



Published in final edited form as:

J Biomed Inform. 2017 February ; 66: 171–179. doi:10.1016/j.jbi.2017.01.004.

Accuracy is in the Eyes of the Pathologist: The Visual Interpretive Process and Diagnostic Accuracy with Digital Whole Slide Images

Tad T. Brunyé, PhD¹, Ezgi Mercan, PhD², Donald L. Weaver, MD³, and Joann G. Elmore, MD, MPH⁴

¹Center for Applied Brain & Cognitive Sciences, Tufts University, Medford, MA

²Department of Computer Science and Engineering, University of Washington, Seattle, WA

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

⁴Department of Medicine, University of Washington School of Medicine, Seattle, WA

Abstract

Digital whole slide imaging is an increasingly common medium in pathology, with application to education, telemedicine, and rendering second opinions. It has also made it possible to use eye tracking devices to explore the dynamic visual inspection and interpretation of histopathological features of tissue while pathologists review cases. Using whole slide images, the present study examined how a pathologist's diagnosis is influenced by fixed case-level factors, their prior clinical experience, and their patterns of visual inspection. Participating pathologists interpreted one of two test sets, each containing 12 digital whole slide images of breast biopsy specimens. Cases represented four diagnostic categories as determined via expert consensus: benign without atypia, atypia, ductal carcinoma in situ (DCIS), and invasive cancer. Each case included one or more regions of interest (ROIs) previously determined as of critical diagnostic importance. During pathologist interpretation we tracked eye movements, viewer tool behavior (zooming, panning), and interpretation time. Models were built using logistic and linear regression with generalized estimating equations, testing whether variables at the level of the pathologists, cases, and visual interpretive behavior would independently and/or interactively predict diagnostic accuracy and efficiency. Diagnostic accuracy varied as a function of case consensus diagnosis, replicating earlier research. As would be expected, benign cases tended to elicit false positives, and atypia, DCIS, and invasive cases tended to elicit false negatives. Pathologist experience levels, case consensus diagnosis, case difficulty, eye fixation durations, and the extent to which pathologists' eyes fixated within versus outside of diagnostic ROIs, all independently or interactively predicted diagnostic accuracy. Higher zooming behavior predicted a tendency to over-interpret benign and atypia cases,

Correspondence to: Tad T. Brunyé, Ph.D., Center for Applied Brain & Cognitive Sciences, Tufts University, 200 Boston Ave., Suite 3000, Medford, MA 02155, Telephone: (617) 306-6262, Fax: (617) 627-3181, tbruny01@tufts.edu.

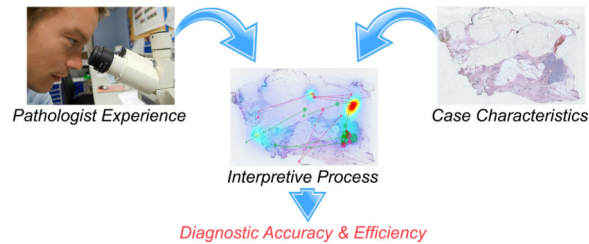
Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Disclosure/Duality of Interest

The authors have no conflict of interests to disclose.

but not DCIS cases. Efficiency was not predicted by pathologist- or visual search-level variables. Results provide new insights into the medical interpretive process and demonstrate the complex interactions between pathologists and cases that guide diagnostic decision-making. Implications for training, clinical practice, and computer-aided decision aids are considered.

Graphical Abstract



Keywords

pathology; breast cancer; decision making; whole slide imaging; cognitive informatics

1. Introduction

Each year millions of women rely upon a pathologist's visual inspection, interpretation, and diagnosis of tissue from their breast biopsies. Accurately diagnosing biopsies on the continuum from benign to invasive cancer requires not only specialized knowledge but also identifying critical regions within the histopathologic material and appropriately classifying perceived visual features by candidate histological diagnoses^{1,2}. This process is exceedingly dynamic and complex; it is also vulnerable to misidentification, misinterpretation, and ultimately diagnostic errors in the form of false positive or false negative interpretations. Because diagnostic errors can contribute to over- or under-treatment, with significant impact on patients, it is important to understand the conditions under which diagnostic errors emerge and how they might be prevented. Information processing frameworks propose that variation in diagnostic accuracy results from the interaction of characteristics of the case, physician experience levels, and the visual search process characterizing interpretation³. This notion provides the conceptual basis for the present study examining the independent and interactive influence of these three elements in predicting the accuracy and efficiency of histological diagnosis. To do so, we tracked the eye movements and image review behaviors employed by pathologists examining whole slide images (WSI) of breast biopsy specimens as they rendered diagnoses. We then evaluated the characteristics of the case, the physician and the visual search process that are associated with accuracy and efficiency.

1.2 Factors influencing diagnostic accuracy

Variation in case-level parameters may be associated with diagnostic accuracy, though few studies have examined this possibility. Several case-level dimensions of breast biopsy specimens include the diagnostic category of tissue (defined by a consensus of experts here), the density of the woman's breast tissue (defined by density reported on mammography exam preceding the biopsy)⁴, and the inherent difficulty of interpreting certain visual

features in the image (defined by a larger group of practicing pathologists who rated the level of difficulty of each test case). In general, breast biopsy specimens can be categorized by diagnosis along a continuum from benign (normal) to invasive carcinoma. For the purposes of this research, we parse diagnoses into the following four categories: benign, atypia, ductal carcinoma in situ (DCIS), and invasive. Each category is characterized by unique cellular and architectural features and extent of disease. Benign indicates normal or hyperplastic cells and architecture, atypia indicates an increased number of cells with abnormal features that are beginning to adopt an abnormal architecture, DCIS indicates abnormal neoplastic cells and architecture confined within the walls of a duct, and invasive indicates that clusters of abnormal malignant cells have invaded beyond the walls of a duct.

In clinical practice, the diagnostic category assigned to a breast specimen carries implications for subsequent monitoring, treatment, quality of life, and prognosis. These diagnostic categories also differentially influence diagnostic accuracy, with atypia and DCIS eliciting lower overall accuracy relative to the benign or invasive cases at the ends of the spectrum⁵. Breast density has been associated with diagnostic accuracy of radiologists when interpreting mammograms^{6,7}, and a relationship between mammographic density and the accuracy of the pathologists has been reported⁵, but not thoroughly examined. Furthermore, whether due to the presence of “borderline” features or complex architecture, cases vary in perceived difficulty², and normative case difficulty ratings may be associated with diagnostic accuracy.

Digital imaging technology is revolutionizing the practice of medicine^{8,9}. The advent of WSI in pathology has removed traditional barriers to studying glass slide interpretation, allowing researchers to understand not only physician magnification and panning behavior but also unobtrusively monitor eye movements to gain high fidelity information regarding the visual interpretive process. Several insights have resulted from this advancement. Tiersma and colleagues found that pathologists reviewing WSI tend to fixate their eyes more frequently on image regions containing relevant (versus irrelevant) diagnostic features, called regions of interest (ROI)¹⁰. More recent research has demonstrated that not only are there salient image regions attracting the visual attention of pathologists¹¹, but pathologist experience levels modulate whether these viewed regions are correctly interpreted and used to make a correct diagnosis. Specifically, trainees (pathology residents) show an early focus on visually salient but diagnostically irrelevant image regions whereas experienced pathologists tend to focus predominantly on regions with diagnostic relevance, and these patterns can in some cases predict diagnostic accuracy¹². Similar experience-related distinctions regarding the accuracy and efficiency of medical interpretation have been found in radiology^{13–17}.

How expertise arises among physicians likely parallels visual expertise in other domains, such as seen with luggage screeners and professional athletes^{18,19}. A number of studies have examined the perceptual and cognitive processes involved in expert visual diagnosis, including seminal work by Norman²⁰ and Ericsson²¹. There are two highly relevant outcomes from this research. First, expertise in visual diagnosis arises from knowledge of exemplars gained through past experience²². The larger and more diverse an expert’s memory for exemplars, the faster they will be able to match a percept with a similar stored

exemplar and perform diagnostic classification²³. Second, expert medical practitioners are better able to flexibly reason, often in a conceptual manner, about the relationships between visual information and diagnostic alternatives²¹. In contrast, novices may rely on memorized biomedical rules that are relatively inflexible and not necessarily grounded in prior perceptual experiences^{24,25}. Thus, in pathology increasing experience interpreting biopsy images results in more finely tuned memory for exemplars, increasing the distinctiveness of exemplar categories, and the efficiency and flexibility with which critical image features are identified and mapped to candidate diagnoses^{26,27}. It is important to realize, however, that developing visual expertise may be faster and ultimately more successful for some individuals than others, suggesting that not all trainees are capable of becoming experts at interpreting and diagnosing images^{28,29}.

The visual search process describes the interactions between the pathologist and the case, considering the pathologist's zooming and panning patterns, and their allocation of visual attention toward image features. To measure the latter, eye tracking devices use infrared light to monitor the spatial and temporal distribution of eye fixations, momentary pauses of the eyes thought to reflect the overt allocation of visual attention over regions of a scene. Eye trackers leverage how we move our eyes to bring particular regions of a scene into high resolution focus, allowing us to perceive and understand fine details³⁰. Research has demonstrated important relationships between the visual search process and diagnostic accuracy. For instance, high diagnostic accuracy levels are reached when pathologists show fewer fixations overall, and less examination of diagnostically irrelevant regions³¹. High diagnostic accuracy is also associated with pathologists spending the majority of time fixating their eyes in regions of high diagnostic relevance^{12,32}. Overall, a more refined visual search process that prioritizes diagnostically relevant areas is related to higher diagnostic accuracy.

1.3 The present study

Extant research suggests that individual case parameters, pathologist experience, and the visual search process may influence diagnostic accuracy, though these studies are limited in a few regards. First, no single study has examined the independent and interactive influence of these elements on the accuracy and efficiency of diagnosis. In contrast, existing research tends to narrowly focus on one (e.g., image navigation patterns¹¹ or experience³²) or two factors at a time, limiting understandings of the full range and influence of pathologist, case, and visual search level factors. Second, most studies examining eye movements of pathologists use restricted sample sizes (e.g., 4 to 7 physicians). Third, most studies restrict zooming and panning behavior, preventing the type of image inspection that would be employed during routine clinical practice^{32,33}. The present study was designed to expand upon this earlier work by using a large sample of pathologists (N = 40), allowing pathologists to freely zoom and pan images during visual inspection, and examining a wider range of pathologist-level, case-level, and visual search processes that may predict diagnostic accuracy and efficiency.

2. Materials and Methods

2.1 Participating pathologists

Forty pathologists were recruited from five U.S. university medical centers, two on the west coast and three on the east coast. Pathologists varied in level of training (14 faculty, 26 residents) and years of experience interpreting breast pathology ($\bar{x} = 3.61$, $\text{min} < 1$, $\text{max} = 14.5$).

2.2 Test cases

Breast biopsy test cases were obtained from women ≥ 40 years of age, using one slide per case that best illustrated the diagnostic features. A set of 24 hematoxylin and eosin (H&E) stained breast specimens was chosen from a larger (240 case) test set^{5,34}. To develop the digital WSI, glass slides were scanned into high resolution digital TIFF format using an iScan Coreo Au scanner³⁵ at 40 \times magnification. Using a modified Delphi technique, each case had a single consensus diagnosis based on the agreement of three expert breast pathologists. These experts also marked one or more ROI on each case to indicate what they considered the “best example(s)” of the consensus diagnosis.

The 24 cases represented four consensus diagnostic categories, to include benign without atypia (4 non-proliferative and proliferative cases), atypia (8 cases), DCIS (8 cases), and invasive cancer (4 cases). Cases also varied in breast density reported on mammograms obtained before the breast biopsy (using BI-RADS⁴), and in standardized ratings of the extent of case difficulty (scale 1–6) based on data gathered from a larger sample of pathologists ($N = 115$). It is worth noting that variation in the number of ROIs across cases was not correlated with reference case difficulty. Furthermore, expert interpretations of these 24 cases were highly concordant (21/24 cases) between the digital and glass versions (for a more in depth examination of this issue, see³⁶).

2.3 WSI viewer and histology form

A custom web browser-based WSI viewer was developed using the Microsoft Silverlight platform (Figure 1a). Like clinical WSI viewers, the system displayed each image at 1 \times magnification and a standard resolution, and allowed images to be zoomed (1–60 \times) and panned while maintaining full resolution. As the pathologist reviewed the image, the software automatically logged panning and zooming behavior over time and output this information to a data file. Note that participating pathologists were always blind to concordance diagnosis and the expert-marked regions of interest (ROIs). Once the pathologist reached an interpretation for a case, a standardized histology form was used to record final diagnosis by selecting from four diagnostic categories (benign, atypia, DCIS, and invasive cancer).

2.4 Eye tracker

We used an unobtrusive, remote eye tracking device (RED-m; SensoMotoric Instruments, Boston, MA) mounted to the bottom of a 22” (NEC EA224WMI-BK IPS LED) flat screen monitor running at 1920 \times 1080 resolution. The eye tracker uses infrared lights and cameras to track eye gaze position at 60Hz with high angular accuracy (0.5 $^\circ$) and spatial resolution

(0.1°). A standard 9-point calibration process was repeated for each participating pathologist to achieve high angular accuracy before they began interpreting the test cases.

2.5 Data collection procedures

The eye tracker and computer monitor were transported to each data collection site where pathologists were scheduled to participate in approximately 45-minute sessions in private conference rooms or offices; thus, each participant interpreted images using the same computer monitor and eye tracker. Following consent, participants completed a brief demographic survey followed by eye tracker calibration performed by watching a dot move between nine successive points on the screen. Participants were then instructed to interpret cases as they would in clinical practice and were randomized to interpret test set A or B. Participants then viewed and interpreted one of the two sets of 12 cases, at full screen, in random order. After each case, they completed the histology form. At the end of the session, they were remunerated with a \$50USD gift card.

2.6 Data processing

All pathologist-level, case-level, and visual search measures are included in Table 1. Data regarding each pathologist included whether they were a resident or faculty member, their self-reported level of experience with digital whole slide imaging, and their years of experience with breast pathology. Data regarding each case included the consensus diagnosis, mammographic breast density, and standardized difficulty ratings as described in section 2.2.

Eye tracking data were output at 60Hz and included raw coordinate eye position data in Cartesian space. Using standard techniques³⁰, raw data were parsed into eye fixations, momentary (> 99ms) pauses of the eye within a predefined angular dispersion (max 100 pixels). Each eye fixation includes a screen location and duration in milliseconds (ms). Data from the WSI viewer, including zoom levels and panning locations, were merged with eye tracking data to produce a single data stream relating image locations and eye fixations. Figure 1b exemplifies eye fixation data overlaid onto a representative case (DCIS consensus diagnosis); each of the 37,645 points indicates a single eye fixation.

Several visual search variables were derived from integrated WSI viewer and eye tracking data. From the WSI viewer, we included several zoom metrics including mean zoom levels, peak zoom level, and standard deviation of zoom behavior. From the eye tracker, we calculated the total number of fixations during case review, and the average duration (in ms) of fixations. We also parsed eye fixations and durations into occurring within versus outside of expert defined ROIs, to include several additional measures: proportion of fixations falling within an ROI, proportion falling outside an ROI, duration of fixations falling within an ROI, duration of fixations falling outside an ROI, and proportion of fixations characterized by re-entering an ROI. Note that because eye blinks tend to vary with levels of effort during visual tasks³⁷, we also exported blink rate for analysis.

2.7 Data consolidation

Because variables at the level of pathologist and visual search behavior tend to show collinearity, dimension reduction was completed using principal components analyses with Varimax rotation. The results of these analyses are detailed in Table 2. At the level of pathologists, the analysis revealed a single component weighted primarily toward career level and years of experience; this component will be referred to as *pathologist experience*. At the level of visual search behavior, the analysis revealed four components. The first was weighted toward overall and ROI specific *fixation durations*. The second was weighted toward relative *fixations within versus outside of ROIs*. The third was weighted toward *zoom behavior*, and the fourth weighted toward the *number of fixations*. Factor scores were calculated for all components and saved as new predictor variables using the Bartlett method³⁸. Case-level variables remained untransformed given the lack of collinearity.

Three outcome variables were considered. First, diagnostic accuracy was assessed by comparing participants' diagnoses to the expert defined consensus diagnosis for each case; concordant diagnoses were coded as a 1 and discordant diagnoses were coded as a 0. Second, we coded over-called (false positive) cases as 1, and under-called (false negative) cases as -1; concordant responses were coded as 0. These codes were averaged to calculate overall over- versus under-calling rates for each diagnostic category. Third, we also considered review time (in sec) as a measure of interpretive efficiency, which was only calculated for accurate (concordant) diagnoses.

2.8 Data analysis

Data analysis proceeded in two phases. First, to assess overall outcomes as a function of consensus diagnosis we conducted a repeated-measures analysis of variance (ANOVA) with four levels of our independent variable (Consensus Diagnosis: Benign, Atypia, DCIS, Invasive). Our intent was to replicate earlier studies demonstrating diagnostic concordance differences as a function of consensus diagnosis, and examine how diagnostic category may additionally influence diagnostic efficiency.

Second, to assess how pathologist-level, case-level, and visual search behavior variables predict diagnostic outcomes, we modeled our data using repeated-measures regressions, implementing the generalized estimating equation (GEE) approach^{39,40}. Each model used two categorical factors (Case Diagnostic Category, Case Mammographic Density) and six continuous covariates (Case Reference Difficulty, Pathologist Experience, Fixation Durations, Fixations in versus outside ROIs, Zoom Behavior, and Number of Fixations). All analyses used forward model selection to minimize quasi-likelihood under the independence model criterion (QIC) and identify 2-way interactions contributing to model fit. Due to unequal representation of the four breast density categories (and model failure due to matrix singularity), we collapsed density into two categories (low, high). For accuracy data, the model used a binary logistic outcome. For over- versus under-calling data, the model used an ordinal logistic outcome. For efficiency, the model used a linear outcome.

3. Results

3.1 Overall diagnostic accuracy & efficiency

Overall diagnostic concordance with consensus, rates of interpretation above and below consensus diagnoses, tendency to over- versus under-call diagnoses, and efficiency data are detailed in Table 3.

Diagnostic concordance varied as a function of diagnostic category. A repeated-measures analysis of variance (ANOVA) demonstrated significant accuracy differences across the four diagnostic categories, $F(3, 117) = 34.77, p < .001, \eta^2 = .47$. Paired samples t-tests demonstrated significant differences ($p < .01$) between all pairwise comparisons.

The rate of over-calling (false positive) versus under-calling (false negative) also varied as a function of diagnostic category, $F(3, 117) = 24.25, p < .001, \eta^2 = .38$. Paired samples t-tests showed over-calling in the benign condition relative to the three other conditions ($p < .001$), and higher under-calling with atypia relative to invasive cancer ($p = .018$). All other comparisons were non-significant ($p > .05$). We also compared the rate of over-calling versus under-calling to 0 (zero) within each category, by conducting a series of one-sample t-tests; all four diagnostic categories showed significant variation from 0 ($p < .03$).

For efficiency, the amount of time reviewing a case varied as a function of diagnostic category, $F(3, 117) = 5.47, p < .001, \eta^2 = .12$. Paired samples t-tests demonstrated significantly (p 's $< .02$) faster review times with benign cases relative to the other three categories. All other comparisons were non-significant ($p > .05$).

3.2 Predicting diagnostic accuracy

The best fitting model (QIC = 496.5) for accuracy included main effects and two two-way interactions, as detailed in Table 4. Significant effects were as follows. First, case consensus diagnostic category predicted accuracy (as seen previously in Table 3). Second, case breast density was related to accuracy, with higher accuracy at higher density levels. Third, more difficult cases predicted lower accuracy. Finally, longer fixation durations and more fixations within versus outside the boundaries of expert defined ROIs, were both related to higher accuracy.

Pathologist experience predicted higher accuracy, but this effect was qualified by an interaction between pathologist experience and diagnostic category. Higher pathologist experience levels predicted higher accuracy, but this effect was most pronounced with Atypia ($\chi^2 = 13.94, p < .001$) and DCIS ($\chi^2 = 8.21, p < .001$) cases.

3.3 Predicting over- and under-calling

The final model for over- versus under-calling included main effects and three two-way interactions, as detailed in Table 5. Significant effects were as follows. First, pathologist experience predicted higher over- versus under-calling rates, but this effect was qualified by an interaction between pathologist experience and fixations within versus outside ROIs. A median split by pathologist experience level showed a cross-over interaction. For lower pathologist experience levels, fixating within versus outside ROIs negatively predicted over-

versus under-calling diagnoses ($\chi^2 = .86, p > .05$); in contrast, for higher experience levels, this relationship was positive-going ($\chi^2 = .44, p > .05$). In other words, increasing levels of pathologist experience predicted over-calling when the ROI was fixated. Neither of these individual patterns, however, reached significance.

Second, case consensus diagnosis negatively predicted over- versus under-calling rates (as seen previously in Table 3), and zoom behavior positively predicted over- versus under-calling. These effects, however, were qualified by an interaction between case consensus diagnosis and zoom behavior. Parsing by diagnostic category, higher zoom behavior predicted over-calling but this effect was most pronounced with atypia ($\chi^2 = 21.84, p < .001$) and DCIS ($\chi^2 = 14.88, p < .001$) cases. In other words, when viewing atypia or DCIS cases, more zooming behavior predicted over-calling case diagnoses.

Third, there was an interaction between case difficulty rating and fixations within versus outside of ROIs. A median split by case difficulty ratings showed that with lower difficulty cases, fixations within versus outside of ROIs negatively predicted over- versus under-calling ($\chi^2 = 12.22, p < .001$); with higher difficulty cases, the relationship was non-significant ($\chi^2 = .45, p > .05$). In other words, lower difficulty cases tended to be under-called when there were more fixations within the ROI, and this pattern was not found with more difficult cases.

3.4 Predicting diagnostic efficiency

The best fitting model (QICC = 19302630) included all main and two-way interactions. There was only one marginal main effect of consensus ($\chi^2 = 6.79, p = .08$), suggesting longer review times with higher diagnostic categories (as seen previously in Table 3).

4. Discussion

This study was the first to explore the independent and interactive influences of pathologist-level, case-level, and visual search-level variables on the accuracy and efficiency of histological diagnosis. Replicating recent research, we found the lowest levels of concordance with expert defined consensus diagnoses when pathologists interpreted atypia and DCIS cases, relative to benign and invasive cancer cases^{5,41}. Discordance tended to be in the form of false negatives with atypia and DCIS; this pattern matches predictions made by visual search theory, which posits that targets with low prevalence rates tend to increase false negatives^{42,43}. Breast biopsy diagnoses more severe than benign without atypia are estimated to only occur in approximately 35% of cases, making them relatively low prevalence in daily practice^{44,45}. When a concordant diagnosis is reached, however, it is done most efficiently with Benign cases relative to the other diagnostic categories.

We found several main and interactive effects of pathologist-level, case-level, and visual search-level variables in predicting accuracy. First, pathologists with higher experience levels showed higher accuracy, but only when reviewing Atypia and DCIS cases. In other words, pathologist experience level is critical for the correct interpretation of diagnostic categories that tend to elicit more discordance⁵. The finely tuned exemplars of diagnostic features that develop with experience appear to be most critical for interpreting cases that

tend to show highest diagnostic variability^{5,46}. Second, pathologists who fixated more within rather than outside of ROIs showed higher overall accuracy. This finding supports earlier research demonstrating that spending less time in non-diagnostic regions³¹, and more time in diagnostic regions^{12,32}, increases accuracy, extending this finding from radiology to a dynamic (zooming, panning) diagnostic pathology task^{14,16,31,32,47}. Third, longer fixation durations predicted higher accuracy, supporting recent research showing longer fixation durations with expert versus novice surgeons⁴⁸. Finally, more difficult cases were related to lower accuracy, and higher density cases were related to higher accuracy. Note that the latter effect was novel and in an unexpected direction, which may be driven by an unequal representation of diagnostic categories in our two mammographic density levels.

We also found that pathologist-level, case-level, and visual search-level variables differentially guide over- versus under-calling within diagnostic categories. With Atypia and DCIS cases, pathologists were more likely to over-call a case when they also showed more zooming behavior. This pattern supports visual search theory suggesting that as observers repeatedly examine a case in detailed depth, the probability of an erroneous “guess” increases⁴⁹. Such erroneous interpretations likely result from a failure to find a target image features that precisely match histopathological features stored in memory. Finally, we also found evidence that when pathologists allocated more fixations to regions outside of critical ROIs during the inspection of cases with lower difficulty ratings, they tended to over-call the diagnosis. This pattern suggests that identifying diagnostically relevant ROIs is important not just for the positive identification of features, but also preventing over-diagnosis when interpreting cases that should not otherwise evoke difficulty on behalf of the pathologist.

4.1 Training and Clinical Implications

Modern training and accreditation programs require evidence that pathology residents demonstrate competence in examining and assessing surgical pathology specimens as part of the Milestones project⁵⁰. Toward this goal, medical educators are seeking methods for characterizing learners’ competence. Understanding overt (e.g., zooming, panning) and covert (e.g., fixation location, fixation duration) pathologist behavior can help identify methods and metrics for monitoring and evaluating the development of visual expertise in pathology. For instance, eye movement patterns can be used to provide feedback regarding the relative allocation of visual attention within versus outside pre-determined diagnostic ROIs⁵¹, allowing trainees and educators to review and learn from the visual interpretive process. In one recent advancement, nursing students were shown their own eye movement behaviors during a debriefing, resulting in improved understanding and learning about how they allocated visual attention during clinical practice⁵². The present results are also relevant to cognitive informatics research, which uses cognitive approaches (such as eye tracking) to better understand human information processing and interactions with computer systems^{53,54}. We demonstrate that image navigation behavior, coupled with eye tracking, can unobtrusively monitor physicians’ visual search and reveal important features of the visual search process and distinctions within and across individuals. As eye tracking systems become increasingly prevalent and flexible, and less obtrusive and expensive, they become more feasible for incorporation into classrooms and clinics^{52,55}. A continuing challenge to

this goal, however, is developing more robust, flexible, reliable, and clinically relevant algorithms for automated image processing and eye movement interpretation⁵⁶.

There are three primary clinical impacts of the present work. First, the extent to which pathologist expertise, eye movements, and zooming behavior were related to diagnosis tended to be most pronounced in Atypia and DCIS cases where the variability of pathology interpretations is relatively high^{5,57}. Misdiagnosis, particularly over-diagnosis (e.g., interpreting a biopsy as DCIS when it is Atypia), can lead to a cascade of unnecessary and costly surveillance, intervention, and treatment procedures^{46,58}. Obtaining formal second opinions (i.e., double reading) may prove advantageous in these circumstances⁵⁹, as may providing trainees with feedback about their image review behaviors that may be disadvantageous for ultimate diagnostic success. Second, while most prior work focuses on how single variables, such as pathologist clinical expertise or eye movement patterns, influence accuracy and efficiency, we demonstrate several important interactions that clinicians may find valuable to consider during practice. As the domain of cognitive informatics proceeds, it will find value in not only considering patterns of image navigation and eye movement behavior, but also how these patterns may vary reliably across individuals and cases. Furthermore, as telemedicine and second opinions become increasingly prevalent, it is important for the informatics domain to consider how distributed, team-based interpretation may impact the features attended to, diagnoses reached, and clinical outcomes^{53,60}.

4.2 Limitations

Though our study uses a larger sample size of 40 pathologists, includes more clinical cases, and details a broader range of variables predictive of accuracy and efficiency relative to prior studies, it also carries a few limitations. First, our testing situation was limited to reviewing a single slide, which contrasts with clinical practice wherein pathologists may review multiple slides per case, or request complementary tests. However, we do note that an expert panel was able reach consensus with these slides alone, and agreed that each slide contained adequate and representative material for rendering a primary diagnosis. It is also true that pathologists may routinely seek second opinions and have more clinical information available during interpretation than our experimental design allowed. Second, the proportion of cases representing each diagnostic category is unreflective of the distribution of cases in clinical practice⁴⁴. Indeed, our cases intentionally oversampled atypia and DCIS diagnoses. Violating pathologists' expectations relative to daily practice may influence interpretive behavior in yet unknown ways⁶¹. Though likely intractable for an eye tracking study of this nature, continuing research may benefit from integrating experimental cases into normal clinical practice. Finally, we note that self-reported level of experience with digital whole slide imaging is subjective, relies upon accurate memory of those experiences, and assumes the frequency of experience is positively related to the quality of experience. Though this variable was only minimally weighted in our Pathologist Experience component and thus did not likely contribute to overall data patterns, it is worth considering the inherent weaknesses of this type of subjective measure.

4.3 Conclusions

Overall, we find unique evidence that diagnostic accuracy and efficiency are influenced by variables at the levels of individual pathologists and cases, and the visual search process that relates the two over time. In some cases, these relationships proved interactive, demonstrating the importance of measuring and monitoring a broader range of pathologist behavior during biopsy interpretation.

Acknowledgments

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under award numbers R01 CA172343, R01 CA140560 and KO5 CA104699. The content is solely the responsibility of the authors and does not necessarily represent the views of the National Cancer Institute or the National Institutes of Health. The authors wish to thank Ventana Medical Systems, Inc., a member of the Roche Group, for use of iScan Coreo Au™ whole slide imaging system, and HD View SL for the source code used to build our digital viewer. For a full description of HD View SL please see <http://hdviewsl.codeplex.com/>.

References

1. Mello-Thoms C, Mello CAB, Medvedeva O, et al. Perceptual analysis of the reading of dermatopathology virtual slides by pathology residents. *Arch Pathol Lab Med*. 2012; 136:551–562. [PubMed: 22540304]
2. Crowley RS, Naus GJ, Stewart J III, et al. Development of visual diagnostic expertise in pathology: An information-processing study. *J Am Med Informatics Assoc*. 2003; 10:39–51.
3. Nodine, CF., Kundel, HL. The cognitive side of visual search in radiology. In: O'Regan, JK., Levy-Schoen, A., editors. *Eye Movements: From Psychology to Cognition*. Amsterdam, Netherlands: Elsevier Science; 1987. p. 572-582. Vol
4. D'orsi, CJ., Bassett, L., Berg, WA., et al. *Breast Imaging Reporting and Data System: ACR BI-RADS-Mammography*. Reston, VA: 2003.
5. Elmore JG, Longton GM, Carney PA, et al. Diagnostic Concordance Among Pathologists Interpreting Breast Biopsy Specimens. *JAMA*. 2015; 313:1122. [PubMed: 25781441]
6. Carney PA. Individual and Combined Effects of Age, Breast Density, and Hormone Replacement Therapy Use on the Accuracy of Screening Mammography. *Ann Intern Med*. 2003; 138:168. [PubMed: 12558355]
7. Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. *Radiology*. 1992; 184:613–617. [PubMed: 1509041]
8. May M. A better lens on disease. *Sci Am*. 2010; 302:74–77.
9. Weiser M. The computer for the 21st century. *Sci Am*. 1991; 265:94–104.
10. Tiersma ESM, Peters AA, Mooij HA, et al. Visualising scanning patterns of pathologists in the grading of cervical intraepithelial neoplasia. *J Clin Pathol*. 2003; 56:677–680. [PubMed: 12944551]
11. Roa-Peña L, Gómez F, Romero E. An experimental study of pathologist's navigation patterns in virtual microscopy. *Diagn Pathol*. 2010; 5:71. [PubMed: 21087502]
12. 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. [PubMed: 25084012]
13. Nodine CF, Kundel Hk, Sherri C Lauver, et al. Nature of Expertise in Searching Mammograms for breast masses. *Acad Radiol*. 1996; 3:1000–1006. [PubMed: 9017014]
14. Krupinski EA. Visual search of mammographic images: influence of lesion subtlety. *Acad Radiol*. 2005; 12:965–969. [PubMed: 16023379]
15. Lesgold, A., Rubinson, H., Feltovich, P., et al. *The Nature of Expertise*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988. Expertise in a complex skill: Diagnosing x-ray pictures; p. 311-342. Vol
16. Krupinski EA. Visual scanning patterns of radiologists searching mammograms. *Acad Radiol*. 1996; 3:137–144. [PubMed: 8796654]

17. Manning D, Ethell S, Donovan T, et al. How do radiologists do it? The influence of experience and training on searching for chest nodules. *Radiography*. 2006; 12:134–142.
18. McCarley JS, Kramer AF, Wickens CD, et al. Visual skills in airport-security screening. *Psychol Sci*. 2004; 15:302–306. [PubMed: 15102138]
19. Gegenfurtner A, Lehtinen E, Säljö R. Expertise Differences in the Comprehension of Visualizations: a Meta-Analysis of Eye-Tracking Research in Professional Domains. *Educ Psychol Rev*. 2011; 23:523–552.
20. Norman GR, Coblenz CL, Brooks L, et al. Expertise in Visual Diagnosis: A review of the Literature. *Acad Med*. 1992; 67:S78–S83. [PubMed: 1388563]
21. Ericsson KA, Lehmann AC. Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annu Rev Psychol*. 1996; 47:273–305. [PubMed: 15012483]
22. Medin DL, Altom MW, Murphy TD. Given versus induced category representations: Use of prototype and exemplar information in classification. *J Exp Psychol Learn Mem Cogn*. 1984; 10:333–352. [PubMed: 6235306]
23. Norman G. Research in clinical reasoning: Past history and current trends. *Med Educ*. 2005; 39:418–427. [PubMed: 15813765]
24. Boshuizen HPA, Schmidt HG. On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. *Cogn Sci*. 1992; 16:153–184.
25. Schmidt HG, Norman GR, Boshuizen HPA. A cognitive perspective on medical expertise: theory and implications. *Acad Med*. 1992; 65:611–621.
26. Hershler O, Hochstein S. The importance of being an expert: Top-down attentional control in visual search with photographs. *Attention, Perception, Psychophys*. 2009; 71:1478–1486.
27. Jaarsma T, Boshuizen HPA, Jarodzka H, et al. Tracks to a Medical Diagnosis: Expertise Differences in Visual Problem Solving. *Appl Cogn Psychol*. 2016; 30:314–322.
28. Joseph RM, Keehn B, Connolly C, et al. Why is visual search superior in autism spectrum disorder? *Dev Sci*. 2009; 12:1083–1096. [PubMed: 19840062]
29. Sobel KV, Gerrie MP, Poole BJ, et al. Individual differences in working memory capacity and visual search: The roles of top-down and bottom-up processing. *Psychon Bull Rev*. 2007; 14:840–845. [PubMed: 18087947]
30. Duchowski, AT. *Eye Tracking Methodology: Theory and Practice*. New York, NY: Springer-Verlag; 2007.
31. Krupinski EA, Graham AR, Weinstein RS. Characterizing the development of visual search experience in pathology residents viewing whole slide images. *Hum Pathol*. 2013; 44:357–364. [PubMed: 22835956]
32. 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. *Hum Pathol*. 2006; 37:1543–1556. [PubMed: 17129792]
33. Bombari D, Mora B, Schaefer SC, et al. What was I thinking? Eye-tracking experiments underscore the bias that architecture exerts on nuclear grading in prostate cancer. *PLoS One*. 2012; 7:e38023. [PubMed: 22666438]
34. Oster N, Carney PA, Allison KH, et al. Development of a diagnostic test set to assess agreement in breast pathology: Practical application of the Guidelines for Reporting Reliability and Agreement Studies (GRRAS). *BMS Womens Heal*. 2013; 13:3.
35. Ventana Medical Systems I. *iScan Coreo Au Product Page*. 2012
36. Elmore JG, Longton GM, Pepe MS, et al. A randomized study comparing digital imaging to traditional glass slide microscopy for breast biopsy and cancer diagnosis. *J Pathol Inform*. 2017
37. Brookings JB, Wilson GF, Swain CR. Psychophysiological responses to changes in workload during simulated air traffic control. *Biol Psychol*. 1996; 42:361–377. [PubMed: 8652753]
38. Hershberger, SL. *Encyclopedia of Statistics in the Behavioral Science*. Vol New York, NY: John Wiley; 2005. Factor scores; p. 636-644.
39. Burton P, Gurrin L, Sly P. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Stat Med*. 1998; 17:1261–1291. [PubMed: 9670414]

40. Hanley JA. Statistical analysis of correlated data using generalized estimating equations: An orientation. *Am J Epidemiol.* 2003; 157:364–375. [PubMed: 12578807]
41. Allison KH, Reisch LM, Carney PA, et al. Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel. *Histopathology.* 2014; 65:240–251. [PubMed: 24511905]
42. Wolfe JM, Horowitz TS, Van Wert MJ, et al. Low target prevalence is a stubborn source of errors in visual search tasks. *J Exp Psychol Gen.* 2007; 136:623–638. [PubMed: 17999575]
43. Van Wert MJ, Horowitz TS, Wolfe JM. Even in correctable search, some types of rare targets are frequently missed. *Attention, Perception, Psychophys.* 2009; 71:541–553.
44. Ballard-Barbash R, Taplin SH, Yankaskas BC, et al. Breast cancer surveillance consortium: A national mammography screening and outcomes database. *Am J Roentgenol.* 1997; 169:1001–1008. [PubMed: 9308451]
45. Ernster VL, Ballard-Barbash R, Barlow WE, et al. Detection of ductal carcinoma in situ in women undergoing screening mammography. *J Natl Cancer Inst.* 2002; 94:1546–1554. [PubMed: 12381707]
46. Bleyer A, Welch HG. Effect of three decades of screening mammography on breast-cancer incidence. *N Engl J Med.* 2012; 367:1998–2005. [PubMed: 23171096]
47. Kundel HL, Nodine CF, Krupinski EA, et al. Using gaze-tracking data and mixture distribution analysis to support a holistic model for the detection of cancers on mammograms. *Acad Radiol.* 2008; 15:881–886. [PubMed: 18572124]
48. Giovinco NA, Sutton SM, Miller JD, et al. A passing glance? Differences in eye tracking and gaze patterns between trainees and experts reading plain film bunion radiographs. *J Foot Ankle Surg.* 2015; 54:382–391. [PubMed: 25441848]
49. Chun MM, Wolfe JM. Just say no: How are visual searches terminated when there is no target present? *Cogn Psychol.* 1996; 30:39–78. [PubMed: 8635311]
50. Nasca TJ, Philibert I, Brigham T, et al. The Next GME Accreditation System — Rationale and Benefits. *N Engl J Med.* 2012; 366:1051–1056. [PubMed: 22356262]
51. Graber ML, Kissam S, Payne VL, et al. Cognitive interventions to reduce diagnostic error: a narrative review. *BMJ Qual Saf.* 2012; 21:535–557.
52. Henneman EA, Cunningham H, Fisher DL, et al. Eye tracking as a debriefing mechanism in the simulated setting improves patient safety practices. *Dimens Crit Care Nurs.* 2014; 33:129–135. [PubMed: 24704737]
53. Patel VL, Kannampallil TG. Cognitive informatics in biomedicine and healthcare. *J Biomed Inform.* 2015; 53:3–14. [PubMed: 25541081]
54. Kannampallil TG, Abraham J, Patel VL. Methodological framework for evaluating clinical processes: A cognitive informatics perspective. *J Biomed Inform.* 2016; 64:342–351. [PubMed: 27847328]
55. Henneman EA, Gawlinski A. Eye-Tracking Technology: An Innovative Approach for Teaching and Evaluating Patient Safety Behaviors. *Nurs Educ Perspect.* 2016; 37:356–357.
56. Mercan E, Aksoy S, Shapiro LG, et al. Localization of diagnostically relevant regions of interest in whole slide images: A comparative study. *J Digit Imaging.* 2016
57. Rosai J. Borderline epithelial lesions of the breast. *Am J Surg Pathol.* 1991; 15:209–221. [PubMed: 1847606]
58. London SJ, Connolly JL, Schnitt SJ, et al. A prospective study of benign breast disease and the risk of breast cancer. *JAMA.* 1992; 267:941–944. [PubMed: 1734106]
59. Geller BM, Nelson HD, Carney Pa, et al. Second opinion in breast pathology: policy, practice and perception. *J Clin Pathol.* 2014; 67:955–960. [PubMed: 25053542]
60. Pantanowitz L, Dickinson K, Evans AJ, et al. American Telemedicine Association clinical guidelines for telepathology. *J Pathol Inform.* 2014; 5:39. [PubMed: 25379345]
61. Egglin TKP, Feinstein AR. Context bias: A problem in diagnostic radiology. *JAMA.* 1996; 276:1752–1755. [PubMed: 8940325]

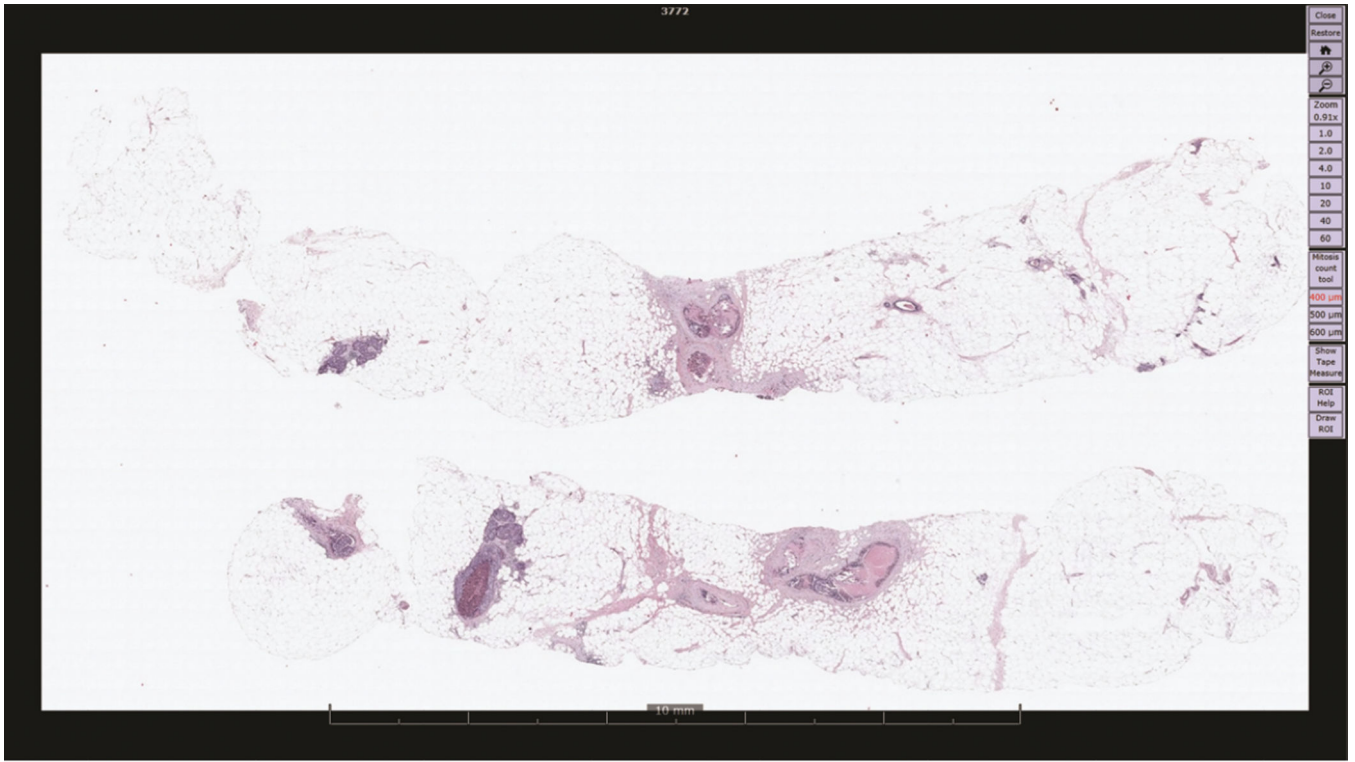
- Pathologists reviewed digital whole slide images and rendered diagnoses
- During interpretation, eye movements and image navigation behavior were recorded
- Pathologist experience positively predicted accuracy
- Case characteristics and eye movement patterns interactively predicted accuracy
- Higher zooming behavior predicted diagnostic over-interpretation

Author Manuscript

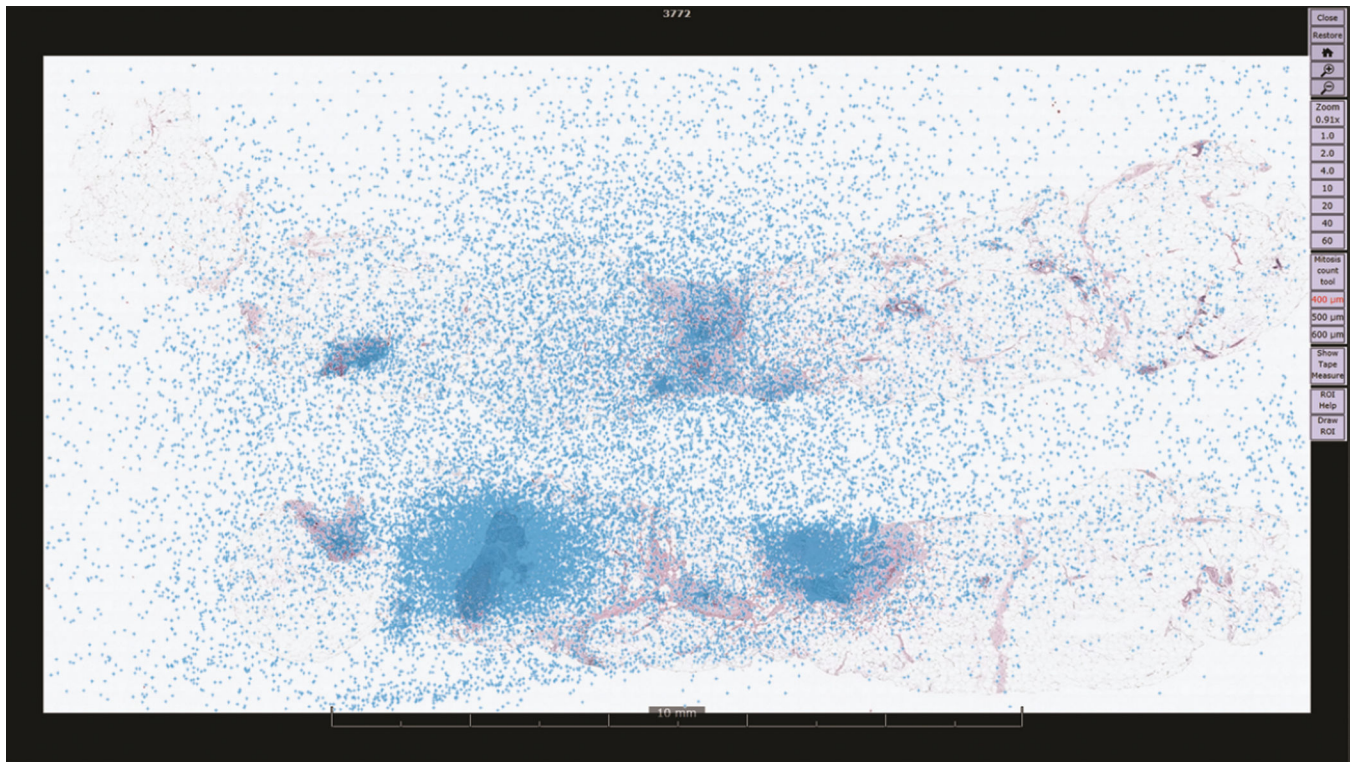
Author Manuscript

Author Manuscript

Author Manuscript



a.



b.

Figure 1.

a. A demonstration of a digitized whole slide image of a breast biopsy specimen with accompanying viewer interface controls. This case was given a diagnosis of ductal carcinoma in situ (DCIS) by an expert consensus panel.

b. Eye fixations (blue dots) overlaid onto the ductal carcinoma in situ (DCIS) case depicted in Figure 1. A total of 37,645 fixations are depicted, representing all 40 participating pathologists.

Table 1

Descriptions of all case-level, pathologist-level, and visual search-level independent variables.

Independent Variable	Level of Measurement	Description
Expert consensus diagnosis	Case	Consensus reference diagnosis derived from expert panel review. Diagnostic categories include: benign without atypia, atypia, DCIS and invasive cancer.
Mammographic breast density	Case	BI-RADS category rating (1–4) collected at the time of mammography preceding the breast biopsy.
Standardized difficulty rating	Case	Mean difficulty rating (1–6) by a separate group (N=115) of pathologists interpreting the glass slides.
Career level (resident, faculty)	Pathologist	Pathologist current appointment at academic medical center.
Digital WSI experience	Pathologist	Pathologist breadth of experience (# of uses) with digital WSI
Breast pathology experience	Pathologist	Pathologist years of experience interpreting breast pathology
Number of fixations	Visual Search Behavior	Total number of fixations made during case review
Mean fixation duration	Visual Search Behavior	Mean duration (in ms) of fixations made during case review
Blink rate	Visual Search Behavior	Mean blink rate (frequency/time) during case review
Proportion fixations in ROI	Visual Search Behavior	Proportion fixations falling within ROI(s) versus total number of fixations
Proportion fixations out ROI	Visual Search Behavior	Proportion fixations falling outside ROI(s) versus total number of fixations
Mean fixation duration in ROI	Visual Search Behavior	Mean duration of fixations falling within ROI(s)
Mean fixation duration out ROI	Visual Search Behavior	Mean duration of fixations falling outside of ROI(s)
Proportion re-entry to ROI	Visual Search Behavior	Proportion fixations re-entering ROI(s) versus total number of fixations
Mean viewer zoom level	Visual Search Behavior	Mean WSI viewer tool zoom level during case review
Peak viewer zoom level	Visual Search Behavior	Peak (maximum) WSI viewer tool zoom level (1–60x) during case review
Standard deviation zoom level	Visual Search Behavior	Standard deviation WSI viewer tool zoom level during case review

Table 2

Independent variables resulting from principal components analysis, with corresponding Eigen values.

PCA-Derived Independent Variable	Eigen Value
Fixation durations	3.05
Fixations within versus outside ROIs	2.45
Zoom behavior	1.84
Number of fixations	1.21
Pathologist clinical experience	1.81

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Overall accuracy, concordance rates (above- vs below-consensus), mean unders versus over-calling rates, and review time (in seconds) as a function of consensus diagnosis. Standard deviations included in parentheses.

Table 3

Consensus Diagnosis	Mean (95%CI) Concordance Rate	Above Consensus Diagnosis (95% CI)	Below Consensus Diagnosis (95% CI)	Mean (95%CI) Under/Over-calling Rate	Mean (95%CI) Review Time (sec)
Benign	71% (61%, 82%)	29% (20%, 40%)	-	.29 (.18, .40)	120.6 (101.6, 139.5)
Atypia	37% (29%, 45%)	21% (15%, 28%)	43% (35%, 50%)	-.22 (-.35, -.09)	193.9 (140.4, 247.5)
DCIS	52% (43%, 61%)	17% (12%, 23%)	31% (25%, 39%)	-.14 (-.25, -.04)	250.4 (164.8, 336.1)
Invasive	94% (88%, 99%)	-	6% (2%, 14%)	-.06 (-.12, -.01)	205.6 (147.9, 263.3)

Table 4

Best fitting generalized estimating equations Type III model effects and parameter estimates, predicting diagnostic accuracy.

Effect	B	SE	Odds Ratio Exp(B)	95% CI	Wald χ^2 (df)	P-value
<i>Type III Model Effects</i>						
Case Consensus Diagnosis					20.69(3)	<.001
Case Breast Density					8.17(1)	.004
Case Difficulty Rating					20.06(1)	<.001
Pathologist Experience					18.58(1)	<.001
Fixation Durations					6.84(1)	.009
Fixations within versus outside ROIs					11.24(1)	.001
Zoom Behavior					3.08(1)	.079
Number of Fixations					2.13(1)	.144
Pathologist Experience × Consensus Diagnosis					9.02(3)	.029
Pathologist Experience × Zoom Behavior					3.49(1)	.062
<i>Model Parameter Estimates</i>						
Case Consensus Diagnosis (reference: benign)						
Atypia	-.629	.357	.533	(.265, 1.07)	3.10(1)	.078
DCIS	-.588	.384	.555	(.262, 1.18)	2.35(1)	.125
Invasive	1.91	.614	6.72	(2.02, 22.4)	9.64(1)	.002
Case Breast Density (reference: BI-RADS 1-2)						
BI-RADS 3-4	.813	.285	2.26	(1.29, 3.94)	8.17(1)	.004
Case Difficulty Rating	-.695	.155	.499	(.368, .677)	20.06(1)	<.001
Pathologist Experience	-.054	.279	.947	(.547, 1.64)	.038(1)	.846
Fixation Durations	.201	.077	1.22	(1.05, 1.42)	6.84(1)	.009
Fixations in versus outside ROIs	.420	.125	1.52	(1.19, 1.94)	11.24(1)	.001

Effect	B	SE	Odds Ratio Exp(B)	95% CI	Wald χ^2 (df)	p-value
Zoom Behavior	-.207	.118	.813	(.645, 1.03)	3.08(1)	.079
Number of Fixations	-.166	.114	.847	(.677, 1.06)	2.13(1)	.144
Pathologist Experience × Consensus Diagnosis						
Atypia	.590	.335	1.80	(.936, 3.48)	3.11(1)	.078
DCIS	.910	.328	2.49	(1.31, 4.72)	7.72(1)	.005
Invasive	1.21	.540	3.37	(1.17, 9.71)	5.05(1)	.025
Pathologist Experience × Zoom Behavior	-.273	.146	.761	(.572, 1.01)	3.49(1)	.062

Table 5

Best fitting generalized estimating equations Type III model effects and parameter estimates, predicting over- versus under-calling rates.

Effect	B	SE	Odds Ratio Exp(B)	95% CI	Wald χ^2 (df)	P-value
<i>Type III Model Effects</i>						
Case Consensus Diagnosis					61.57(3)	<.001
Case Breast Density					2.15(1)	.143
Case Difficulty Rating					3.20(1)	.073
Pathologist Experience					5.79(1)	.016
Fixation Durations					.029(1)	.864
Fixations within versus outside ROIs					1.23(1)	.267
Zoom Behavior					36.29(1)	<.001
Number of Fixations					.389(1)	.533
Pathologist Experience × Fixations in/out ROIs					7.98(1)	.005
Case Consensus Diagnosis × Zoom Behavior					34.21(3)	<.001
Case Difficulty Rating × Fixations in/out ROIs					9.74(1)	.002
<i>Model Parameter Estimates</i>						
Case Consensus Diagnosis (referent: benign)						
Atypia	-1.93	.258	.146	(.088, .241)	55.76(1)	<.001
DCIS	-1.17	.271	.184	(.109, .313)	39.05(1)	<.001
Invasive	-1.61	.349	.200	(.101, .397)	21.18(1)	<.001
Case Breast Density (reference: BI-RADS 1-2)						
BI-RADS 3-4	.253	.173	1.29	(.918, 1.81)	2.15(1)	.143
Case Difficulty Rating	.256	.143	1.29	(.976, 1.71)	3.20(1)	.073
Pathologist Experience	.244	.101	1.28	(1.05, 1.56)	5.79(1)	.016
Fixation Durations	.015	.089	1.02	(.853, 1.21)	.029(1)	.864
Fixations in versus outside	-.334	.301	.716	(.397, 1.29)	1.23(1)	.267

Effect	B	SE	Odds Ratio Exp(B)	95% CI	Wald χ^2 (df)	p-value
ROIs						
Zoom Behavior	.627	.230	1.87	(1.19, 2.94)	7.42(1)	.006
Number of Fixations	.065	.104	1.07	(.870, 1.31)	.389(1)	.533
Pathologist Experience \times Fixations in/out ROIs	-.176	.062	.838	(.742, .947)	7.98(1)	.005
Case Consensus Diagnosis \times Zoom Behavior						
Atypia	.489	.281	1.63	(.940, 2.83)	3.03(1)	.082
DCIS	.191	.259	1.21	(.728, 2.01)	.539(1)	.463
Invasive	-.476	.229	.621	(.396, .974)	4.29(1)	.038
Case Difficulty Rating \times Fixations in/out ROIs	.358	.115	1.43	(1.14, 1.79)	9.74(1)	.002