifying cases, including a Phase I glass to Phase II glass reproducibility study arm as a benchmark, and employing a 9-month wash-out period between phases to reduce recall bias when assessing reproducibility. We also compared pathologists' accuracy using a carefully defined expert consensus reference standard. Finally, the investigators have no associations with manufacturers of digital WSI instruments or viewing platforms except that one commercial manufacturer provided use of a scanner to digitally archive the glass slides.

It is possible that the slightly lower accuracy with digital WSI imaging that we noted can be corrected with experience. Among pathologists reporting prior experience using digital WSI, we noted a nonsignificant trend for higher accuracy of digital interpretations than for pathologists who reported no experience with WSI, even after accounting for the effects of other pathologist-level characteristics. However, participating pathologists had limited experience with the digital format as it is not currently approved for primary diagnostic use in the U.S. by the FDA. It may be too early to address whether experience with the digital format results in improved diagnostic accuracy.

In the digital format, pathologists were more likely to deem a case challenging and spent more time interpreting cases compared to pathologists using glass slides – circumstantial evidence suggesting that experience with the technology may be an issue. Technological improvements to image acquisition and standardized display systems, coupled with physician education and experience using digital WSI, may reduce performance gaps between the formats. While no learning curve was noted in performance during this study, gaining experience requires time, and sixty cases without an educational intervention may be inadequate. The absence of a learning curve has been noted by others,[22] though an improvement in accuracy after completing an educational intervention was reported in one study.[32]

Many areas of pathology are challenging and might benefit from digital technology. Pathologists are understandably concerned about the high level of difficulty of breast pathology[24] and the high risk for medical malpractice when a cancer diagnosis is a possibility.[33] Pathologists are also likely to desire a second opinion to improve clinical care on breast cases more often than being required by existing laboratory policies.[34] Digital technology could, therefore, be an important tool to facilitate second opinions on these challenging cases.

Pathologists interpret differently using traditional glass slide microscopy versus digital WSI format. Behind the microscope, small finger movements reposition the slide, and eye saccades scan the microscopic field; the remainder of the head and body are stationary. Digital viewing requires larger hand movements to pan and zoom and greater head and eye movements to scan all areas of the image. In addition, for pathologists wearing corrective lenses, particularly bifocals or variable focus lenses, constant corrections are needed to maintain focus. Implementation studies in Sweden suggest job-specific ergonomics may be improved by incorporating the digital format.[35]

Special considerations in designing quality assessment studies

Technologic improvements in design and image quality are occurring quickly in this field. Going forward, proposed technical performance parameters and regulation of digital imaging have been outlined by the FDA and discussed by others.[11,36,37] One potential limitation to this study is that each pathologist completed the histology evaluation remotely using their own microscope and computer, with no standardization. We do not have information on their workstation and monitor specifications or internet and bandwidth capabilities. The scanner we used is no longer commercially available, and scanner technology is rapidly updating. However, the digital whole slide images were acquired using a 40× objective lens and research staff carefully reviewed the digital scan image of each slide to avoid errors introduced during digital scanning and to assure quality.

In a randomized study design such as ours, other limiting factors apply equally to both glass and digital formats. For example, our study included one slide per case, assessment of performance in a testing situation instead of actual clinical setting, and a higher proportion of benign proliferative, atypia, and DCIS cases than usual clinical practice, as well as a relatively small number of invasive cancer cases. While these can be considered limitations, these limiting factors were equally present in the digital and the glass format testing.

Implications

Digital imaging technology has revolutionized medicine and is an important emerging adjunct to traditional light microscopy that might greatly aid the practice of pathology. We noted that diagnoses of invasive breast carcinoma are highly reproducible using both glass and digital formats. However, clinical practice includes a broad spectrum of cases, including those in the middle diagnostic categories, and these cases are often more challenging to diagnose even in the traditional glass slide format. As noted in this study, the more challenging high-risk and preinvasive lesions (atypia and DCIS) may have lower predictive value using a digital format compared with a glass slide format. We encourage future studies evaluating the effect(s) of the digital format on patient outcomes to include the full spectrum of cases and consider the randomized design features presented in our study.

Acknowledgments

The collection of cancer and vital status data was supported in part by several state public health departments and cancer registries throughout the U.S. For a full description of these sources, please see: http://www.breastscreening.cancer.gov/work/acknowledgement.html. The authors thank Ventana Medical Systems, Inc. (Roche Group) for iScan Coreo Au™ whole slide imaging system use. HD View SL was used as the source code to build our digital viewer; see http://hdviewsl.codeplex.com/for details. The authors also thank the participating pathologists.

Financial support and sponsorship

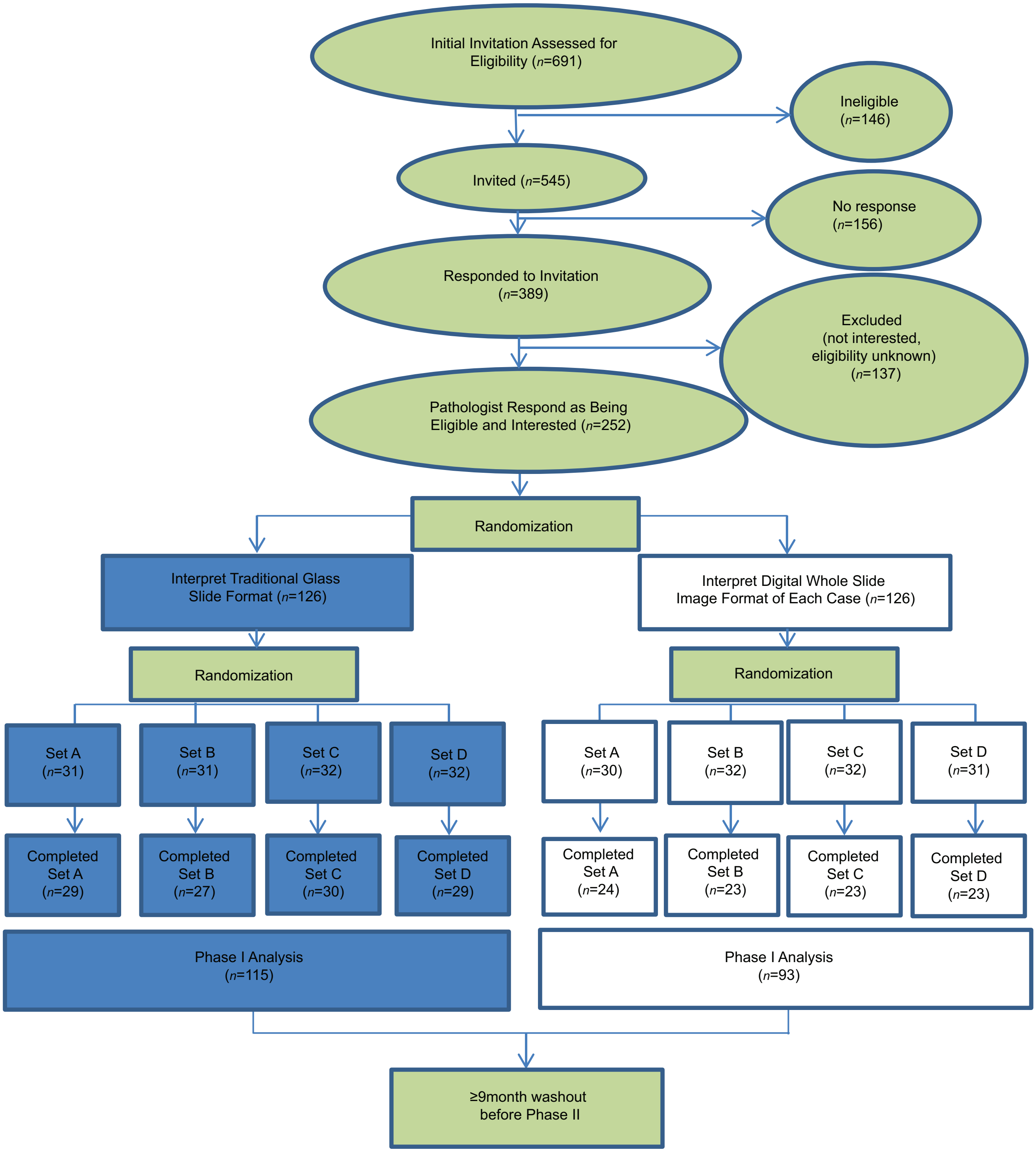
This study was supported by the National Cancer Institute (R01CA140560, R01CA172343, K05CA104699, U01CA86082, U01CA70013) and the NCI-funded Breast Cancer Surveillance Consortium (HHSN261201100031C).

Conflicts of interest

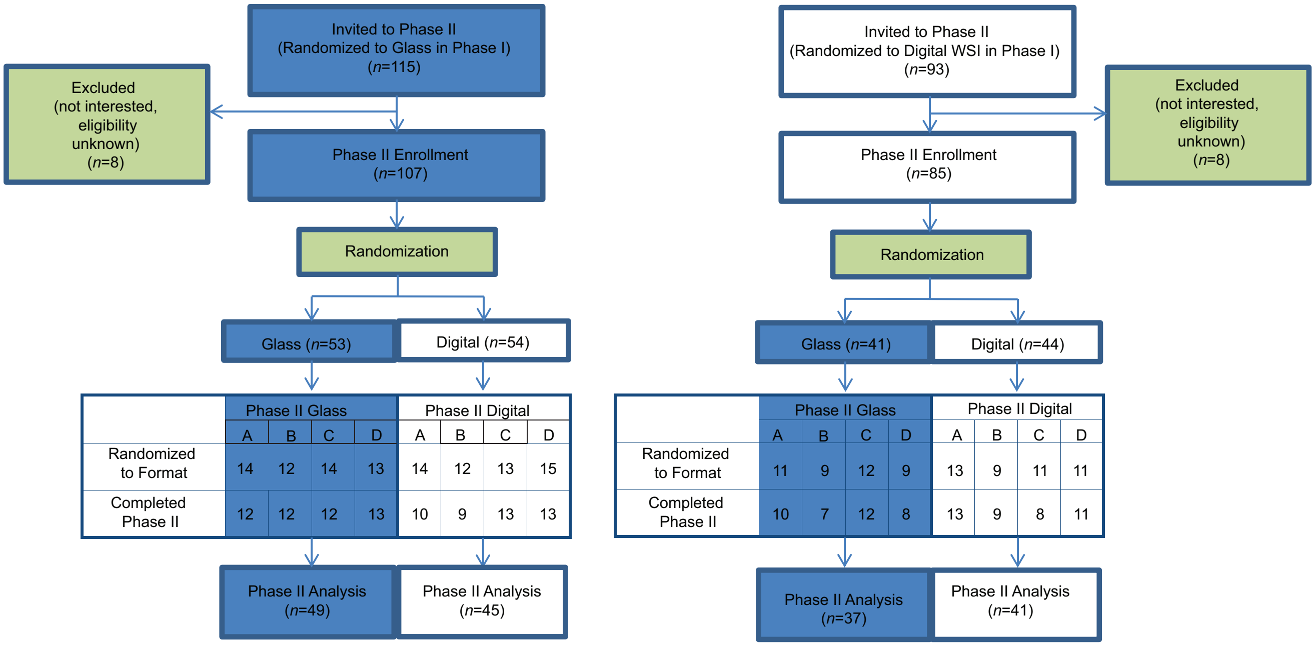
There are no conflicts of interest.

References

1. Kennecke HF, Speers CH, Ennis CA, Gelmon K, Olivotto IA, Hayes M. Impact of routine pathology review on treatment for node-negative breast cancer. J Clin Oncol 2012;30:2227-31.
2. Khazai L, Middleton LP, Goktepe N, Liu BT, Sahin AA. Breast pathology second review identifies clinically significant discrepancies in over 10% of patients. J Surg Oncol 2015;111:192-7.
3. Newman EA, Guest AB, Helvie MA, Roubidoux MA, Chang AE, Kleer CG, *et al.* Changes in surgical management resulting from case review at a breast cancer multidisciplinary tumor board. Cancer 2006;107:2346-51.
4. Marco V, Muntal T, García-Hernandez F, Cortes J, Gonzalez B, Rubio IT. Changes in breast cancer reports after pathology second opinion. Breast J 2014;20:295-301.
5. Romanoff AM, Cohen A, Schmidt H, Weltz CR, Jaffer SM, Nagi CS, *et al.* Breast pathology review: Does it make a difference? Ann Surg Oncol 2014;21:3504-8.
6. Staradub VL, Messenger KA, Hao N, Wiley EL, Morrow M. Changes in breast cancer therapy because of pathology second opinions. Ann Surg Oncol 2002;9:982-7.
7. Hamilton PW, Wang Y, McCullough SJ. Virtual microscopy and digital pathology in training and education. APMIS 2012;120:305-15.
8. Dee FR. Virtual microscopy in pathology education. Hum Pathol 2009;40:1112-21.
9. Pantanowitz L, Valenstein PN, Evans AJ, Kaplan KJ, Pfeifer JD, Wilbur DC, *et al.* Review of the current state of whole slide imaging in pathology. J Pathol Inform 2011;2:36.
10. Huisman A, Looijen A, van den Brink SM, van Diest PJ. Creation of a fully digital pathology slide archive by high-volume tissue slide scanning. Hum Pathol 2010;41:751-7.
11. Titus K. Regulators Scanning the Digital Scanner. CAP Today; 2012. p. 7. Available from: http://www.captodayonline.com/Archives/0112/0112a_regulators.html. [Last cited on 2016 Oct 13].
12. Allen TC. Digital pathology and federalism. Arch Pathol Lab Med 2014;138:162-5.
13. Tetu B, Evans A. Canadian licensure for the use of digital pathology for routine diagnoses: One more step toward a new era of pathology practice without borders. Arch Pathol Lab Med 2014;138:302-4.
14. Mooney E, Hood AF, Lampros J, Kempf W, Jemec GB. Comparative diagnostic accuracy in virtual dermatopathology. Skin Res Technol 2011;17:251-5.
15. Reyes C, Ikpatt OF, Nadji M, Cote RJ. Intra-observer reproducibility of whole slide imaging for the primary diagnosis of breast needle biopsies. J Pathol Inform 2014;5:5.
16. Gui D, Cortina G, Naini B, Hart S, Gerney G, Dawson D, *et al.* Diagnosis of dysplasia in upper gastro-intestinal tract biopsies through digital microscopy. J Pathol Inform 2012;3:27.
17. Ozluk Y, Blanco PL, Mengel M, Solez K, Halloran PF, Sis B. Superiority of virtual microscopy versus light microscopy in transplantation pathology. Clin Transplant 2012;26:336-44.
18. Nassar A, Cohen C, Agersborg SS, Zhou W, Lynch KA, Barker EA, *et al.* A multisite performance study comparing the reading of immunohistochemical slides on a computer monitor with conventional manual microscopy for estrogen and progesterone receptor analysis. Am J Clin Pathol 2011;135:461-7.
19. Jukic DM, Drogowski LM, Martina J, Parwani AV. Clinical examination and validation of primary diagnosis in anatomic pathology using whole slide digital images. Arch Pathol Lab Med 2011;135:372-8.
20. Jen KY, Olson JL, Brodsky S, Zhou XJ, Nadasdy T, Laszik ZG. Reliability of whole slide images as a diagnostic modality for renal allograft biopsies. Hum Pathol 2013;44:888-94.
21. House JC, Henderson-Jackson EB, Johnson JO, Lloyd MC, Dhillon J, Ahmad N, *et al.* Diagnostic digital cytopathology: Are we ready yet? J Pathol Inform 2013;4:28.
22. Buck TP, Dilorio R, Havrilla L, O'Neill DG. Validation of a whole slide imaging system for primary diagnosis in surgical pathology: A community hospital experience. J Pathol Inform 2014;5:43.
23. Silverstein MJ, Recht A, Lagios MD, Bleiweiss IJ, Blumencranz PW, 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;209:504-20.
24. Elmore JG, Longton GM, Carney PA, Geller BM, Onega T, Tosteson AN, *et al.* Diagnostic concordance among pathologists interpreting breast biopsy specimens. JAMA 2015;313:1122-32.
25. 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;13:3.
26. Feng S, Weaver DL, Carney PA, Reisch LM, Geller BM, Goodwin A, *et al.* A framework for evaluating diagnostic discordance in pathology discovered during research studies. Arch Pathol Lab Med 2014;138:955-61.
27. 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;65:240-51.
28. American College of Radiology. Breast Imaging Reporting and Data System (BI-RADS). Reston, VA: American College of Radiology; 1993.
29. Elmore JG, Nelson HD, Pepe MS, Longton GM, Tosteson AN, Geller B, *et al.* Variability in pathologists' interpretations of individual breast biopsy slides: A population perspective. Ann Intern Med 2016;164:649-55.
30. Weaver DL, Rosenberg RD, Barlow WE, Ichikawa L, Carney PA, Kerlikowske K, *et al.* Pathologic findings from the breast cancer surveillance consortium: Population-based outcomes in women undergoing biopsy after screening mammography. Cancer 2006;106:732-42.
31. Pantanowitz L, Sinard JH, Henricks WH, Fatheree LA, Carter AB, Contis L, *et al.* Validating whole slide imaging for diagnostic purposes in pathology: Guideline from the College of American Pathologists Pathology and Laboratory Quality Center. Arch Pathol Lab Med 2013;137:1710-22.
32. Jones NC, Nazarian RM, Duncan LM, Kamionek M, Lauwers GY, Tambouret RH, *et al.* Interinstitutional whole slide imaging teleconsultation service development: Assessment using internal training and clinical consultation cases. Arch Pathol Lab Med 2015;139:627-35.
33. Reisch LM, Carney PA, Oster NV, Weaver DL, Nelson HD, Frederick PD, *et al.* Medical malpractice concerns and defensive medicine: A nationwide survey of breast pathologists. Am J Clin Pathol 2015;144:916-22.
34. Geller BM, Nelson HD, Carney PA, Weaver DL, Onega T, Allison KH, *et al.* Second opinion in breast pathology: Policy, practice and perception. J Clin Pathol 2014;67:955-60.
35. Thorstenson S, Molin J, Lundström C. Implementation of large-scale routine diagnostics using whole slide imaging in Sweden: Digital pathology experiences 2006-2013. J Pathol Inform 2014;5:14.
36. Parwani AV, Hassell L, Glassy E, Pantanowitz L. Regulatory barriers surrounding the use of whole slide imaging in the United States of America. J Pathol Inform 2014;5:38.
37. Technical Performance Assessment of Digital Pathology Whole Slide Imaging Devices: Guidance for Industry and Food and Drug Administration Staff. Rockville, MD: Federal Register: U. S. Food and Drug Administration; c2016. Avaiable from: http://www.fda.gov/ucm/groups/fdagov-public/@fdagov-meddev-gen/documents/document/ucm435355.pdf. [Last cited on 2016 Oct 13].



Appendix 1a: Pathologist recruitment and randomization for Phase I



Appendix 1b: Phase II detailed flow diagram for pathologist randomization

Appendix 2: Rates of over- and under-interpretation and agreement with the reference diagnosis for glass interpretation and digital interpretation using the digital consensus reference interpretations

Consensus reference interpretation based on digital slide format[a]	Glass interpretation			Digital interpretation		
	Rate of over- and under-interpretation compared to the reference diagnosis		Agreement with reference diagnosis	Rate of over- and under-interpretation compared to the reference diagnosis		Agreement with reference diagnosis
	Over-interpretation Rate (95% CI)	Under-Interpretation Rate (95% CI)	Rate (95% CI)	Over-interpretation Rate (95% CI)	Under-interpretation Rate (95% CI)	Rate (95% CI)
Benign without atypia	18 (15-20)	-	82 (80-85)	20 (17-23)	-	80 (77-83)
Atypia	19 (16-22)	36 (32-39)	46 (42-49)	22 (19-26)	34 (30-38)	44 (39-48)
DCIS	3 (2-4)	16 (14-18)	81 (78-83)	2 (2-3)	19 (17-22)	78 (76-81)
Invasive breast cancer	-	1 (0-3)	99 (97-100)	-	4 (2-7)	96 (93-98)

DCIS: Ductal carcinoma *in situ*, CI: Confidence interval

Appendix 3: Associations between pathologist and case characteristics and rates of agreement with expert consensus reference diagnosis when 115 pathologists interpreted breast biopsy cases in glass format, and 93 participants interpreted in digital format

Pathologist characteristics (*n*=115 glass, 93 WSI)	Number of pathologists	Number of interpretations	Percentage of diagnoses over-interpreted (95% CI)	Percentage of diagnoses under-interpreted (95% CI)	Agreement rate with reference diagnosis (95% CI)	*P*[a]
Age						
Glass						
<40	16	960	9 (6-13)	18 (14-22)	74 (69-78)	0.16[b]
40-49	41	2460	10 (8-12)	13 (11-16)	77 (74-80)	
50-59	42	2520	11 (9-13)	13 (11-16)	76 (73-79)	
60+	16	960	9 (6-15)	20 (16-25)	70 (66-74)	
Digital						
<40	12	720	14 (8-23)	18 (12-26)	68 (61-74)	0.98[b]
40-49	29	1740	13 (9-16)	15 (12-19)	72 (68-76)	
50-59	32	1920	12 (9-15)	17 (14-20)	71 (68-74)	
60+	20	1200	13 (9-17)	18 (14-22)	69 (65-74)	
Academic affiliation						
Glass						
None	87	5220	11 (9-12)	15 (14-17)	74 (72-76)	0.007[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)	
Digital						0.96[c]
None	66	3960	12 (10-14)	17 (15-20)	71 (68-73)	
Adjunct affiliation	18	1080	13 (9-17)	15 (12-20)	72 (67-77)	
Primary academic	9	540	15 (8-27)	16 (10-25)	68 (61-75)	
Estimated number of breast cases interpreted per week						
Glass						
<5	31	1860	11 (8-14)	17 (15-21)	72 (68-75)	0.001[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)	
Digital						
<5	16	960	13 (9-19)	22 (17-28)	65 (60-69)	<0.001[d]
5-9	47	2820	13 (10-15)	16 (14-19)	71 (68-74)	
10-19	22	1320	13 (10-18)	16 (12-20)	71 (67-75)	
20+	8	480	9 (5-16)	12 (8-18)	79 (72-84)	
Practice size[e]						
Glass						
1-9 pathologists	68	4080	10 (8-12)	16 (14-19)	74 (71-76)	0.034
≥10 pathologists	47	2820	9 (8-12)	13 (11-15)	78 (75-80)	
Digital						
1-9 pathologists	66	3960	13 (11-15)	18 (16-20)	69 (67-72)	0.06
≥10 pathologists	27	1620	12 (9-16)	14 (12-17)	74 (70-77)	
Expertise in breast pathology[f]						
Glass						
Nonexpert	88	5280	10 (9-12)	16 (14-17)	74 (72-76)	0.055
Expert	27	1620	9 (7-12)	12 (9-16)	79 (75-82)	
Digital						
Nonexpert	74	4440	13 (11-16)	17 (15-19)	69 (67-72)	0.02
Expert	19	1140	9 (6-14)	15 (11-20)	76 (71-80)	
Experience with digital pathology						
Glass						
No	63	3780	10 (8-12)	15 (13-18)	75 (72-77)	0.61
Yes	52	3120	10 (8-12)	14 (12-17)	76 (73-78)	

Contd...

Appendix 3: Contd...

Pathologist characteristics (*n*=115 glass, 93 WSI)	Number of pathologists	Number of interpretations	Percentage of diagnoses over-interpreted (95% CI)	Percentage of diagnoses under-interpreted (95% CI)	Agreement rate with reference diagnosis (95% CI)	P^a
Digital						
No	46	2760	12 (10-15)	19 (16-22)	69 (66-72)	0.09
Yes	47	2820	13 (10-16)	15 (12-17)	72 (69-75)	
Test case patient characteristics (*n*=240 test cases)	**Number of cases**	**Number of interpretations**	**Percentage of diagnoses overinterpreted (95% CI)**	**Percentage of diagnoses underinterpreted (95% CI)**	**Agreement rate with reference diagnosis (95% CI)**	P^a
Patient age at time of biopsy (years)						
Glass						
40-49	118	3391	11 (9-13)	14 (12-16)	76 (73-78)	0.45
≥50	122	3509	9 (8-11)	16 (14-18)	75 (73-77)	
Digital						
40-49	118	2744	13 (11-16)	16 (14-19)	71 (68-73)	0.81
≥50	122	2836	12 (10-14)	17 (15-20)	71 (68-73)	
Breast density						
Glass						
Low	118	3391	8 (7-10)	14 (12-16)	77 (75-80)	<0.001
High	122	3509	11 (10-13)	16 (14-18)	73 (71-75)	
Digital						
Low	118	2744	12 (10-14)	16 (14-18)	73 (70-75)	<0.001
High	122	2836	13 (11-16)	18 (16-20)	69 (66-71)	
Type of biopsy						
Glass						
Core needle	138	3953	11 (9-13)	14 (13-16)	75 (73-77)	0.35
Excisional	102	2947	9 (7-10)	15 (13-18)	76 (74-78)	
Digital						
Core needle	138	3207	15 (12-17)	15 (13-17)	70 (68-73)	0.61
Excisional	102	2373	10 (8-12)	19 (17-22)	71 (68-74)	
Pathologist assessment of test case		**Number of interpretations**	**Percentage of diagnoses overinterpreted (95% CI)**	**Percentage of diagnoses underinterpreted (95% CI)**	**Agreement rate with reference diagnosis (95% CI)**	P^a
Difficulty rating						
Glass						
Low (1-3)		4829	6 (5-7)	13 (11-15)	81 (79-83)	<0.001
High (4-6)		2071	19 (17-22)	19 (16-22)	62 (59-64)	
Digital						
Low (1-3)		3432	8 (6-9)	15 (13-17)	77 (75-79)	<0.001
High (4-6)		2148	20 (17-23)	19 (17-22)	60 (58-63)	
Case considered "borderline"						
Glass						
No		5097	7 (5-8)	13 (11-15)	81 (79-82)	<0.001
Yes		1803	19 (17-23)	20 (17-23)	60 (57-64)	
Digital						
No		4208	9 (8-11)	15 (13-17)	75 (73-78)	<0.001
Yes		1372	22 (19-26)	22 (19-25)	56 (53-60)	
Second opinion desired						
Glass						
No		4449	6 (5-7)	12 (11-14)	82 (80-84)	<0.001
Yes		2451	17 (15-20)	20 (17-23)	63 (60-66)	

Contd...

Appendix 3: Contd...

Pathologist characteristics (*n*=115 glass, 93 WSI)	Number of pathologists	Number of interpretations	Percentage of diagnoses over-interpreted (95% CI)	Percentage of diagnoses under-interpreted (95% CI)	Agreement rate with reference diagnosis (95% CI)	*P*[a]
Digital						
No		3208	7 (6-9)	15 (13-17)	78 (76-80)	<0.001
Yes		2372	20 (17-23)	20 (17-22)	61 (58-64)	
Confidence in assessment						
Glass						
High		5640	8 (7-9)	13 (12-15)	79 (77-80)	<0.001
Low		1260	19 (15-24)	21 (17-26)	60 (55-65)	
Digital						
High		4385	11 (9-13)	16 (14-18)	73 (71-75)	<0.001
Low		1195	18 (15-23)	20 (17-24)	61 (57-66)	

[a]*P* value for covariate effect on agreement rate, [b]A test for trend based on a logistic regression model, which includes a single 4-category ordinal variable for pathologist age category, [c]*P* value comparing none versus 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]<10 versus ≥10 other pathologists in the same laboratory who also interpret breast tissue, [f]Expertise defined as self-reported completion of a fellowship in breast pathology and/or their peers considering them an expert in breast pathology

Appendix 4: Reproducibility of interpretations: Intraobserver agreement between interpretations of the same case in Phase I and Phase II by diagnostic format used for interpretation. Data are shown by the reference diagnosis of the case (*n*=172 pathologists with a total of 20,640 individual case assessments)[a]

Diagnostic format		Number of pathologists (*n*)	Number of interpretations (*n*)	Percentage agreement between Phase I and Phase II (95% CI) Reference diagnosis				Overall agreement
Phase I	Phase II			Benign without atypia	Atypia	DCIS	Invasive	
Glass	Glass	49	5880	87 (84-90)	59 (56-62)	85 (83-87)	95 (91-98)	79 (77-81)
Glass	Digital	45	5400	81 (78-84)	62 (58-65)	81 (77-84)	97 (94-98)	77 (75-79)
Digital	Glass	37	4440	83 (80-85)	59 (54-64)	82 (77-85)	93 (87-96)	76 (74-79)
Digital	Digital	41	4920	78 (73-82)	56 (51-60)	79 (76-82)	95 (91-97)	73 (71-76)

[a] ≥9 months between Phase I and Phase II. An interpretation by a participating pathologist was considered “in agreement” if the pathologist diagnosed the case in the same category in Phase I and Phase II; the diagnosis did not necessarily need to agree with the reference standard. When the format changed between phases, the pathologists’ average overall agreement between diagnoses in the two phases was 77% (95% CI: 75-78), compared with 79% (95% CI: 77-81) when glass slides were used in both phases (*P*=0.08). The agreement of pathologists interpreting the same cases using the same format in both phases was 73% (95% CI: 71-76) for digital, compared to 79% (95% CI: 77-81) for glass (*P*<0.001). CI: Confidence interval

Appendix 5: Probability that a pathologist’s interpretation of a single-slide breast biopsy specimen will be verified by the reference consensus interpretation in the U.S. population of women aged 50-59 years having screening mammography

Glass format

Pathologist interpretation	Probability of reference consensus interpretation (95% CI), %[a]				Total, %
	Benign without atypia	Atypia	DCIS	Invasive breast cancer	
Benign without atypia	**97.1 (96.7-97.4)**	2.1 (1.9-2.4)	0.6 (0.5-0.7)	0.2 (0.0-0.4)	100
Atypia	53.6 (47.9-58.3)	**37.8 (33.6-42.7)**	8.6 (7.0-10.5)	0.0 (0.0-0.0)	100
DCIS	9.5 (5.7-13.6)	9.0 (7.8-10.2)	**69.6 (64.4-75.3)**	11.8 (7.6-15.7)[b]	100
Invasive breast cancer	1.6 (0.7-2.7)	0.1 (0.0-0.1)	0.6 (0.4-0.9)	**97.7 (96.5-98.7)**	**100**

Digital format

Pathologist interpretation	Probability of reference consensus interpretation (95% CI), %[a]				Total, %
	Benign without atypia	Atypia	DCIS	Invasive breast cancer	
Benign without atypia	**95.7 (95.0-96.4)**	2.2 (2.0-2.4)	1.0 (0.9-1.1)	1.1 (0.5-1.7)	100
Atypia	62.7 (56.6-67.8)	**27.8 (23.9-32.5)**	8.8 (7.0-10.8)	0.8 (0.0-2.5)	100
DCIS	21.0 (15.2-26.4)	9.8 (8.4-11.2)	**57.1 (50.6-64.8)**	12.2 (7.4-16.5)[b]	100
Invasive breast cancer	2.1 (0.8-3.8)	0.1 (0.1-0.2)	0.5 (0.3-0.7)	**97.2 (95.6-98.6)**	**100**

[a]Boldface values indicate probabilities of verification by the reference consensus interpretation (i.e., predictive values), [b]This estimate may have been influenced by one case of DCIS with focal microinvasion that was difficult to identify and was frequently diagnosed as DCIS by study participants. The reference panel noted that this microinvasive focus would not significantly change the treatment or outcome. DCIS: Ductal carcinoma *in situ*, CI: Confidence interval