Journal of Digital Imaging

Characterizing Diagnostic Search Patterns in Digital Breast Pathology: Scanners and Drillers

--Manuscript Draft--

Manuscript Number:	JDIM-D-17-00032R1
Full Title:	Characterizing Diagnostic Search Patterns in Digital Breast Pathology: Scanners and Drillers
Article Type:	Original Paper
Funding Information:	National Cancer Institute (RO1 CA172343) — Dr Joann G Elmore National Cancer Institute (US) (RO1 CA140560) — Dr Joann G Elmore National Cancer Institute (K05 CA104699) — Dr Joann G Elmore
Abstract:	Following a baseline demographic survey, 87 pathologists interpreted 240 digital whole slide images of breast biopsy specimens representing a range of diagnostic categories from benign to atypia, ductal carcinoma in situ, and invasive cancer. A web-based viewer recorded pathologists' behaviors while interpreting a subset of 60 randomly selected and randomly ordered slides. To characterize diagnostic search patterns, we used the viewport location, time stamp, and zoom level data to calculate four variables: average zoom level, maximum zoom level, zoom level variance, and scanning percentage. Two distinct search strategies were confirmed: Scanning is characterized by panning at a constant zoom level, while drilling involves zooming in and out at various locations. Statistical analysis was applied to examine the associations of different visual interpretive strategies with pathologist characteristics, diagnostic accuracy, and efficiency. We found that females scanned more than males, and age was positively correlated with scanning percentage, while the facility size was negatively correlated. Throughout 60 cases, the scanning percentage and total interpretation time per slide decreased, and these two variables were positively correlated. The scanning percentage was not predictive of diagnostic accuracy. Increasing average zoom level, maximum zoom level, and zoom variance were correlated with over-interpretation.
Corresponding Author:	Ezgi Mercan University of Washington UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Washington
Corresponding Author's Secondary Institution:	
First Author:	Ezgi Mercan, MSc
First Author Secondary Information:	
Order of Authors:	Ezgi Mercan, MSc Linda G Shapiro, PhD Tad T Brunye, PhD Donald L Weaver, M.D. Joann G Elmore, M.D., M.P.H.
Order of Authors Secondary Information:	
Author Comments:	

ABSTRACT

Following a baseline demographic survey, 87 pathologists interpreted 240 digital whole slide images of breast biopsy specimens representing a range of diagnostic categories from benign to atypia, ductal carcinoma *in situ*, and invasive cancer. A web-based viewer recorded pathologists' behaviors while interpreting a subset of 60 randomly selected and randomly ordered slides. To characterize diagnostic search patterns, we used the viewport location, time stamp, and zoom level data to calculate four variables: average zoom level, maximum zoom level, zoom level variance, and scanning percentage. Two distinct search strategies were confirmed: *Scanning* is characterized by panning at a constant zoom level, while *drilling* involves zooming in and out at various locations. Statistical analysis was applied to examine the associations of different visual interpretive strategies with pathologist characteristics, diagnostic accuracy, and efficiency. We found that females scanned more than males, and age was positively correlated with scanning percentage, while the facility size was negatively correlated. Throughout 60 cases, the scanning percentage and total interpretation time per slide decreased, and these two variables were positively correlated. The scanning percentage was not predictive of diagnostic accuracy. Increasing average zoom level, maximum zoom level, and zoom variance were correlated with over-interpretation.

Keywords: digital pathology, diagnostic decision making, breast cancer, breast histopathology, whole slide imaging, diagnostic interpretation.

BACKGROUND AND SIGNIFICANCE

Digital imaging technologies have revolutionized clinical medicine, particularly within diagnostic radiology. In pathology, digital whole slide images (WSI) are well established and have proven efficient and reliable for research[1], education[2–5], and archiving[6], and are now being utilized in pathologic diagnosis[7,8]. Although they are not approved by the U.S. FDA for primary pathologic diagnosis, digital WSI are increasingly used to obtain second opinions remotely. In addition to their advantages in clinical settings, the use of computers to interpret digital WSI provides a unique opportunity to study pathologists' viewing behaviors and better understand how their interpretive strategies relate to diagnostic accuracy and efficiency.

Pathologic diagnosis is a complex process characterized by visual search and interpretation strategies. Previous research concerning the visual search patterns of physicians has focused on volumetric lung images,[9–11] mammography,[12] and breast pathology.[13,14] The method of investigation has usually included eye-tracking or video recordings of physicians interpreting medical images in a setting controlled by the experimenter. Three outcomes from published research are relevant to the present study. First, physicians reviewing medical images tend to adopt one of two search strategies: *drilling* versus *scanning*. Drilling involves restricting a search to a region of interest and zooming in to high magnification levels. Conversely, scanning involves maintaining a particular zoom level while searching relatively broad regions of interest.[11] Second, search strategies change as a function of acquired experience in an expert domain[9,15] and prior experience with novel review formats.[16] Third, certain visual search strategies have been associated with greater diagnostic accuracy and efficiency. In radiology, physicians who use a drilling search pattern tend to show higher accuracy and efficiency when detecting lung nodules in volumetric images,[11,17] though no research has explored drilling and scanning strategies by pathologists reviewing non-volumetric images.

To address this knowledge gap, our study attempts to provide an initial understanding of the interpretative strategies pathologists use when reviewing digital slides of breast biopsy specimens. In this study, we investigated three aims. First, we considered how various pathologist characteristics are associated with the two image review strategies (drilling and scanning) identified in the extant cognitive science literature[11]. Second, we tracked how these image review strategies may change as pathologists gain experience with the digital imaging format. Finally, we examined the extent to which each interpretive strategy is associated with diagnostic accuracy and efficiency.

While digital slides are becoming a powerful adjunct tool for breast pathology, understanding the diagnostic processes used by pathologists as they interpret cases may provide insight to improve the education and training of pathologists and lead to the development of computational tools that can aid in the diagnostic decision making process.

MATERIALS AND METHODS

Data were collected as part of the Breast Pathology (B-Path) and Digital Pathology (digiPATH) studies. The detailed explanation of methods used for test case development and recruitment of participant pathologists has been previously described[18,19] and is briefly summarized below.

Case Selection

The 240 excisional (N=102) and core (N=138) breast biopsy specimens were selected from pathology registries in Vermont and New Hampshire using a random sampling stratified by woman's age, breast density (N=118 low density and N=122 high density), and initial diagnosis. New glass slides were prepared from the selected tissue blocks.

The newly prepared glass slides were scanned at 40x magnification (iScan Coreo, Ventana Medical Systems, Tuscon, AZ, USA) to create digital WSIs, which were then reviewed by a research technician and a breast pathologist to ensure consistency and quality. A web-based digital viewer, which was developed specifically for this study, allowed users to pan the image and zoom in or out (up to 40x actual and 60x digital magnification), providing an interface similar to industry-sponsored WSI viewers but enhanced with study-specific data collection capabilities.

Expert Consensus Diagnosis

The digital WSIs were independently interpreted by three experienced breast pathologists to determine independent diagnoses and representative regions of interest (ROIs); these pathologists then established a consensus diagnosis for each case following a modified Delphi approach in subsequent webinars and in-person meetings.[19,20] Cases spanned a wide range of diagnostic categories: benign without atypia (N=60), atypia (N=80), ductal carcinoma *in situ* (DCIS) (N=78), and invasive cancer (N=22). See Supplementary Table 1 for details.

Participants

More than 200 pathologists from across the U.S. (Alaska, Maine, Minnesota, New Hampshire, New Mexico, Oregon, Vermont, and Washington), who regularly interpret breast biopsy specimens in their clinical practices, were invited to participate in the study. Each participant completed a baseline survey that included demographic data, experience with breast pathology, and perceptions about breast cancer interpretation.

Each participant was randomly assigned to interpret the cases in glass or digital format. A small portion of the participants did not complete the study. In this work, we are using the data collected from 87 pathologists who were assigned to digital format.

Data Collection on Interpretations

The 240 cases were arranged into 4 sets of 60 cases each that preserved the distribution of diagnostic categories and breast densities of the overall case set. Participants were randomly assigned to interpret one of the four test sets. The order of the 60 cases was randomized for each participant, and they interpreted each case independently, considering histopathological features as well as accompanying information regarding patient age and biopsy type. After viewing each case, participants were instructed to select all applicable diagnoses on an electronic histology form listing 14 possible diagnostic interpretations. The same categorical mapping scheme was used for participant diagnoses as was used for the expert consensus diagnoses (see Supplementary Table 1).

The study was conducted in two phases so that each participant interpreted the same test set twice, either in glass slide format or digital slide format or both. The study is explained in detail in [21]. Participants were not informed that they were seeing the same cases in phase II and the cases were presented in a different order for each participant and also were randomly reordered in phase II.

Detailed tracking data were automatically logged by the web-based digital viewer. As pathologists navigated each slide, the viewer software logged their coordinate positions in the digital WSI, their magnification (zoom) levels, and time stamps.

Tracking data were collected only for those interpreting the cases in digital format in the phase II. Half of the participants in phase II were then asked to electronically annotate the digital WSI with an ROI supporting the highest order (most severe) diagnosis while the other half were not asked to mark an ROI on the digital image. This was done to control for any potential impact of the ROI placement task on the diagnostic decision-making process. The participants randomized to mark the ROI used a tool in the web-based viewer to draw a rectangular ROI following their diagnostic interpretation. The relationship between ROI identification and diagnostic concordance was explored in [20].

Tracking Data Analysis

A viewport scene is a rectangular part of the image that is visible on the pathologist's computer monitor at any time during an interpretation. The time spent on each viewport scene was calculated using logged timestamps. If an entry exceeded a total duration of one minute, it was excluded under the assumption that the pathologist was not actively interpreting during that time. From the tracking logs, several variables were calculated to characterize the viewing behaviors of each participant, as described below.

Average zoom level, maximum zoom level, and zoom level variance: The web-based viewer allowed zoom levels from 1x to 60x. For each interpretation, viewport tracking logs provided a variable number of zoom level values depending on pathologists' interpretive behavior; for this reason, summary statistics were used to describe zoom level behavior during each interpretation. Average and maximum zoom levels, as well as zoom level variance, were calculated for each interpretation. For each interpretation, we calculated the average zoom level by summing the zoom level values of all viewport scenes and dividing by the number of viewport scenes. Similarly, we calculated the maximum zoom level of each interpretation and the standard deviation of the zoom level variable as the zoom level variance.

Scanning percentage: We quantified scanning behavior by calculating the percentage of log entries associated with panning behavior (i.e., changing viewport scene coordinates) in each interpretation. Unlike average zoom level, maximum zoom level and zoom level variance, scanning percentage considers the changes of zoom level in consecutive log entries, regardless of the zoom level itself. In other words, scanning percentage quantifies a behavior that can manifest at different zoom levels. Scanning percentage approaches 100% when the pathologist pans across different areas of the digital image at a constant zoom, and it approaches 0% when zooming in and out at different locations, with less panning or infrequent but long distance pans at a low zoom magnification. For analysis, the scanning percentages were grouped into five categorical variables (0-20%, 20-40%, 40-60%, 60-80%, 80-100%).

Analysis

To assess how pathologist demography influenced interpretive strategy, we modeled our data using repeated-measures regressions, implementing the generalized estimating equation (GEE) approach. The model included 10 categorical predictors (factors), as detailed in Table 1. The model used scanning percentage as a linear dependent variable (outcome).

To assess how case order within each set of 60 cases influenced viewing behaviors, we again modeled our data using repeated-measures regressions, implementing the GEE approach. We implemented two models, both including interpretation order as the continuous predictor. We used a linear dependent variable (outcome) for both models: scanning percentage for the first model and total interpretation time per case for the second model.

To assess how interpretive strategy influenced diagnostic outcome, we conducted four separate repeated-measures analyses of variance (ANOVA) with four variables that describe the interpretative behaviors. Each model included one of four continuous variables (average zoom, maximum zoom level, zoom level variance, or scanning percentage) and one of three categorical dependent variables for diagnostic outcome (over-interpretation compared to the expert consensus diagnosis, concordance with the expert consensus diagnosis, and under-interpretation compared to the expert consensus diagnosis). To assess the effect of interpretative behaviors on diagnostic efficiency, we used a repeated-measures ANOVA with a continuous dependent variable (time) and one of five independent categorical variables (scanning percentage: 0-20%, 20-40%, 40-60%, 60-80%, or 80-100%).

RESULTS

Viewport tracking data from 87 pathologists, who completed 60 cases in the digital format, were analyzed, producing a total of 5,220 interpretations and approximately 1.03 million viewport log entries. 907 entries were excluded because they exceeded 1-minute in total duration.

Tracking logs were visualized and analyzed to summarize the interpretive strategy of each pathologist. Figure 1 contrasts visualizations representing two different pathologists. The pathologist represented on the left, a *scanner*, chose a consistent zoom level and systematically panned to investigate the whole image. The scanner pathologist used the same zoom level on the majority of their cases. In contrast, the pathologist represented on the right, a *driller*, zoomed out periodically, selected a new area to view, then zoomed in again. The driller pathologist zoomed in and out on different regions throughout their interpretations. It could be argued that the driller scanned the image with eye movements (rather than screen pans) at a lower resolution to determine areas for drilling. Some of the scanning

versus drilling strategies may reflect the pathologist's comfort level when scanning with eye movements at lower magnifications. The scanning percentage for the visualization on the left is close to 100%, while it is closer to 0% for the visualization on the right.

Table 1. Characteristics and average scanning percentages of pathologists (N=87)

Variable		Number of Pathologists	Average Scanning Percentage	p-value	Wald Chi-Square
Age at survey (years)					
	30-39	10 (11%)	69%	0.041	8.251
	40-49	25 (29%)	77%		
	50-59	36 (41%)	75%		
	60+	16 (18%)	70%		
Gender					
	Male	57 (66%)	70%	0.035	4.439
	Female	30 (34%)	82%		
Affiliation with academic medical center					
	Yes	19 (22%)	77%	0.642	0.216
	No	68 (78%)	73%		
Facility size					
	< 10 pathologists	55 (63%)	76%	0.019	5.484
	≥ 10 pathologists	32 (37%)	69%		
Fellowship training in surgical or breast pathology					
	No	41 (47%)	75%	0.076	3.141
	Yes	46 (53%)	73%		
Do your colleagues consider you an expert in breast pathology?					
	No	70 (80%)	73%	0.103	2.666
	Yes	17 (20%)	79%		
Breast pathology experience (years)					
	<20	65 (75%)	76%	0.073	3.210
	≥ 20	22 (25%)	68%		
Number of breast cases per week					
	<5	19 (22%)	73%	0.490	1.426
	5-9	36 (41%)	75%		
	≥ 10	32 (37%)	72%		
Marked an ROI					
	Yes	44 (51%)	71%	0.565	0.330
	No	43 (49%)	77%		
How confident are you in your assessments of breast cases?					
	1 (very confident)	13 (15%)	67%	0.100	7.783
	2	43 (49%)	75%		
	3	21 (24%)	75%		
	4	8 (9%)	77%		
	5 (not confident at all)	2 (2%)	83%		

Pathologist Demographics and Viewing Behaviors

Overall, pathologists tended to show scanning percentages exceeding 50% (μ = 74%, σ= 16%), demonstrating a disproportionate trend toward scanning rather than drilling. This pattern was confirmed with a one-sample t-test comparing to 50%, $t(86) = 13.53$, $p < .001$. However, this pattern also varied significantly as a function of certain pathologist demographics.

The GEE model goodness-of-fit was 1140.98 (QIC), with three significant main effects. First, age positively predicted increasing scanning percentage ($\chi^2 = 8.25$, $p < .05$), with higher age groups showing increasingly higher scanning percentages. Second, there were higher scanning percentages among female versus male pathologists ($\chi^2 = 4.44$, $p < .05$). Finally, facility group size negatively predicted scanning percentage ($\chi^2 = 5.48$, $p < .05$), with pathologists working in larger facility groups showing lower scanning percentages. No other patterns reached traditional ($\alpha = .05$) significance levels.

Interpretation Order

The GEE model showed a significant negative relationship between case position and scanning percentage ($\chi^2 = 16.01$, $p < .001$), with scanning percentage decreasing over the course of the 60 cases (see Figure 2). The total time spent on an interpretation of each case also decreased on average with interpretation order. The participants interpreted later cases in less time compared to earlier cases, ($\chi^2 = 67.36$, $p < .001$). In a previous study, we found that the diagnostic concordance with the expert panel does not change significantly over the sixty cases interpreted digitally [21].

Table 2. Zoom and scanning variables by concordance with expert consensus diagnosis.

Consensus Diagnosis	Concordance with Consensus	Number of Interpretations	Average Zoom Level	p-value	Maximum Zoom Level	p-value	Zoom Level Variance	p-value	Scanning Percentage	p-value
all	under	760	7.89	≤ 0.001	24.93	≤ 0.001	6.22	≤ 0.001	75%	0.574
	agree	3672	8.86		27.29		6.98		74%	
	over	788	9.94		31.87		8.10		73%	
benign	under	-	-	≤ 0.001	-	≤ 0.001	-	≤ 0.001	-	0.278
	agree	933	6.64		22.10		5.28		75%	
	over	348	9.71		31.49		7.98		72%	
atypia	under	492	7.39	≤ 0.001	23.21	≤ 0.001	5.78	≤ 0.001	76%	0.276
	agree	882	8.88		27.51		7.03		72%	
	over	384	10.00		31.79		8.13		73%	
DCIS	under	259	8.74	≤ 0.001	28.14	0.001	7.03	0.003	73%	0.365
	agree	1386	8.36		27.66		6.88		74%	
	over	56	10.95		34.82		8.64		78%	
invasive	under	9	10.40	0.312	26.67	0.79	7.73	0.763	74%	0.652
	agree	471	14.72		36.10		10.57		75%	
	over	-	-		-		-		-	

Diagnostic Concordance with Expert Consensus Diagnosis

The mean values of the average zoom level, maximum zoom level, zoom level variance, and scanning percentage variables for interpretations are shown by expert consensus diagnosis and concordance with expert consensus diagnosis in Table 2. Supplementary Table 2 provides detailed results of ANOVA tests.

Over-interpretation was associated with increased drilling (average zoom level, maximum zoom level, and zoom level variance). Average zoom level, maximum zoom level, and zoom level variance were higher than the expert consensus diagnosis for over-interpretations and were lower than the expert consensus diagnosis for under-interpretations. The trend was replicated in benign, atypia, and invasive cases. For DCIS cases, both over-interpretation and under-interpretation were associated with higher zoom values. All associations except those for invasive cases were statistically significant.

No association was noted between scanning percentage and accuracy (Table 2). Supplementary Figure 1 shows the average over-interpretation and under-interpretation rates within different scanning percentage groups.

Diagnostic Efficiency

Efficiency to arrive at an accurate diagnosis was negatively predicted by the extent to which pathologists followed a scanning strategy; in other words, higher scanning percentage was associated with lower efficiency. A repeated-measures ANOVA revealed a main effect of scanning percentage category, $F(4, 52) = 6.72$, $p < .001$, demonstrating significantly higher case review times as a function of increased scanning percentage. This pattern is depicted in Figure 3. Follow-up paired t-tests demonstrated significant differences between all pairwise category comparisons, with the exception of the first (0-20%) versus second (20-40%) categories, and fourth (60-80%) versus fifth (80-100%) categories. In contrast, rates of diagnostic concordance with the expert consensus diagnosis showed no significant difference across scanning percentage groups.

DISCUSSION

The field of pathology has begun adopting the digital WSI format as it offers great potential for teaching[2–5] and research[1], as well as archival purposes[6] and gathering second opinions[7,8]. To better understand the visual search patterns used in breast pathology, 87 pathologists across the U.S. interpreted 60 digital WSIs of breast biopsies representing a range of diagnostic categories, amounting to 5,220 individual independent interpretations for analysis.

A web-based viewer tracked and recorded the interpretive behaviors of pathologists as they viewed each digital WSI. The viewer provided pathologists with two possible actions: zooming and panning. Zooming in to an area allowed pathologists to examine cytological, cellular, and nuclear structural details, thereby revealing those that are not as visible to the human eye at lower magnification, but also limiting the portion of the whole slide image viewable on the screen. The panning action allowed pathologists to view neighboring areas of the whole slide image that were not viewable on the screen at higher magnifications.

Combinations of both actions were used by all pathologists to interpret the digital WSI, but interpretive patterns emerged when we analyzed the tracking logs. Specifically, we found that participants varied in their extent of panning and zooming behaviors over time and across cases. *Drilling* behavior showed a relative tendency to zoom in on a particular region, use panning actions sparingly to examine that region, and then zoom out to a lower magnification. In contrast, *scanning* behavior showed a relative tendency to use panning actions to systematically explore the complete image at a constant, and relatively low, magnification. We conceptualize drilling and scanning behavior as two complementary strategies falling at the ends of a bipolar continuum. To quantify image review behavior along this continuum, we calculated the proportion of case review behavior indicative of scanning (i.e., scanning percentage). We wanted to explore potential explanations for the interpretative strategies through their correlation with diagnostic accuracy and efficiency, as well as determining if these patterns change over time.

A number of pathologist demographic characteristics were associated with changes in scanning percentage, including age, gender, and facility size. Higher age was positively correlated with increased scanning percentage, females scanned more than males, and pathologists from smaller facility sizes had higher scanning percentages. It may have been the case that younger participants had relatively more prior experience with similar computer interfaces or image manipulation tools (e.g., mapping software, digital slide viewers, image editing software), thereby making them more comfortable with image drilling behaviors.[22] Although not a statistically significant trend, pathologists with higher scanning percentages also reported lower baseline confidence in their breast pathology skills. This finding suggests that increased scanning may be related to personality-level (e.g., neuroticism[23]) and/or situation-level (e.g., anxiety[24]) factors. The scanning percentage and total time per slide decreased as pathologists gained experience throughout the set of 60 cases. This suggests a learning curve where participants who started with a scanning-based strategy adopted a more hybrid approach of scanning and drilling as they interpreted through the digital images. This learning curve may be due to prior inexperience with digital slides and computer-based viewing systems that pathologists began to overcome through their experience in this study. Previous research shows a learning curve for interpreting mammograms before and after residency, suggesting a correlation between interpretive behavior and experience.[25] The participants who marked an ROI at the end of their interpretations had lower scanning percentages than those who did not mark an ROI, but the difference was not statistically significant. In other words, we could not find a link between the additional task of looking for a region of interest and the visual search strategy of the participants.

We also noted a pattern of over-interpretation at higher zoom levels. For all diagnostic categories except invasive cancer, the cases that were over-interpreted based on the expert consensus diagnosis had higher values of average zoom level, maximum zoom level, and zoom level variance. This relationship aligns with some research in the cognitive science and visual search literature. Specifically, when observers repeatedly examine a visual scene in detail, the probability of making an erroneous "guess" increases.[26] These inaccurate interpretations likely result from a failed match between perceived image features and stored histopathological features in their memory.

In order to analyze the association of scanning with accuracy and total interpretation time, we divided image interpretations into five categories based on scanning percentage. As scanning percentage increased, so did interpretation time, though rates of over- and under-interpretation were not affected. Scanning was found to be a less efficient strategy for diagnostic interpretation, and the results with the learning curve indicated that pathologists adopted a more balanced and efficient strategy as they progressed through the set of 60 cases.

There are a few reasons why scanning may prove a less efficient, and sometimes less effective [11], method for searching visual images. Scanning at a moderate magnification level involves constantly monitoring and updating past and current positions relative to the entire image space, when only small portions of the overall image can be seen at a time. As demonstrated in prior literature, this type of constant monitoring can be very intensive for working memory, particularly when it is done simultaneously to a more important (primary) task (i.e., identifying malignancy) [27,28]. In contrast, drilling enables a pathologist to focus attention on a single well-defined region at a time: examining a single region of interest in great depth and detail, and then iteratively returning to low magnification and examining the next region. In this manner, the searcher need only remember which salient region(s) they have or have not already "drilled into," which involves monitoring and updating only a representation of salient regions in the low magnification space. The present results speak to the relative efficiency of drilling, suggesting support for this possibility; however, no research has specifically examined the relative memory cost of employing drilling versus scanning search strategies.

Recent research on volumetric lung images revealed that radiologists adopt distinct visual search strategies during interpretation.[11] Though this earlier research used eye-tracking to monitor and interpret visual search patterns, our findings suggests that similar distinctions can be ascertained by recording zooming and panning behaviors. We expect this is specifically the case with 2D digital pathology images. Indeed, these images require pathologists to zoom in and out dramatically in order to magnify breast tissue and reveal specific structural and cellular features. This process

results in high-density zooming and panning data, which is likely uncharacteristic of viewing behavior with narrow slices of volumetric images. The unique characteristics of these breast biopsy digital WSIs may explain why our data did not suggest any influence of visual search strategy (i.e. scanning percentage) on diagnostic accuracy, unlike earlier research with volumetric lung radiographs.[11,17] Of course, when attempting to identify specific structural or cellular features that were viewed or neglected during the interpretive process, eye-tracking is an invaluable technique.

Several notable works in pathology studied the diagnostic search patterns on digital slides[12,13,15,29–31]. The work of Krupinski et al. on breast pathology suggested a link between expertise and search patterns [12,13,15,31]. In our study, the participants with more than 20 years of experience had lower scanning percentages yet the correlation was not significant ($p = 0.1$). However, our participant cohort included only practicing pathologists in comparison to the studies by Krupinski et al. which recruited trainees and experts to examine the changes in the search patterns. The work of Treanor et al. compared the localization errors with interpretation errors in oesophageal biopsies and found a trend in which lower zoom levels are correlated with inaccurate diagnosis [29]. Our findings suggest that an opposite trend exists in breast pathology, where higher zoom levels are correlated with over-diagnoses of the pre-invasive lesions and there seems to be a "happy medium" of magnification for an accurate diagnosis. The existence of a link between magnification and diagnostic accuracy is an important insight but the nature of the relationship depends on the biopsy type and the visual characteristics of the tissue. Finally, Mello-Thoms et al. described a "focused and efficient" strategy that correlated with the correct outcome in dermatopathology [30]. In our study, we found that drilling is the more efficient strategy in terms of interpretation time but we did not find any links to the diagnostic outcome. Putting the differences in breast and skin biopsies aside, the selection of the diagnostic categories, the difficulty of the cases and the demographics of the pathologists are all important factors when comparing two studies. To the best of our knowledge, our work is the first to use an objective quantification of the viewing behavior in a large study.

Limitations and Strengths

This study was limited to one slide per case, which does not reflect actual clinical practice—a factor that may influence diagnostic accuracy but does not preclude evaluation of interpretive strategies. However, the one-slide-per-case study design reduced the workload of participants and allowed them to interpret more images representing a variety of tissue characteristics. This study also diverged from clinical practice in the distribution of diagnostic categories among the cases participants interpreted. Atypia and DCIS cases were oversampled compared to actual clinical prevalence, with the purpose of better understanding the interpretation of these diagnostically difficult non-invasive cases. Previous research shows that atypia and DCIS cases are more likely to be overinterpreted or misinterpreted, so it is crucial to understand interpretive behaviors on these diagnoses[18]. A possible limitation was the participants' prior inexperience with the digital format or digital viewers. Although the field of pathology has begun to incorporate digital WSI, most U.S. pathologists are still inexperienced with software for digital WSI interpretation, making it difficult to dissociate the relative contributions of experience with the digital format versus expertise in breast pathology on drilling versus scanning. Similarly, some variation between pathologists may be attributable to participants using their own computer monitors; it is therefore possible that identical monitors may standardize the pathologists' experience viewing digital WSI. However, identical monitors do not reflect actual clinical practice, where monitors vary at the level of the practice and, often, between pathologists at the same facility.

Limitations aside, this study is a timely and unique investigation of pathologists' interpretive strategies with digital media. Strengths of this study include the large sample size of breast biopsy cases (N=240) representing a full spectrum of diagnostic categories from benign and atypia to DCIS and invasive cancer, interpreted and diagnosed by three expert pathologists to define diagnostic accuracy. Another strength is the large number of practicing pathologists (N=87) from across the U.S. while the previous studies in the literature had recruited medical students and residents in small numbers, i.e. 4 to 11 pathologists. The use of a web-based viewer allowed participants to use their own computers in their own time, which is as close to the real-world practice of digital pathology as possible. Furthermore,

the order of the 60 cases was random for each participant, which allowed us to see a learning curve for the digital slide viewer without case biases attributable to interpretive difficulty or severity of diagnosis.

CONCLUSIONS

We identified two distinct interpretive strategies as pathologists viewed digital whole slides of breast biopsy specimens: *scanning*, where the pathologist pans at a constant zoom level, and *drilling*, where the pathologist zooms in and out repeatedly. Our analysis of pathologist characteristics indicated that scanning was more common among women and older pathologists. The facility size, defined as the number of pathologists who worked in the same facility as the participant, was also a significant predictor of the scanning percentage with the participants from smaller facilities scanning more. One possible explanation for this correlation is that the participants from larger facilities could share cases with their colleagues, obtain second opinions and learn from each other; they would have more experience interpreting breast biopsies. Those who reported less confidence in their interpretation of breast tissue tended to spend more time scanning but the correlation was not statistically significant.

Regarding accuracy and efficiency, we found that scanning is associated with longer interpretation time on average, yet scanners and drillers had similar levels of accuracy compared to the consensus reference diagnoses. Through our unique study design that randomized the order of cases, we also observed that scanning may be more common at the beginning of a pathologist's experience interpreting cases in the digital format, while a more balanced strategy of both scanning and drilling is adopted by the end of the 60 cases. We found that average zoom level, maximum zoom level, and zoom level variance for an interpretation increased from under-interpretation to concordance and from concordance to over-interpretation. In other words, when participants under-interpreted a case, they used lower magnifications and changed the zoom level less, as compared to concordant interpretations. Similarly, when they over-interpreted a case, they used higher magnifications and changed the zoom level more, as compared to concordant interpretations. This trend was preserved for all diagnostic categories of breast tissue.

In conclusion, this study demonstrates that two different search strategies are employed by pathologists, and these strategies can be explained by a pathologist's demographics, breast pathology perceptions, and prior experience viewing the digital format. The interpretive strategy can affect the diagnostic outcome and the efficiency of the diagnostic process. These findings motivate further research in medical decision-making and computerized decision support systems as digital pathology is adopted more widely.

ACKNOWLEDGEMENTS

Research reported in this publication was supported by the National Cancer Institute awards 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. We 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/.

COMPETING INTERESTS

The authors have no competing interests to declare.

REFERENCES

1 Irshad H, Veillard A, Roux L, *et al.* Methods for nuclei detection, segmentation, and classification in digital histopathology: A review-current status and future potential. *IEEE Rev Biomed Eng* 2014;**7**:97–114. doi:10.1109/RBME.2013.2295804

2 Yin F, Han G, Bui MM, *et al.* Educational Value of Digital Whole Slides Accompanying Published Online Pathology Journal Articles: A Multi-Institutional Study. *Arch Pathol Lab Med* 2016;**140**:694–7. doi:10.5858/arpa.2015-0366-OA

3 Saco A, Bombi JA, Garcia A, *et al.* Current Status of Whole-Slide Imaging in Education. *Pathobiology* 2016;**83**:79–88. doi:10.1159/000442391

4 Kumar RK, Freeman B, Velan GM, *et al.* Integrating histology and histopathology teaching in practical classes using virtual slides. *Anat Rec - Part B New Anat* 2006;**289**:128–33. doi:10.1002/ar.b.20105

5 Bruch LA, De Young BR, Kreiter CD, *et al.* Competency assessment of residents in surgical pathology using virtual microscopy. *Hum Pathol* 2009;**40**:1122–8. doi:http://dx.doi.org/10.1016/j.humpath.2009.04.009

6 Gutman D, Cobb J, Somanna D. Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *... Med Informatics ...* 2013;**20**:1091–8. doi:10.1136/amiajnl-2012-001469

7 Al-Janabi S, Huisman A, Van Diest PJ. Digital pathology: Current status and future perspectives. Histopathology. 2012;**61**:1–9. doi:10.1111/j.1365-2559.2011.03814.x

8 Pantanowitz L, Valenstein PN, Evans AJ, *et al.* Review of the current state of whole slide imaging in pathology. *J Pathol Inform* 2011;**2**:36. doi:10.4103/2153-3539.83746

9 Brunyé TT, Carney PA, Allison KH, *et al.* Eye Movements as an Index of Pathologist Visual Expertise: A Pilot Study. *PLoS One* 2014;**9**:e103447. doi:10.1371/journal.pone.0103447

10 Bahlmann C, Patel A, Johnson J, *et al.* Automated detection of diagnostically relevant regions in H&E stained digital pathology slides. Proc. SPIE, Med. Imaging. 2012;**8315**:831504. doi:10.1117/12.912484

11 Drew T, Vo ML, Olwal A, *et al.* Scanners and drillers: characterizing expert visual search through volumetric images. *J Vis* 2013;**13**:1–13. doi:10.1167/13.10.3

12 Tourassi G, Voisin S, Paquit V, *et al.* Investigating the link between radiologists' gaze, diagnostic decision, and image content. *J Am Med Inform Assoc* 2013;**20**:1067–75. doi:10.1136/amiajnl-2012-001503

13 Krupinski EA, Graham AR, Weinstein RS. Characterizing the development of visual search expertise in pathology residents viewing whole slide images. *Hum Pathol* 2013;**44**:357–64. doi:10.1016/j.humpath.2012.05.024

14 Crowley RS, Naus GJ, Stewart J, *et al.* Development of visual diagnostic expertise in pathology: An information-processing study. *J Am Med Informatics Assoc* 2003;**10**:39–51. doi:10.1197/jamia.M1123

15 Krupinski E a, Weinstein RS. Changes in visual search patterns of pathology residents as they gain experience. In: *Proceedings of SPIE*. 2011. 79660P. doi:10.1117/12.877735

16 Velez N, Jukic D, Ho J. Evaluation of 2 whole-slide imaging applications in dermatopathology. *Hum Pathol* 2008;**39**:1341–9. doi:10.1016/j.humpath.2008.01.006

17 Wen G, Drew T, Wolfe JM, *et al.* Computational assessment of visual search strategies in volumetric medical images strategies in volumetric medical images. *J Med Imaging* 2016;**3**. doi:10.1117/1.JMI.3.1.015501

18 Elmore JG, Longton GM, Carney PA, *et al.* Diagnostic Concordance Among Pathologists Interpreting Breast Biopsy Specimens. *Jama* 2015;**313**:1122. doi:10.1001/jama.2015.1405

19 Oster N V, Carney P a, Allison KH, *et al.* Development of a diagnostic test set to assess agreement in breast pathology: practical application of the Guidelines for Reporting Reliability and Agreement Studies (GRRAS).

2013;**13**:3.http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3610240&tool=pmcentrez&rendertype=abstract

20 Nagarkar DB, Mercan E, Weaver DL, *et al.* Region of interest identification and diagnostic agreement in breast pathology. Mod. Pathol. 2016;:1–8. doi:10.1038/modpathol.2016.85

21 Elmore J, Longton G, Pepe M, *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. doi:10.4103/2153-3539.201920

22 Elias SM, Smith WL, Barney CE. Age as a moderator of attitude towards technology in the workplace: Work motivation and overall job satisfaction. *Behav Inf Technol* 2012;**31**:453–67. doi:10.1080/0144929X.2010.513419

23 Newton T, Slade P, Butler NM, *et al.* Personality and performance on a simple visual search task. *Pers Individ Dif* 1992;**13**:381–2. doi:10.1016/0191-8869(92)90119-A

24 Wu S, Zhong S, Liu Y. Deep residual learning for image steganalysis. *Multimed Tools Appl* Published Online First: 15 February 2017. doi:10.1007/s11042-017-4440-4

25 Miglioretti DL, Gard CC, Carney PA, *et al.* When Radiologists Perform Best: The Learning Curve in Screening Mammogram Interpretation. *Radiology* 2009;**253**:632–40. doi:10.1148/radiol.2533090070

26 Chun MM, Wolfe JM. Just say no: how are visual searches terminated when there is no target present? *Cogn Psychol* 1996;**30**:39–78. doi:10.1006/cogp.1996.0002

27 Miyake A, Friedman NP, Emerson MJ, *et al.* The Unity and Diversity of Executive Functions and Their Contributions to Complex 'Frontal Lobe' Tasks: A Latent Variable Analysis. *Cogn Psychol* 2000;**41**:49–100. doi:https://doi.org/10.1006/cogp.1999.0734

28 Turner ML, Engle RW. Is working memory capacity task dependent? *J Mem Lang* 1989;**28**:127–54. doi:https://doi.org/10.1016/0749-596X(89)90040-5

29 Treanor D, Lim CH, Magee D, *et al.* Tracking with virtual slides : a tool to study diagnostic error in histopathology. 2009;:37–45. doi:10.1111/j.1365-2559.2009.03325.x

30 Mello-thoms C, Mello CAB, Medvedeva O, *et al.* Perceptual Analysis of the Reading of Dermatopathology Virtual Slides by Pathology Residents. ;:551–62. doi:10.5858/arpa.2010-0697-OA

31 Krupinski EA, Tillack AA, Richter L, *et al.* Eye-movement study and human performance using telepathology virtual slides . Implications for medical education and differences with experience B. 2006;:1543–56. doi:10.1016/j.humpath.2006.08.024

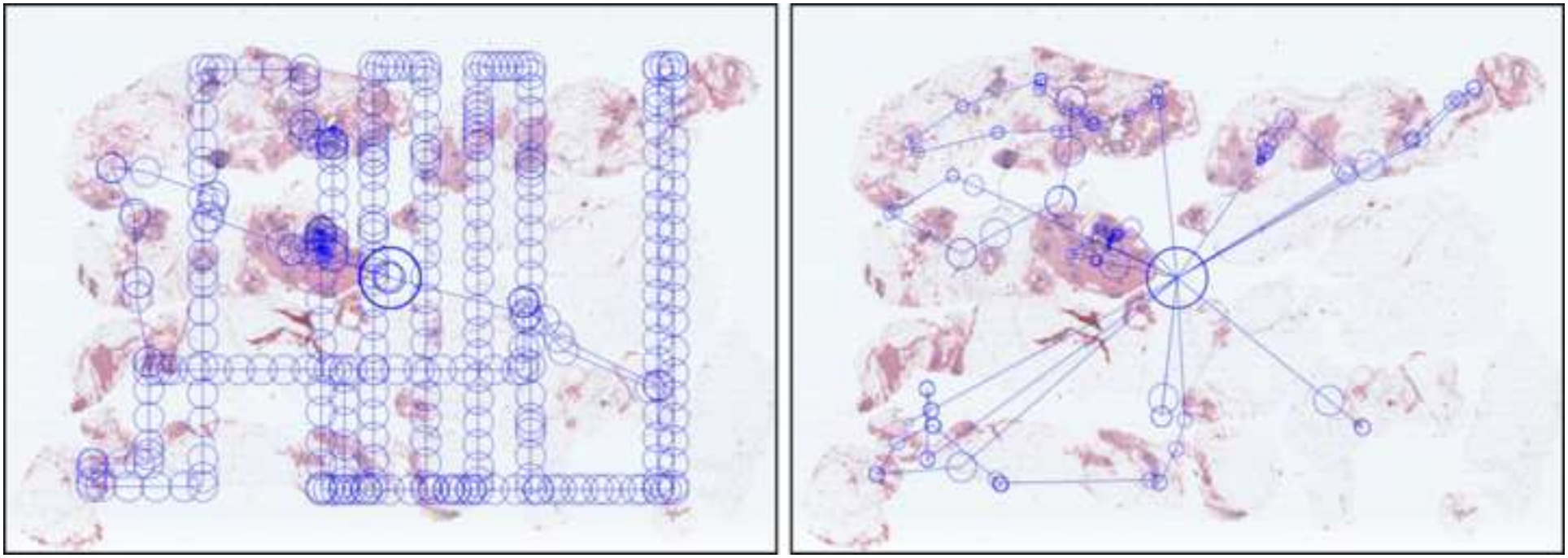
Figure 1. Visualization of viewport tracking logs a scanner (left) and a driller (right) on the same image. Each participant starts at the center of the image with a zoom level of 1x. The rings indicate the center of each viewport, the size of the rings indicate the zoom level (the larger the ring, the lower the zoom level), the thickness of the rings indicate the time spent at that viewport, and the lines connect consecutive viewports.

Figure 2. Average scanning percentage of 87 pathologists during the interpretation of 60 test cases. The order of the 60 cases was randomized for each pathologist so that *n*-th case included a random sampling of cases from all diagnostic categories.

Figure 3. Average total time of interpretation in five categories of scanning percentage.

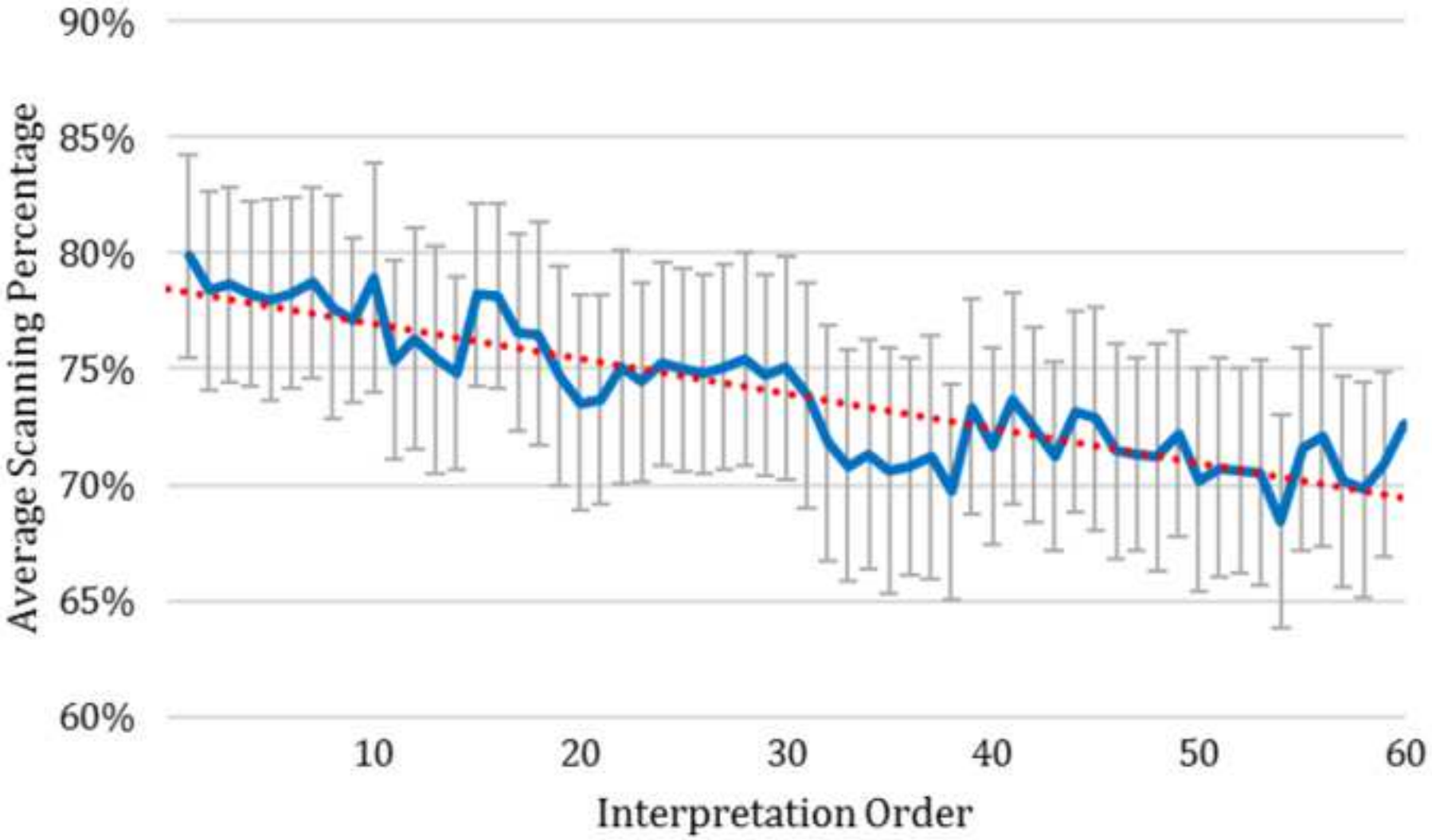
Figure

Click here to download Figure Fig1.tif



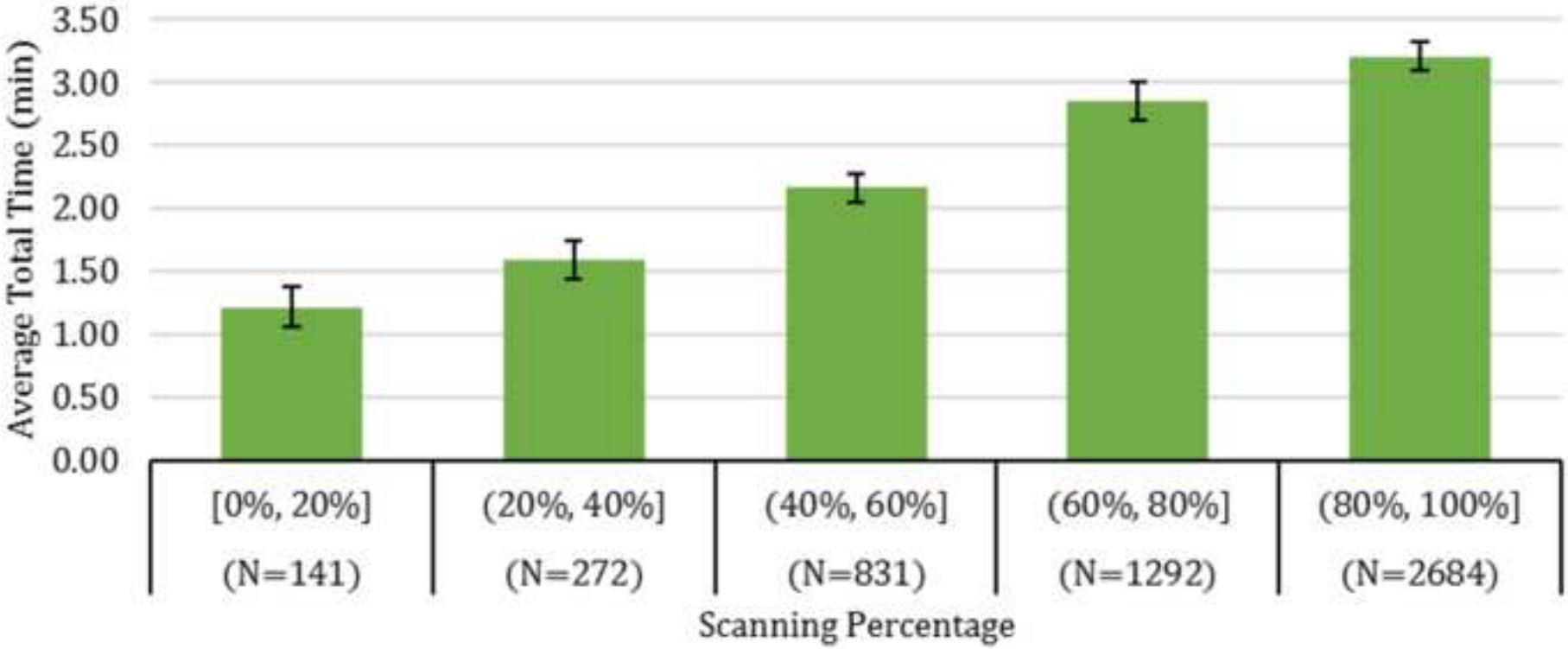
Figure

Click here to download Figure Fig2.tif



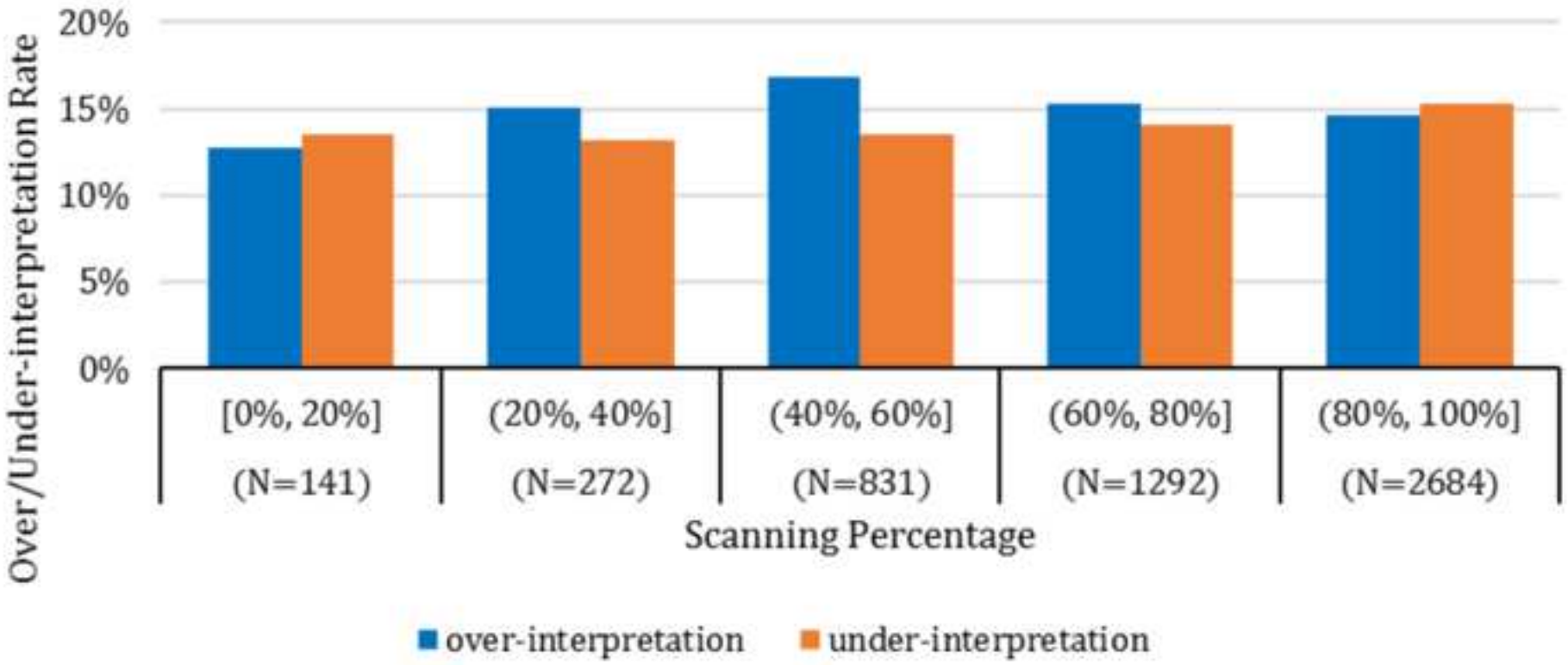
Figure

Click here to download Figure Fig3.tif



Supplementary Figure 1

Click here to download Supplementary Material FigSupp1.tif



Supplementary Table 1: Diagnostic category mapping scheme.

Diagnostic Interpretation	Diagnostic Category
Non-proliferative Changes Only	Benign without Atypia
Fibroadenoma (FA)[a]	Benign without Atypia
Intraductal Papilloma w/o Atypia (IP)	Benign without Atypia
Usual Ductal Hyperplasia (UDH)	Benign without Atypia
Columnar Cell Hyperplasia/ Columnar Cell Change (CCH/CCC)	Benign without Atypia
Sclerosing Adenosis	Benign without Atypia
Radial Scar/Complex Sclerosing Lesion	Benign without Atypia
Flat Epithelial Atypia (FEA)[b]	Atypia
Atypical ductal hyperplasia (ADH)	Atypia
Intraductal Papilloma with Atypia (IPA)	Atypia
ALH[c]	Atypia
Ductal Carcinoma in situ (DCIS)	DCIS
LCIS[c]	DCIS
Invasive (ductal or lobular or other special type)	Invasive

[a] FA is grouped with Benign without Atypia. FA is technically a proliferative lesion but has little associated risk.
[b] FEA was grouped with ADH because FEA may lead to excision in some institutions.
[c] ALH is grouped with ADH in the atypia category and LCIS is grouped with DCIS following traditional cancer progression schemes.

Supplementary Table 2. Detailed statistical outcomes of repeated-measures ANOVA for assessing how interpretative behavior affects diagnostic outcome

		factor: diagnostic concordance					Error		
		Sum of Squares	df	Mean Square	F	p-val	Sum of Squares	df	Mean Square
All	Average Zoom Level	160.712	2	80.356	33.304	≤ **0.001**	415.008	172	2.413
	Maximum Zoom Level	1999.169	2	999.585	44.022	≤ **0.001**	3905.499	172	22.706
	Zoom Level Variance	143.269	2	71.635	40.067	≤ **0.001**	307.514	172	1.788
	Scanning Percentage	0.003	2	0.002	0.557	0.574	0.466	172	0.003
Benign	Average Zoom Level	415.116	1	415.116	72.503	≤ **0.001**	458.039	80	5.725
	Maximum Zoom Level	3396.427	1	3396.427	55.561	≤ **0.001**	4890.373	80	61.130
	Zoom Level Variance	305.678	1	305.678	53.414	≤ **0.001**	457.824	80	5.723
	Scanning Percentage	0.007	1	0.007	1.193	0.278	0.482	80	0.006
Atypia	Average Zoom Level	179.997	2	89.998	17.569	≤ **0.001**	788.896	154	5.123
	Maximum Zoom Level	2661.627	2	1330.814	31.937	≤ **0.001**	6417.232	154	41.670
	Zoom Level Variance	173.739	2	86.870	20.329	≤ **0.001**	658.087	154	4.273
	Scanning Percentage	0.013	2	0.007	1.298	0.276	0.776	154	0.005
DCIS	Average Zoom Level	114.040	2	57.020	9.379	≤ **0.001**	364.778	60	6.080
	Maximum Zoom Level	678.115	2	339.057	8.196	**0.001**	2482.221	60	41.370
	Zoom Level Variance	47.336	2	23.668	6.325	**0.003**	224.531	60	3.742
	Scanning Percentage	0.016	2	0.008	1.024	0.365	0.472	60	0.008
Invasive	Average Zoom Level	10.156	1	10.156	1.164	0.312	69.803	8	8.725
	Maximum Zoom Level	7.347	1	7.347	0.076	0.790	777.518	8	97.190
	Zoom Level Variance	0.796	1	0.796	0.097	0.763	65.456	8	8.182
	Scanning Percentage	0.002	1	0.002	0.219	0.652	0.066	8	0.008

RE: **Characterizing Diagnostic Search Patterns in Digital Breast Pathology: Scanners and Drillers**
June 11, 2017

Dear Dr. Honeyman-Buck,

Thank you for giving us the opportunity to revise our manuscript. We appreciate the insightful feedback provided by each reviewer. We have addressed the questions and incorporated the suggested changes into the manuscript. We have responded specifically to each suggestion below and pointed to the corresponding changes in the revised manuscript.

Comments to the author (if any):

Reviewer #1: Overall Impression:

This is a very interesting paper that seeks to determine whether two search strategies previously observed in radiologists when reading 3D volumetric images can be found when pathologists read digital breast slides. The authors managed to get a fine cohort of pathologists (87) that have each read 60 cases chosen from a 240 cases database. Overall the study is well designed and the results are quite interesting. What I found myself mostly commenting on was on what was not in the paper, namely, a discussion between memory use and search strategy, a comparison of the findings reported here with what has been previously published (and a proposal for the differing results), etc. I believe that this paper has the potential to be very widely used by the scientific community, and I would like to see it better integrated with current findings so as to allow the readers to situate the authors' findings in the context of what is currently known.

Response: Thank you for your kind comments. We have included a comparison of the findings of our study with similar studies from the literature in the discussion section.

Abstract:

Page 1, line 12: The authors say that "two distinct search strategies were identified", but in reality it seems that, from previous work, they were already expecting to identify these two particular search strategies, that is, they were looking to confirm their presence in the digital breast slides. Hence, I don't think that the word "identified" is correct here; perhaps "confirmed" would better convey the fact that these were expected findings, not novelties.

Response: This is a very good point. Although we identified the patterns before the study on 3D volumetric lung images by Drew et al. was brought to our attention, we found them to be very similar. We changed the wording as suggested.

Materials and Methods:

Pp 3, line 5. Where is "Appendix 1"? I cannot find it.

Response: This is now clearly labeled as Supplementary Table 1.

Pp 3, line 15. Where is the information pertaining breast density? Supplementary Table 1 only contains diagnostic information, not information about breast density. Please provide this information, or at least a distribution of breast densities for each case set of 60 cases.

Response: The information about the breast density has been added to this paragraph.

Pp 3, line 22. Why only half of the participants was asked to mark an ROI? I can find no explanation for this, and it makes no sense to me. Please clarify.

Response: Only half of the participants were asked to provide an ROI to control for any potential impact of drawing an ROI on the diagnostic decision-making process. One could argue that the task of picking a final ROI might affect the search pattern and cognitive process during the interpretation. However, in our analysis, we did not find any correlation between the ROI selection task and diagnostic search pattern. We have now clarified in the revised manuscript the reason for randomizing the pathologists to marking ROIs.

Pp 3, line 37. Authors say that "tracking data were collected only for those interpreting the cases in digital WSI format in the second phase". Why was this done? To reduce memory effects? Please clarify.

Response: We did not have funding to develop the system to gather the tracking data during Phase I. It took our team about 2 years to collect all of the Phase I data and during that time we submitted grant proposals to the NIH/National Cancer Institute. During the very labor intensive 2 year process of collecting the Phase I data from participants we were able to obtain additional funding from the National Cancer Institute to develop the then employ the tracking system in time for Phase II data collection. If the reviewer wonders why it took about 2 years to collect Phase I data, we will also explain. While the digital image interpretations (the diagnosis) was fairly easy to collect in Phase I, the scheduling and logistics of data collection for the glass slide interpretations was hard due to having to schedule mailing the same glass slides to different participants to interpret on their own microscopes, have the participants return the original slides to us and then have our staff clean the slides and randomize the slide order into a new package to be sent to the next participant.

Pp 4, line 9. Where is the group 60-80%? It is present in the figures, just not in the categorical definition given here.

Response: This was a typo. It is corrected.

Pp 4, line 31. Same issue, missing 60-80%.

Response: This was a typo. It is corrected.

Results:

Pp 6, "Interpretation order". While these results are very interesting, the question I really have is, did the interpretation order affect the diagnostic accuracy??? I think it is nice to know that they scanned less and spent less time looking at the cases, but were there any diagnostic consequences from such behavior???

Response: This is a very relevant clinical question. In this study, our focus was on the underlying causes of the different diagnostic search strategies and their impact on the diagnostic accuracy and efficiency. In a previous study, we found that there was no learning curve over the sixty cases interpreted digitally in terms of diagnostic accuracy (Elmore et al., A randomized study comparing digital imaging to traditional glass slide microscopy for breast biopsy and cancer diagnosis, *Journal of Pathology Informatics*, 2017). We have also noted an interesting impact of the context of previous cases interpreted on the diagnostic accuracy (Frederick, P. D. et al. The Influence of Disease Severity of Preceding Clinical Cases on Pathologists ' Medical Decision Making. 91–100 (2017).). When pathologists have recently evaluated a high grade lesions such as invasive cancer, they are more likely to provide a higher diagnosis on a case (and vice a versa – when they have seen a few benign cases their diagnoses drop to lower level classes on subsequent cases). We have been able to study these interesting aspects of diagnostic behavior given our careful study design, randomizing each pathologists to a different order of case presentation. A sentence was added to the relevant results section.

Also in the Results, but with no specific page. I may have missed this, but what were the ROIs that half the pathologists drew used for? I don't seem to be able to find any analysis that takes advantage of this information.

Response: In a previous study, we explored the relationship between the ROI identification and diagnostic accuracy (Nagarkar et al., Region of interest identification and diagnostic agreement in breast pathology, *Modern Pathology*, 2016). This study found a correlation between the overlap of participant's ROI to the expert consensus ROI and the diagnostic concordance. A sentence was added to the relevant section.

We are also using ROIs from the expert panel in an image analysis study where we are developing and validating image features that can differentiate pre-invasive lesions of the breast: atypia and DCIS. Our automated diagnosis system achieves results comparable to the participants' performance on the same dataset. We are working on a manuscript to publish our findings.

Discussion:

Pp 7, lines 35-41. I imagine that the memory costs associated with scanning and drilling must be exceedingly different, and this may bear some discussing. To make a simplistic association, scanning seems to be like looking for diamonds in a somewhat flat mine - one that may have many caves, so one has to slowly progress from room to room and to keep track (in memory) of which rooms contained promising findings, while drilling is like looking for diamonds in a mine that is full of shafts, so in addition of remembering where the promising findings are, one has to remember at what level they were seen. In other words, more information has to be held in memory if the search is to be successful, so there seems to be a trade-off between a more time consuming strategy (scanning) versus a more memory-taxing strategy (drilling). If the authors agree, may they please expand on the role of memory in these two strategies?

Response: We thank you for this interesting insight. You are likely correct that drilling versus scanning behavior imparts variable working memory load, with one being more taxing than the other. As you suggest, drilling might involve a higher working memory load due to changing position in not only *xy* space (i.e., as a scanner would), but also *z* (depth) space.

But we also would like to entertain the opposite possibility, that scanning would involve a higher working memory load. Specifically, drilling enables a pathologist to focus on a single well-defined region at a time: examining a single region of interest in great depth and detail, and then iteratively returning to low magnification and examining the next region. In this manner, the searcher need only remember which salient region(s) they have or have not already "drilled into," which involves monitoring and updating only a representation of salient regions in the low magnification space. In contrast, a scanner moves across very broad sections of an image at moderate magnification, requiring them to constantly monitor and update their past and current position relative to the entire image space at moderate magnification, when only small portions of the overall image can be seen at a time. This might incur a high working memory load given the continuous locational monitoring necessary to sustain a strategic search.

To our knowledge no research has directly examined these possibilities in the context of visual search. Surprisingly, one recent study demonstrates that people searching visual images actually have very poor memory for where they have (and have not) visited (Vo et al., 2016; see also Horowitz & Wolfe, 1998). This result suggests that the memory load associated with location monitoring and updating is quite low regardless of strategy (i.e., drilling versus scanning). Thus, while there is no definitive answer to this question, we have included a discussion of this intriguing future research direction in the discussion section.

Pp 8, line 30-31. The authors say that "… may explain why our data did not suggest any influence of visual search strategy on diagnostic accuracy". Where exactly in the Results section was this discussed? Please clarify.

Response: This point might have created confusion due to the poor choice of words. The term "visual search strategy" referred to the scanning percentage variable that quantifies the drilling/scanning behavior and in Table 2, we showed that there was no correlation between the scanning percentage and the diagnostic concordance.

Still in the Discussion, but not on a specific page. Other researchers have investigated digital slide exploration in Whole Slide Imaging (for example, see 1. Krupinski EA, Tillack AA, Richter L, et al. Eye-movement study and human performance using telepathology virtual slides: implications for medical education and differences with expertise. Hum Pathol. 2006;37(12):1543-1556. 2. Treanor D, Lim CH, Magee D, Bulpitt A, Quirke P. Tracking with virtual slides: a tool to study diagnostic error in histopathology. Histopathology. 2009; 55(1):37-45. 3. Mello-Thoms C, Mello CAB, Medvedeva O, et al. Perceptual analysis of the reading of dermatopathology slides by pathology residents. Arch Path Lab Med 2012; 136:551-562), maybe not necessarily in breast slides - although

Krupinski's study was in breast - and they have come up with several conclusions linking search strategy with diagnostic accuracy and pathologist experience level. As such, it is surprising that no such link (to diagnostic accuracy) was found in this study, but I would encourage the authors to explore what has been found in the literature and at least hypothesize why they believe their study did not observe these relationships. Was it because mostly they used experienced pathologists, while the previous studies focused on the differences between experts and trainees? Was it the type of images used? I believe that it would be important to situate the authors' findings in the context of what others have reported, and pose some explanations for the differences in the results.

Response: This is a very good point. We added a paragraph discussing the findings of these studies. Our study and analysis are significantly different from the studies in the literature. The most relevant one by Mello-Thoms et al. was not in breast slides but the discussion of the different findings certainly adds to our paper.

Conclusions:

Fine as written.

Figures and Tables:

Fine.

Reviewer #2: This well-written paper describes the behavioural data gathered from a large sample of pathologists (87) as they view, zoom, pan and interpret breast biopsy slides. The aims of the study are sound - the introduction clearly and concisely outlines recent work on 'drillers' and 'scanners' that was derived from eye-tracking studies using volumetric medical images (e.g., Drew et al., 2013). Although this is not an eye-tracking study and therefore the precise visual attention components of the observers remains unknown, the paper nonetheless documents the different scanning/navigation strategies of observers and a comprehensive analysis examines what factors relate to diagnostic accuracy (in this case, agreement with an expert panel). Overall I think this is a very intriguing and well thought out study that makes a useful contribution to this growing field.

I only have a few issues with this paper, many of which can be easily resolved.

1. Sample selection criteria

First off, I applaud the authors in recruiting so many pathologists to take part in their study. Low sample size is a common problem with medical image perception research and so it is important to find that larger scale studies are underway.

Response: Thank you very much.

One minor point is that in the paper (p3 line 9) the authors indicate that over 200 participants were invited and yet only much later (p4 line 35) does the reader find out that only 87 participants data were analyzed. What remains unclear is whether this 87 are the only participants out of the 200+ sample that actually took part, or whether more than 87 took part but there were some screening or data issues that meant some participants were excluded. This could just be clarified briefly on p3 in the participant data section. Also in the tracking data analysis section (p3 line 42) the authors indicate that trials were excluded if slides were inactive for more than 5 minutes - how many trials were excluded because of this?

Response: This study was part of a larger project exploring the diagnostic concordance on digital and glass slides in breast pathology. To ensure a blind review, we excluded the references to our previous work. From the 200

pathologists invited, half of them were assigned glass slides, and a small portion of them failed to complete the study. The subset we used in this study, namely the 87-pathologist cohort, was assigned to interpret digital slides in the second phase of the study. We added an explanation and related references to the results section.

907 entries were removed because of the time limit. We only removed the entries exceeding 1-minute (not 5 minutes, this was a typo), but used the rest of the entries from the same interpretation session. We included this information in the results section.

2. Unclear rationale for why only half of participants annotated with ROI

On p3 line 22-25 the authors indicate that "Half of the participants were then asked to electronically annotate the digital WSI with an ROI supporting the highest order (most severe) diagnosis". It remains unclear as to the rationale as to why did not all participants do this. Please clarify. What is gained by having only half of participants do this second localization part of the task? The ROI analysis is in the Table 1 analysis, but its non-sig results are not discussed or explained. What did you expect to find with this deliberate manipulation? For example, the way I understand it (based on p3 line 18-20) your participants had to select one or more diagnostic categories from the list of 14 provided in supplement table 1, but only half of participants had to actually localize the target region - therefore there is a chance that the other half were correct (i.e., agree with the expert panel), but based on the decision of different visual features. This issue needs clarifying.

Response: Only half of the participants were asked to provide an ROI to control for the potential impact of drawing an ROI on the diagnostic decision-making process. One could argue that the task of picking a final ROI might affect the search pattern and cognitive process during the interpretation. However, in our analysis, we did not find any correlation between the ROI selection task and diagnostic search pattern. In a previous study, we explored the relationship between the ROI identification and diagnostic accuracy (Nagarkar et al., Region of interest identification and diagnostic agreement in breast pathology, *Modern Pathology*, 2016). This study found a correlation between the overlap of participant's ROI to the expert consensus ROI and the diagnostic concordance.

3. Unclear criteria as to what precisely counts as a driller or a scanner

The main analysis is always based on the binned categories of scanning percentages (0-20%, 20-40%, etc) that are defined p3 line 57 onwards (note that these 5 binned categories are nicely displayed with n values in Figure 3, yet on p4 line 9 you say there are only 4 categories - it seems you have missed out the 60-80% here?). However, throughout the paper the authors discuss that 'some participants are drillers and some participants are scanners', and then proceed to describe the behavioural characteristics of these two groups (e.g., discussion section p8 line 35 - "some participants, namely drillers…", p8 line 37 - "participants at the other end of the spectrum, namely scanners…"). However, what is never actually made clear in the paper is what precisely makes a pathologist count as a driller or a scanner. How much scanning % is required to be a 'scanner' instead of a driller? Across all images and participants the average scanning rate is around 75%, so what proportion of your 87 pathologists would be considered scanners and what proportion would be considered drillers? Fig 1 is a visual representation of what a driller and a scanner apparently looks like. But on what basis were these 2 pictures pulled out from the 5000+ and a decision made as to which is which. E.g. based on the authors qualitative/subjective decision that the one on the left is a scanner and that the one on the right is a driller? Or based on some precise ratio of scanning / zooming / panning? According to p4 line 36, it does appear that these categorizations were made based on subjective decisions: "Tracking logs were visualized and analysed to summarize the interpretative strategy of each pathologist". If the scanner/driller categorizations were made based on specific zoom/pan criteria you would not have had to visualize it beforehand. The upshot of all this is that it would be worth running a kappa analysis to establish whether other observers can reliably differentiate scanners from drillers.

The authors acknowledge in the discussion p7 line 39 "As scanning and drilling are complementary strategies, participants used a combination of both, but in different ratios. We wanted to explore potential explanations for the two interpretative strategies through their correlation with diagnostic accuracy and efficiency, as well as determining if these patterns change over time".

A key part of the rationale you put forward was that scanning and drilling strategies could have differential effects on diagnostic accuracy, but if the definition of these categories is unreliable then you can see the problem this causes. E.g. your conclusion p9 line 9-10 states "we identified two distinct interpretive strategies….". Strictly speaking this is not the case as all your analysis and shown data is based on the 5 binned categories (0-20%, 20-40% etc), not the distribution of 'driller' performance versus 'scanner' performance'. For that conclusion to hold you would have to justify and explain how scanners and drillers are categorized / analysed. Another one from the conclusion is p9 line 19 "yet scanners and drillers had similar levels of accuracy compared to the consensus reference diagnoses". What specific analysis are you drawing this from? I can see nothing in any of the tables or analysis that specifically shows how accurate scanners are compared to drillers, nor how these two relate to your expert panel.

Response: Scanning and drilling are terms used to describe two semi-distinct strategies for searching an image. Participating pathologists tended to use a combination of both drilling and scanning, using one or the other strategy on different images. To our knowledge, only a qualitative technique has been used to parse search patterns into scanning versus drilling, involving the plotting and visual inspection of search behavior. This method tends to correlate with quantitative techniques derived from eye tracking data (Drew et al., 2013). Because we did not use an eye tracker, rather than using the qualitative approach we devised a continuous measure that assesses the percent of image review behavior indicative of scanning. We believe this is a more powerful approach in that it eliminates the subjectivity of other techniques, more accurately describes the continuous nature of the drilling-scanning continuum, affords regression-based analyses rather than between-groups comparisons, and allows differences in search behaviors to be seen both within and across pathologists. Because we are not attempting to parse pathologists into driller versus scanner groups, we have revised the test to make it clearer that we conceptualize this distinction as bipolar and continuous. Our figures are intended only to depict the two polar ends of this continuum. We have made these points clearer in the revised manuscript, and apologize for any confusion.

4. Over interpretation of non-sig findings relating to confidence

In the discussion p7 line 50+ the authors indicate that there was no statistically significant difference between scanning percentages and confidence ratings. Nevertheless, in the next sentence the authors then continue to interpret these findings as if they are significant, and link higher scanning strategies to personality subtypes. This does not make sense. There were no statistical differences in scanning % as a function of confidence ratings. I can see the descriptive trend, but it is important to not conflate non-significant descriptive trends with statistically significant findings.

Likewise in the conclusion p9 line 11/12, the authors summarise their significant effects of gender and age on scanning (but not facility size?), and then once again treat the non-significant findings of confidence rating as if they were significant "those who reported less confidence in their interpretation of breast tissue tended to spend more time scanning". This last statement is misleading as it is not statistically accurate according to the authors own analysis. Moreover, having these misleading trend statements can set a dangerous precedent, as the reader may focus on the conclusion remarks when summarising and citing your work own, and therefore incorporate and perpetuate this misinformation.

Response: We apologize for our over-interpretation of the data. We made corrections to the text to prevent any misunderstanding and revised the manuscript to better reflect the statistical outcomes.

Signed Review

Damien Litchfield

UNIVERSITY of WASHINGTON

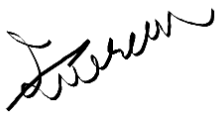
March, 13 2017

RE: Characterizing Diagnostic Search Patterns in Digital Breast Pathology: Scanners and Drillers

To Whom It May Concern,

Please see individual conflict of interest declarations from authors of the study attached. The authors declare that they have no conflict of interest.

Sincerely,



Ezgi Mercan

Paul G. Allen School of Computer Science and Engineering
University of Washington

Springer

Disclosure of potential conflicts of interest

Authors must disclose all relationships or interests that could have direct or potential influence or impart bias on the work. Although an author may not feel there is any conflict, disclosure of all relationships and interests provides a more complete and transparent process, leading to an accurate and objective assessment of the work. Awareness of a real or perceived conflicts of interest is a perspective to which the readers are entitled. This is not meant to imply that a financial relationship with an organization that sponsored the research or compensation received for consultancy work is inappropriate. For examples of potential conflicts of interests *that are directly or indirectly related to the research please* visit:
springer.com/gp/authors-editors/journal-author/journal-author-helpdesk/before-you-start

All authors of papers submitted to Journal of Digital Imaging (name of journal) must complete this form and disclose any real or perceived conflict of interest.

Please complete one form per author. The corresponding author collects the conflict of interest disclosure forms from all authors. The corresponding author will include a summary statement in the text of the manuscript in a separate section before the reference list, that reflects what is recorded in the potential conflict of interest disclosure form(s). Please note that you cannot save the form once completed. Kindly print upon completion, sign, and scan to keep a copy for your files.

The corresponding author should be prepared to send potential conflict of interest disclosure form if requested during peer review or after publication on behalf of all authors (if applicable).

☑ I have no potential conflict of interest.

Category of disclosure	Description of Interest/Arrangement

Article title Characterizing Diagnostic Search Patterns in Digital Breast Pathology: Scanners and Drillers

Manuscript No. (if you know it) JDIM-D-17-00032

Author name Ezgi Mercan

Are you the corresponding author? ■ Yes ☐ No

Herewith I confirm that the information provided is accurate.

Author signature  Date 3/6/2017

Springer

Disclosure of potential conflicts of interest

Authors must disclose all relationships or interests that could have direct or potential influence or impart bias on the work. Although an author may not feel there is any conflict, disclosure of all relationships and interests provides a more complete and transparent process, leading to an accurate and objective assessment of the work. Awareness of a real or perceived conflicts of interest is a perspective to which the readers are entitled. This is not meant to imply that a financial relationship with an organization that sponsored the research or compensation received for consultancy work is inappropriate. For examples of potential conflicts of interests *that are directly or indirectly related to the research please* visit:
springer.com/gp/authors-editors/journal-author/journal-author-helpdesk/before-you-start

All authors of papers submitted to Journal of Digital Imaging
(name of journal) must complete this form and disclose any real or perceived conflict of interest.

Please complete one form per author. The corresponding author collects the conflict of interest disclosure forms from all authors. The corresponding author will include a summary statement in the text of the manuscript in a separate section before the reference list, that reflects what is recorded in the potential conflict of interest disclosure form(s). Please note that you cannot save the form once completed. Kindly print upon completion, sign, and scan to keep a copy for your files.

The corresponding author should be prepared to send potential conflict of interest disclosure form if requested during peer review or after publication on behalf of all authors (if applicable).

☑ I have no potential conflict of interest.

Category of disclosure	Description of Interest/Arrangement

Article title Characterizing Diagnostic Search Patterns in Digital Breast Pathology: Scanners and Drillers

Manuscript No. (if you know it) JDIM-D-17-00032

Author name Linda Shapiro

Are you the corresponding author? ☐ Yes ☒ No

Herewith I confirm that the information provided is accurate.

Author signature Linda Shapiro Date 3/6/2017

Springer

Disclosure of potential conflicts of interest

Authors must disclose all relationships or interests that could have direct or potential influence or impart bias on the work. Although an author may not feel there is any conflict, disclosure of all relationships and interests provides a more complete and transparent process, leading to an accurate and objective assessment of the work. Awareness of a real or perceived conflicts of interest is a perspective to which the readers are entitled. This is not meant to imply that a financial relationship with an organization that sponsored the research or compensation received for consultancy work is inappropriate. For examples of potential conflicts of interests *that are directly or indirectly related to the research please* visit:
springer.com/gp/authors-editors/journal-author/journal-author-helpdesk/before-you-start

All authors of papers submitted to Journal of Digital Imaging
(name of journal) must complete this form and disclose any real or perceived conflict of interest.

Please complete one form per author. The corresponding author collects the conflict of interest disclosure forms from all authors. The corresponding author will include a summary statement in the text of the manuscript in a separate section before the reference list, that reflects what is recorded in the potential conflict of interest disclosure form(s). Please note that you cannot save the form once completed. Kindly print upon completion, sign, and scan to keep a copy for your files.

The corresponding author should be prepared to send potential conflict of interest disclosure form if requested during peer review or after publication on behalf of all authors (if applicable).

☑ I have no potential conflict of interest.

Category of disclosure	Description of Interest/Arrangement

Article title Characterizing Diagnostic Search Patterns in Digital Breast Pathology: Scanners and Drillers

Manuscript No. (if you know it) JDIM-D-17-00032

Author name Tad Brunye

Are you the corresponding author? ☐ Yes ☒ No

Herewith I confirm that the information provided is accurate.

Author signature _____ Date 3/6/2017

Springer

Disclosure of potential conflicts of interest

Authors must disclose all relationships or interests that could have direct or potential influence or impart bias on the work. Although an author may not feel there is any conflict, disclosure of all relationships and interests provides a more complete and transparent process, leading to an accurate and objective assessment of the work. Awareness of a real or perceived conflicts of interest is a perspective to which the readers are entitled. This is not meant to imply that a financial relationship with an organization that sponsored the research or compensation received for consultancy work is inappropriate. For examples of potential conflicts of interests *that are directly or indirectly related to the research please* visit:
springer.com/gp/authors-editors/journal-author/journal-author-helpdesk/before-you-start

All authors of papers submitted to Journal of Digital Imaging
(name of journal) must complete this form and disclose any real or perceived conflict of interest.

Please complete one form per author. The corresponding author collects the conflict of interest disclosure forms from all authors. The corresponding author will include a summary statement in the text of the manuscript in a separate section before the reference list, that reflects what is recorded in the potential conflict of interest disclosure form(s). Please note that you cannot save the form once completed. Kindly print upon completion, sign, and scan to keep a copy for your files.

The corresponding author should be prepared to send potential conflict of interest disclosure form if requested during peer review or after publication on behalf of all authors (if applicable).

☑ I have no potential conflict of interest.

Category of disclosure	Description of Interest/Arrangement

Article title Characterizing Diagnostic Search Patterns in Digital Breast Pathology: Scanners and Drillers

Manuscript No. (if you know it) JDIM-D-17-00032

Author name Donald L. Weaver, MD

Are you the corresponding author? ☐ Yes ☒ No

Herewith I confirm that the information provided is accurate.

Author signature [signature] Date 3/6/2017

Springer

Disclosure of potential conflicts of interest

Authors must disclose all relationships or interests that could have direct or potential influence or impart bias on the work. Although an author may not feel there is any conflict, disclosure of all relationships and interests provides a more complete and transparent process, leading to an accurate and objective assessment of the work. Awareness of a real or perceived conflicts of interest is a perspective to which the readers are entitled. This is not meant to imply that a financial relationship with an organization that sponsored the research or compensation received for consultancy work is inappropriate. For examples of potential conflicts of interests *that are directly or indirectly related to the research please* visit:
springer.com/gp/authors-editors/journal-author/journal-author-helpdesk/before-you-start

All authors of papers submitted to Journal of Digital Imaging
(name of journal) must complete this form and disclose any real or perceived conflict of interest.

Please complete one form per author. The corresponding author collects the conflict of interest disclosure forms from all authors. The corresponding author will include a summary statement in the text of the manuscript in a separate section before the reference list, that reflects what is recorded in the potential conflict of interest disclosure form(s). Please note that you cannot save the form once completed. Kindly print upon completion, sign, and scan to keep a copy for your files.

The corresponding author should be prepared to send potential conflict of interest disclosure form if requested during peer review or after publication on behalf of all authors (if applicable).

☑ I have no potential conflict of interest.

Category of disclosure	Description of Interest/Arrangement

Article title Characterizing Diagnostic Search Patterns in Digital Breast Pathology: Scanners and Drillers

Manuscript No. (if you know it) JDIM-D-17-00032

Author name J Elmore

Are you the corresponding author? ☐ Yes ■ No

Herewith I confirm that the information provided is accurate.

Author signature _____ Date 3/6/2017

Characterizing Diagnostic Search Patterns in Digital Breast Pathology: Scanners and Drillers

Ezgi Mercan[1], Linda G. Shapiro[1], Tad T. Brunyé[2], Donald L. Weaver[3], Joann G. Elmore[4]
[1] Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA
[2] Department of Psychology, Tufts University, Medford, MA, USA
[3] Department of Pathology and UVM Cancer Center, University of Vermont, Burlington, VT, USA
[4] Department of Medicine, University of Washington School of Medicine, Seattle, WA, USA

Correspondence to:
Ezgi Mercan
Paul G. Allen Center for Computing
185 Stevens Way
Seattle, WA 98195
Email: ezgi@cs.washington.edu
Phone: 206-953-1711

Keywords: diagnosis, interpretation, breast cancer, whole slide imaging, pathology, biopsy

Conflict of Interest: The authors declare that they have no conflict of interest.