Eye tracking reveals expertise-related differences in the time-course of medical image inspection and diagnosis

Tad T. Brunyé,[a,b,*] Trafton Drew,[c] Kathleen F. Kerr,[d] Hannah Shucard,[d] Donald L. Weaver,[e] and Joann G. Elmore[f]

[a]Center for Applied Brain and Cognitive Sciences, Medford, Massachusetts, United States
[b]Tufts University, Department of Psychology, Medford, Massachusetts, United States
[c]University of Utah, Department of Psychology, Salt Lake City, Utah, United States
[d]University of Washington, Department of Biostatistics, Seattle, Washington, United States
[e]University of Vermont, Larner College of Medicine and UVM Cancer Center, Department of Pathology, Burlington, Vermont, United States
[f]University of California, Los Angeles, David Geffen School of Medicine, Department of Medicine, Los Angeles, California, United States

Abstract

Purpose: Physicians' eye movements provide insights into relative reliance on different visual features during medical image review and diagnosis. Current theories posit that increasing expertise is associated with relatively holistic viewing strategies activated early in the image viewing experience. This study examined whether early image viewing behavior is associated with experience level and diagnostic accuracy when pathologists and trainees interpreted breast biopsies.

Approach: Ninety-two residents in training and experienced pathologists at nine major U.S. medical centers interpreted digitized whole slide images of breast biopsy cases while eye movements were monitored. The breadth of visual attention and frequency and duration of eye fixations on critical image regions were recorded. We dissociated eye movements occurring early during initial viewing (prior to first zoom) versus later viewing, examining seven viewing behaviors of interest.

Results: Residents and faculty pathologists were similarly likely to detect critical image regions during early image viewing, but faculty members showed more and longer duration eye fixations in these regions. Among pathology residents, year of residency predicted increasingly higher odds of fixating on critical image regions during early viewing. No viewing behavior was significantly associated with diagnostic accuracy.

Conclusions: Results suggest early detection and recognition of critical image features by experienced pathologists, with relatively directed and efficient search behavior. The results also suggest that the immediate distribution of eye movements over medical images warrants further exploration as a potential metric for the objective monitoring and evaluation of progress during medical training.

© *2020 Society of Photo-Optical Instrumentation Engineers (SPIE)* [DOI: 10.1117/1.JMI.7.5.051203]

Keywords: medical informatics; pathology; decision-making; medical image interpretation; eye-tracking; diagnosis; medical education; accuracy.

Paper 20148SSR received Jun. 1, 2020; accepted for publication Oct. 14, 2020; published online Oct. 27, 2020.

1 Introduction

When physicians interpret medical images, they engage a dynamic interplay between the perception of their visual world and their application of emergent and crystallized expertise in a domain.[1–4] This process is fundamental to image interpretation and diagnosis across a range of

*Address all correspondence to Tad T. Brunyé, tbruny01@tufts.edu

2329-4302/2020/$28.00 © 2020 SPIE

specialties including pathology, neurology, cardiology, dermatology, and radiology. With the advent of digital medical imaging, researchers and clinicians can gain new insights into the visual interpretive process by tracking image manipulation and eye movements.[1,2] These insights can lay an empirical foundation for quantitatively measuring the interpretive process, predicting when it might fail, and realizing new ways to optimize interpretation in education and clinical practice.[5–7]

Tracking image manipulation and eye movements during digital whole slide image review has revealed interesting features of the interpretive process and several potential methods for measuring performance in students and other trainees. A current paradigm parses the visual interpretive process into search, recognition, and decision-making; errors can emerge at each stage of this process.[1,2,8,9] Search errors involve failing to fixate the eyes on a critical image region, such as a radiologist who does not fixate a lung nodule on a chest x-ray. Recognition errors refer to a failure to recognize what is seen in a region, such as a pathologist who fixates upon atypical glands on a breast biopsy but does not recognize it as abnormal. Decision errors involve failing to map recognized features to an appropriate diagnosis, such as a cardiologist who accurately recognizes a pattern on an echocardiogram (ECG) but gives an incorrect diagnosis.[10] Eye tracking has introduced the potential to distinguish these three types of errors.

Eye tracking during medical image interpretation has also revealed differences in search, recognition, and decision patterns of novice versus expert diagnosticians. There are semidistinct strategic search patterns, with some observers scanning an image at low power and others tending to drill down to discrete image regions at high power.[11–13] Differences in search strategy tend to associate with physician expertise level[12,13] and may relate to diagnostic accuracy. Compared to experts, novice diagnosticians also tend to be more distracted by salient yet diagnostically irrelevant regions of an image,[14,15] spend less time interpreting challenging image regions,[16,17] spend more time overall on image inspection,[18,19] and show less global scanning and more local interrogation of (often irrelevant) image features.[20,21]

Eye-tracking studies also commonly find differences in local versus global processing. These differences have, in turn, been important for developing theory in this area and to understand the development of expertise in medical image interpretation. When experienced diagnosticians first view a medical image, they typically employ a broad sweep of visual attention across the scene. This broad sweep is thought to involve both foveated and parafoveal attention, with overt fixations on some regions and covert attention shifts toward other regions.[22] This breadth-first process is particularly evident in the most experienced observers, affords efficient Gestalt-like perception of the entire image, and is thought to serve to prioritize and set goals for subsequent review.[1,2,22,23] The process is also thought to result in multiple diagnostic hypotheses being formed, each of which is considered during subsequent feature interrogation.[14] Two extant theories describe this process. First, the information-reduction hypothesis proposes that experts have highly selective information processing that rapidly omits task-irrelevant information in favor of task-relevant information.[24] Second, the holistic model of image perception proposes that experts engage a broad initial visual scan of an image, quickly extracting information from disparate regions.[23] Following this initial breadth-first stage, experienced observers tend to spend more time examining diagnostically relevant, challenging, and often visually less salient regions of the image and less time exploring other areas. A few studies suggest that training in the breadth-first search strategies that are characteristic of expert searchers can accelerate the transition from novice to expert.[25,26]

However, a comprehensive meta-analysis of expertise studies using eye tracking across sports, medicine, and transportation domains found mixed support for some predictions of these theories.[27] For example, there was very weak evidence that experts show longer fixation durations on relevant versus irrelevant areas, and the time to first fixate a critical region was not reliably predicted by expertise. There was moderate evidence for experts displaying more and longer fixations on relevant regions than novices, supporting the information reduction hypothesis, and experts were sometimes faster to fixate relevant regions, supporting the holistic model of image perception. It is important to realize, however, that extant theories of medical image interpretation may not be suitable for application to domains outside of medicine (e.g., sports, transportation), or even to medical tasks with varied visual formats and demands.[28] The meta-analysis also found high effect size heterogeneity, and most effects were poor to moderate in strength, emphasizing the importance of increasing sample sizes for continuing research. Indeed, the studies in the meta-analysis examined an average of only 11 experts and 12 novices.

In pathology, digital whole slide imaging (WSI) provides an ideal format for exploring predictions of a breadth-first model of expert interpretation. With WSI, digital slide scanners are used to develop high-resolution digital versions of glass slides, with viewer tools that allow observers to zoom and pan images on a computer monitor. In this manner, observers are viewing a two-dimensional (2D) image but are able to zoom to very high levels (e.g., 40×) while maintaining resolution and viewing more detailed histopathological features of the image. By coupling the tracking of digital image manipulation (zooming and panning) with eye tracking, we can gain insights into the interpretive process from a perceptual and cognitive perspective. With digital WSI, the initial display of an image is at low power (1× magnification), affording a broad scan of the biopsy prior to selecting a region for high power examination (i.e., zooming). WSI affords an opportunity to examine observer behavior upon first examination, when early impressions are formed and decisions are made regarding image regions to prioritize for further examination.

Leveraging the opportunities provided by WSI, this study recruited a relatively large[1,27,29] sample size ($N = 92$) and specifically examined the early viewing period when an image is first displayed on the screen, prior to the first zoom behavior. This allowed us to test three primary predictions of a breadth-first model of expertise. First, we examined patterns of early image viewing behavior, comparing the behavior of residents versus experienced (faculty) pathologists. In line with previous research, we expected that residents would show evidence of relatively narrow attentional breadth during early viewing relative to faculty, including lower amplitude saccades and lower percentage fixation coverage of the tissue. Second, in our cross-sectional data we examined whether any of these variables show trends by year of residency. We expected that more senior residents would show behavior increasingly resembling the experts, compared to more junior residents. Third, we asked whether years of experience and early viewing behavior are associated with residents' diagnostic accuracy, examining the critical link between residency-based changes in visual behavior and diagnostic outcomes.

2 Methods

2.1 *Participants*

To derive a sample size estimate, we used aggregate effect size outcomes from a meta-analysis aimed at understanding expertise-related eye movement (e.g., fixation duration and time to first fixate) differences in several specialized professional domains, including medicine.[27] With a mean effect size ($r = 0.29$) derived from the meta-analysis, an α of 0.05 and power ($1 - \beta$) of 0.80, the sample size estimate suggested a minimum of 18 participants per group.

We collected data from 92 pathologists at nine major university medical centers located across the U.S., from February 2019 through October 2019. Breast pathology experience ranged from 72 residents with relatively limited experience with breast pathology to 20 faculty members who were comfortable interpreting breast pathology (Table 1). All participants provided written informed consent, and all study procedures were approved by the appropriate Institutional Review Boards (IRB), with the University of California, Los Angeles acting as the IRB of record (Protocol #18-000327).

2.2 *Materials and Equipment*

A set of 33 hematoxylin and eosin-stained digital WSI were selected from a larger test set of 240 cases developed in earlier studies. Each case included a consensus reference diagnoses, and most included one or more diagnostic regions of interest (dROI) as previously described.[30,31] Digital WSI were developed by scanning glass slides using an iScan Coreo Au digital slide scanner[32] at 40× objective magnification. To ensure our cases captured a range of clinically relevant diagnoses, selected cases included consensus diagnoses spanning five diagnostic categories: benign without atypia (4 cases), atypia (10 cases), low-grade ductal carcinoma *in situ* (lg-DCIS; 10 cases), high-grade DCIS (4 cases), and invasive carcinoma (5 cases).

Table 1 Participant demographic details for the 90 participants (70 residents, 20 faculty) included in analyses, following the removal of data from two residents due to failure of the eye-tracking system (see Sec. 3).

Participant group	Variable	Data
Residents	Year of residency training	Year 1: $N = 19$
		Year 2: $N = 25$
		Year 3: $N = 17$
		Year 4: $N = 9$
	Approximate weeks of breast pathology training	Year 1: $x = 3.03$
		Year 2: $x = 5.26$
		Year 3: $x = 7.21$
		Year 4: $x = 9.11$
	Sex	Male: $N = 34$
		Female, other, or undisclosed: $N = 36$
Experienced pathologists	Total years of experience interpreting breast pathology	<1 to 4 years: $N = 4$
		5 to 9 years: $N = 5$
		10 to 19 years: $N = 8$
		20+ years: $N = 3$
	Percentage of breast cases in current case load	<10%: $N = 4$
		10% to 24%: $N = 3$
		25% to 49%: $N = 4$
		50% to 74%: $N = 7$
		75% or more: $N = 2$
	Fellowship trained in breast pathology?	Yes: $N = 7$
		No: $N = 13$
	Sex	Male: $N = 8$
		Female: $N = 12$
		Other: $N = 0$
		Undisclosed: $N = 0$

To select 33 cases from the set of 240, we examined histology form data gathered from a total of 54 pathologists who previously interpreted the same cases on glass slides.[31] All 54 pathologists were fellowship-trained in breast pathology and/or their peers considered them an expert. Using these extant data, we selected 33 cases that were frequently diagnosed in one of our five diagnostic categories. To reduce the challenges associated with the interpretation of certain histopathological features, we eliminated cases frequently (>30%) diagnosed as lobular carcinoma *in situ* (LCIS), atypical lobular hyperplasia, or flat epithelial atypia (FEA). Cases assigned to the DCIS category were parsed into low-grade versus high-grade based upon nuclear grade and the presence of necrosis; cases with low or intermediate nuclear grade and no necrosis were considered low-grade, and cases with high nuclear grade and/or necrosis were considered high-grade.[33]

One invasive carcinoma case with high concordance (93%) when interpreted by prior pathologists in the glass slide format was selected for use as a practice case.[30] To limit the amount of time a participant would spend reviewing cases in the study (∼1 h), we divided the 32 cases into three test sets with equal distributions across consensus diagnostic categories: two benign cases, four atypia cases, six DCIS cases (four low-grade and two high-grade), and two invasive cases. This distribution of cases was intended to maximize sampling of the inherently challenging diagnostic categories that tend to elicit diagnostic discordance (i.e., atypia, low-grade DCIS[31,34]). Five cases, one for each diagnostic category, were used in all three test sets, and the remaining nine cases were unique to each test set. Most cases contained one or more dROIs, decided by an expert panel of pathologists using a modified Delphi technique, described previously.[30,34] The dROI was considered the image region(s) most representative of the highest (most clinically serious) consensus diagnosis for the case.

To enable WSI review, we developed a digital slide viewer that displayed navigable (zoom, pan) images, allowing up to 60× magnification while maintaining resolution. The viewer allowed image navigation with the computer mouse, using the scroll wheel to zoom and a click-and-drag movement to pan. It also provided buttons that could be used to manually adjust the zoom, a ruler overlay for scale, a mitotic field tool, and a drawing tool allowing participants to annotate the image. The viewer continuously logged image manipulation data (current zoom level and position in image) to a custom SQL database stored locally on the Dell Precision laptop computer used to run the study.

A histology form was developed from prior studies to collect diagnostic information from participants after they viewed each case.[31] The histology form collected information about diagnostic category (five categories), diagnostic specifics, additional information for some diagnostic categories (e.g., type of atypia, tubule formation, nuclear grade, necrosis, mitotic activity, and Nottingham score), difficulty and confidence ratings (scale from 1 to 6), yes/no responses whether the participant considered their diagnosis borderline between two diagnoses, and whether they would seek a second opinion about the case (Table 12 in Appendix). To assess accuracy, we compared participants' diagnostic responses to the consensus reference diagnosis (i.e., benign, atypia, low-grade DCIS, high-grade DCIS, invasive) for each case.

We used the remote eye-tracking device (RED) system, which is a noninvasive and portable eye tracking system manufactured by SensoMotoric Instruments (Boston, Massachusetts). The system uses an array of infrared lights and cameras to track eye position at 250 Hz with relatively high gaze position accuracy (0.4°) using a nine-point calibration process. For data collection, we had two systems, each of which included the RED device mounted to the bottom of a color-calibrated 22" Dell liquid crystal display (1920×1080 resolution) computer monitor (model U2417H). All analyses averaged the left and right eye position to reduce variable error, a measure referred to as the version signal.[35]

2.3 *Data Collection Locations and Procedures*

One of two investigators (TTB or TD) visited each of the nine data collection sites with one of the data collection systems (eye tracker, monitor, and computer). Data collection was completed with each participant one at a time in a private room (office or conference room) after web-based consenting and baseline survey. The baseline survey included questions about career level, confidence, and experience with digital imaging for primary diagnosis and experience with breast pathology. Upon arrival for a session, the experimenter would explain the study to the participant and answer any questions. The participant would then complete the nine-point eye tracker calibration, watching a white dot move between nine points on a gray background. The practice case was then displayed to help participants become familiar with the viewer controls, image navigation, annotation, and completing the histology form. After this introduction and practice, one of the three test sets (containing 14 cases) was presented, one case at a time in random order, at full screen. After viewing each case, the participants completed the histology form. Figure 1 shows a pathologist interpreting an image while their eyes are being tracked. Each participant was compensated for their time with a $50USD gift card.

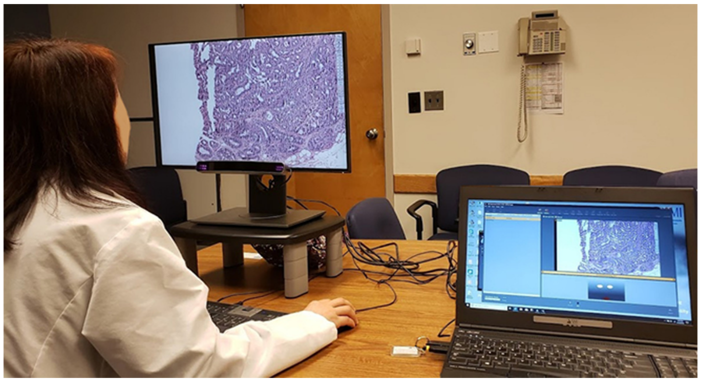

Fig. 1 A pathologist interpreting a digitized whole slide breast biopsy image (face obscured for privacy), with eyes being tracked. The screen on the right shows the SMI software used for data collection.

2.4 *Data Scoring and Analysis*

The early image viewing period is defined as the time when the case was at lowest power, before the first zoom (to a higher magnification level) was made. The early viewing period is therefore a variable time when the entire image is present on the screen for the participant, during which the participant can conduct an initial visual scan of the full biopsy specimen at low power. For terminology, early viewing behaviors occur before the first zoom behavior, whereas later viewing behaviors occur after the first zoom behavior and until the pathologist closes the case. The eye tracking system captures and outputs eye fixations and saccades over time. Fixations are momentary pauses (≥ 80 ms) of the eye in a constrained (2° visual angle) space, and saccades are ballistic movements of the eye between successive fixations.

By comparing the Cartesian coordinates of each fixation with the current viewer position in the image, we could ascertain whether participants were directing their eyes toward diagnostically critical regions (i.e., fixating on a dROI) versus other regions of the image. To distinguish eye fixations on tissue versus background space, we developed custom software (using the Python programming language[36]) that allowed us to outline tissue regions on each case and classify each fixation as falling within tissue or within background space.

We calculated seven measures that describe eye behavior during early image viewing (Table 2). First, we calculated the duration of the early viewing period, in seconds. Second, we calculated two measures that indicate the breadth of early viewing: saccade amplitude and breadth of fixation coverage. Saccade amplitude indicates the distance between successive eye fixations, in degrees of arc. Breadth of fixation coverage across the tissue was calculated by dividing the image into a 16×9 grid (each cell 120×120 px) and calculating the proportion of distinct grid cells that were fixated, including only cells that contained tissue (excluding white

Table 2 Each viewing behavior variable of interest, its level of analysis, and brief description.

Variable of interest	Description
Viewing duration	Mean duration (ms) of the early viewing period.
Saccade amplitude	Mean saccade amplitudes (degrees of arc) during the early viewing period.
Coverage	Mean coverage (%) during the early viewing period.
Probability of dROI fixation	Binary variable indicating whether at least one early viewing period fixation was on the dROI.
Time to first dROI fixation	Time (ms) to first fixation in dROI during the early viewing period.
Fixation count on dROI	Mean fixation count on dROI during the early viewing period
Fixation duration dROI	Mean fixation duration (ms) on dROI during the early viewing period.

space). This approach was intended to examine the possibility that an observer could, theoretically, examine a breadth of tissue through a sequence of small-amplitude saccades, rendering the saccade amplitude measure a less meaningful measure of early viewing breadth. We then calculated four additional measures that captured attention toward the dROI. First, a binary indicator of whether a pathologist fixated or did not fixate on the dROI during early case inspection. Second, a measure of time during early case inspection before a pathologist's first dROI fixation, in seconds. Third and fourth, if a pathologist fixated at least once on the dROI during early viewing, we calculated the number and duration of fixations in the dROI. To assess diagnostic accuracy, each categorized diagnosis made by pathologists was compared to one of four consensus reference diagnosis categories: benign, atypia, DCIS, or invasive.

We conducted three primary statistical analyses, using $\alpha = 0.05$ as the threshold for statistical significance and additionally noting effects that were marginally significant ($0.05 < P \leq 0.10$). First, we compared each of the seven early viewing behaviors between residents and faculty. To do so, for each variable we fit a generalized estimating equation (GEE) model with the behavior variable as the outcome. For continuous variables we used linear models. We $\log_2$-transformed duration, Saccade amplitude, time to first dROI fixation, and dROI fixation duration due to extremely right-skewed distributions (Fig. 2). For the dichotomous variable recording

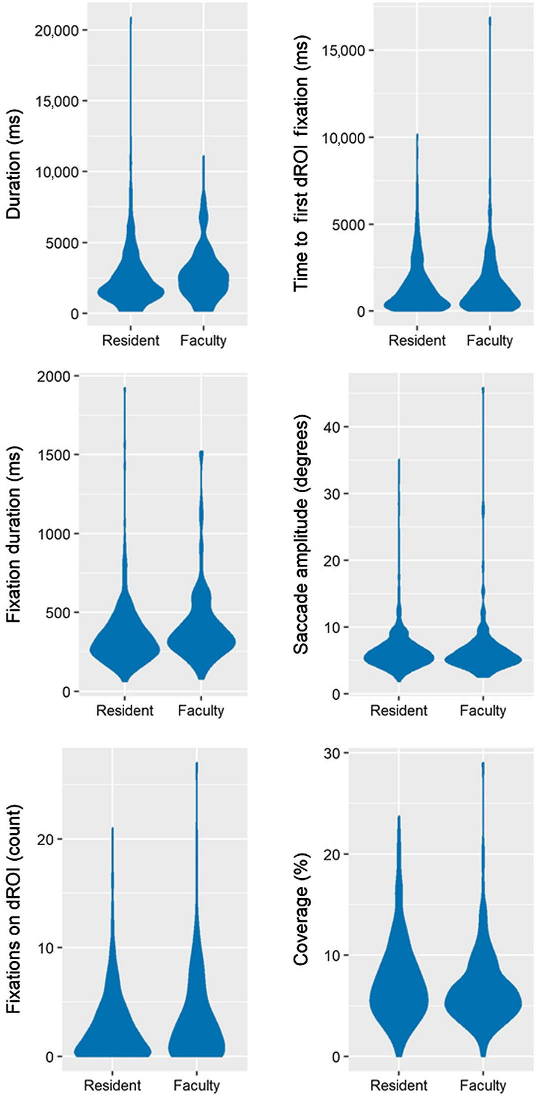

Fig. 2 Violin plots depicting pretransform data for the six continuous early viewing variables (before first zoom), separated by resident and faculty. Sample proportions for the binary variable (fixated on dROI) were 0.73 (residents) and 0.72 (faculty).

whether or not a participant fixated on the dROI during early viewing, we used a binomial model with logit link function (i.e., logistic regression). The predictor of interest in each model was an indicator variable of faculty status (1 for faculty, 0 for residents). Because not all study participants viewed the same cases, each model included case ID as a fixed effect, to yield the desired interpretation of comparing faculty and residents interpreting the same case. For the binomial model, we used conditional logistic regression using the cases as the strata. Since each participant interpreted 14 cases, we treated each participant as a cluster in the GEE model. For each variable, we report the estimated regression coefficient that represents the contrast between residents and faculty interpreting the same case along with its associated Wald p-value.

The second set of analyses asked whether year of residency is associated with each of our seven viewing variables. For these GEE analyses, residents were coded as year in the pathology training program as 1, 2, 3, or 4, and year was used as a predictor of each of the seven viewing variables. Otherwise, the statistical models are the same as with the first set of analyses comparing residents to faculty. Note that years of resident experience was strongly associated with self-reported experience with breast pathology (Table 1), and only the former more objective measure was analyzed. One post-sophomore fellow was categorized as a first-year resident, and one fellow was categorized as a fourth-year resident.

The third set of analyses examined associations between residents' viewing behaviors and diagnostic accuracy. We also examined the association between year of residency and diagnostic accuracy. For these analyses, each of the eight variables —year of residency (1 to 4) and the seven viewing behavior variables—was used, in turn, as the predictor of interest in a GEE model. The model used logit link with accuracy as the binary outcome, included case as a fixed effect, and clustered on participants.

3 Results

At the level of participants, the eye tracking system failed to save any data for two participating residents, so their data were excluded from analysis (leaving 90 participants: 70 residents and 20 faculty). At the level of individual cases, the eye tracking system failed to capture data for 45 of the 1260 case interpretations (3.6% of all cases), so data from these cases were also removed from analysis.

Overall, residents were nearly evenly distributed across the case test sets ($N = 22$, 24, 24, respectively), as were faculty ($N = 5$, 7, 8, respectively). For analysis, two case categories were excluded: benign and invasive. Benign cases were excluded due to a lack of consensus dROIs (i.e., there is no single critical image feature to detect), and thus the inability to calculate many of the early viewing measures. Invasive cases were excluded due to (near) perfect accuracy (96% to 100%) that did not vary between residents and faculty, and the fact that dROIs covered nearly the entire tissue space (making these cases less useful for examining early attention to critical regions). Analyses therefore focused on generally more challenging cases, atypia and DCIS, with 10 atypia cases, and 14 DCIS cases (with each participant viewing 4 atypia cases and 6 DCIS cases).

For the analysis of diagnostic accuracy, one DCIS case was excluded from analysis due to floor-level accuracy (0%), leaving 23 cases for analysis (10 atypia, 13 DCIS). Overall, median accuracy among residents was 25% (mean = 26%), and 60% (mean = 54%) among faculty. As would be expected, a conditional logistic regression with case ID as a fixed effect showed that on average faculty had 2.4 times higher odds of an accurate diagnosis than residents interpreting the same case (95% CI: 1.9, 3.1; $P < 0.0001$).

3.1 *Viewing Behavior: Resident Versus Faculty Patterns*

Figure 2 shows pretransformed data for six early viewing variables, parsed by resident versus faculty pathologists (see Table 6 in Appendix for numeric summaries). Results of the statistical analyses are detailed in Table 3, showing that, on average, resident pathologists had shorter fixation durations on the dROI compared to faculty. Specifically, the geometric mean of fixation duration on the dROI was, on average, 16% lower among residents compared to faculty (95% CI: 4% to 30% lower, $P < 0.01$).

Table 3 Associations between career level (residents versus faculty) and early viewing behavior. For each variable, a positive value of $\hat{\beta}$ indicates higher values observed in faculty pathologists compared to residents, on average. Negative values indicate the reverse.

Variable	Analysis scale	$\hat{\beta}$ for experienced	*StErr*	*P*
Viewing duration (s)	Log_2	0.224[a]	0.156	0.15
Saccade amplitude (°)	Log_2	−0.049[a]	0.084	0.56
Coverage (%)	Untransformed	−0.861[b]	0.461	0.06
Fixated dROI (binary)	NA (Categorical)	−0.035[c]	0.098	0.70
Time to first dROI fixation (s)	Log_2	−0.018[a]	0.163	0.91
Fixation count on dROI (n)	Untransformed	0.623[b]	0.338	0.07
Fixation duration dROI (ms)	Log_2	0.219[a]	0.082	<0.01

[a] $2^{\hat{\beta}}$ estimates the ratio of geometric means for faculty pathologists compared to residents interpreting the same case.
[b] $\hat{\beta}$ estimates the difference in means for faculty pathologists compared to residents interpreting the same case.
[c] $\exp(\hat{\beta})$ estimates the odds ratio comparing faculty pathologists to residents interpreting the same case.

Two additional effects were marginally significant, suggesting that residents tend to show higher coverage of the image compared to faculty, and suggesting that residents tend to have fewer fixations on the dROI compared to faculty. Of course, given these effects did not meet our alpha criterion (0.05), we cannot conclude the population effects are nonzero; instead, we point to the pattern as a potential direction for future inquiry.

Exploratory follow-up analyses examined later viewing behavior differences between residents and faculty, including only eye movements after the first zoom; pretransformed later viewing data are included in Table 8 in Appendix. As detailed in Table 9 in Appendix, residents had longer duration later viewing than faculty, and smaller overall saccade amplitudes; they also showed marginally longer time to first fixation on the dROI.

3.2 *Viewing Behavior: Residency Trends*

We assessed trends in the seven viewing behaviors by year of residency (years 1 to 4) among the 70 resident (nonfaculty) participants (see Table 7 in Appendix for numeric summaries of these variables stratified by year of residency).

Results of the statistical analyses are detailed in Table 4. On average, each year of residency was associated with 10% higher odds of fixating the dROI (95% CI 1% to 20%, $P = 0.03$) when comparing residents viewing the same case. The trend is shown in Fig. 3. Note that Fig. 3 is intended to depict the overall trend captured in the formal analysis; it is difficult to compare one specific year to another (e.g., year 2 versus 3) given high heterogeneity in breast pathology curricula across training institutions. No other early viewing behaviors were significantly associated with year of residency.

Exploratory follow-up analyses examined later viewing behavior trends by residency year, including only eye movements occurring after the first zoom behavior. As detailed in Table 10 in Appendix, each year of residency was associated with 2% higher odds of fixating the dROI during later viewing (95% CI 0% to 4%, $P = 0.03$), a similar trend as early viewing behavior.

3.3 *Viewing Behavior: Associations with Diagnostic Accuracy*

Results of the statistical analyses are detailed in Table 5. Year of residency was positively associated with accuracy ($P = 0.05$). On average, comparing residents 1 year apart interpreting the same case, the more senior residents have 16% higher odds of accurate diagnosis (95% CI for odds ratio, 1.00 to 1.36). No other variable was significantly associated with accuracy.

Table 4 Trends in viewing behaviors during early viewing by year of residency (1 to 4). For each variable, a positive value of $\hat{\beta}$ indicates a trend toward higher values among residents in later years of residency compared to more junior residents, on average. Negative values indicate the reverse.

Variable	Analysis scale	$\hat{\beta}$ per year	*StErr*	*P*
Viewing duration (s)	Log_2	0.105[a]	0.064	0.10
Saccade amplitude (°)	Log_2	−0.037[a]	0.027	0.17
Coverage (%)	Untransformed	0.484[b]	0.302	0.11
Fixated dROI (binary)	NA (categorical)	0.095[c]	0.045	0.03
Time to first dROI fixation (s)	Log_2	−0.019[a]	0.084	0.82
Fixation count on dROI (n)	Untransformed	0.093[b]	0.159	0.56
Fixation duration dROI (ms)	Log_2	0.001[a]	0.039	0.98

[a] $2^{\hat{\beta}}$ estimates the ratio of geometric means for residents 1 year apart in year of residency interpreting the same case.

[b] $\hat{\beta}$ estimates the difference in means comparing residents 1 year apart in year of residency interpreting the same case.

[c] $\exp(\hat{\beta})$ estimates the odds ratio comparing residents 1 year apart in year of residency interpreting the same case.

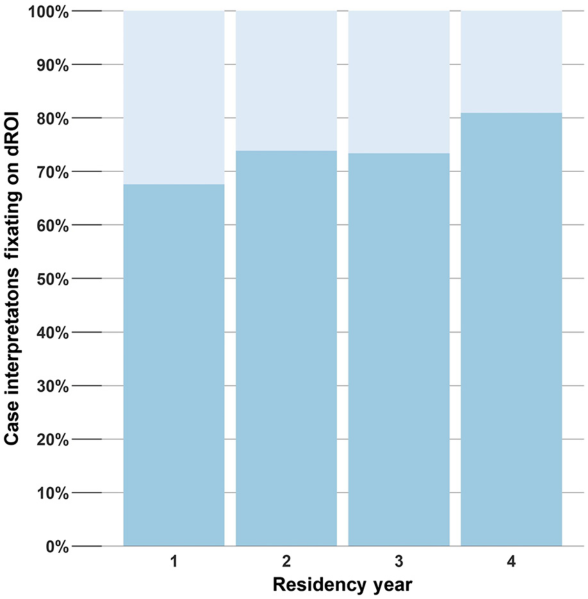

Fig. 3 Proportion of case interpretations fixating on the dROI by year of residency (1 to 4).

Exploratory follow-up analyses examined associations between later viewing behavior and diagnostic accuracy. As detailed in Table 11 in Appendix, no variable was significantly associated with accuracy.

4 Discussion

This study was designed to assess interpretive viewing patterns that characterize expertise and expertise development and possibly relate to diagnostic accuracy. In the largest study of its kind, we analyzed the behavior of 90 pathologists at nine different academic medical centers, ranging in expertise from first-year resident to attending physician with over 20 years' experience

Table 5 Associations between year of residency and early viewing behaviors and accuracy. For each variable, a negative value of $\hat{\beta}$ indicates a trend toward lower values associated with higher accuracy. Positive values indicate the reverse.

Variable	Analysis scale	$\hat{\beta}$	*StErr*	*P*
Year of residency (1 to 4)	NA (categorical)	0.151	0.077	0.05
Viewing duration (s)	Log_2	−0.067	0.054	0.14
Saccade amplitude (°)	Log_2	−0.053	0.153	0.72
Coverage (%)	Untransformed	−0.029	0.021	0.17
Fixated dROI (binary)	NA (categorical)	0.055	0.229	0.80
Time to first dROI fixation (s)	Log_2	−0.009	0.049	0.84
Fixation count on dROI (n)	Untransformed	−0.032	0.029	0.28
Fixation duration dROI (ms)	Log_2	−0.093	0.138	0.46

interpreting breast pathology. With this large and representative sample, we found mixed support for extant research and theory.

Research using eye tracking has suggested that relatively experienced diagnosticians employ a broader initial scan of a specimen than their less experienced counterparts,[11,14,22,23,37,38] supporting the holistic model of image perception. This broad initial scan is thought to enable consideration of multiple regions for subsequent focused inspection,[2] promoting an explore-exploit process that involves an early search for reward sources and then actively interrogating an identified source.[39] To investigate evidence of this pattern, we examined two measures of early viewing breadth. First, we calculated mean saccade amplitudes, with previous research demonstrating generally larger saccade amplitudes with higher experience levels.[20,27] Second, we adopted a coverage measure from previous research to quantify the extent to which broader regions of a scene are examined.[40,41] In contrast to some earlier work, our data showed that residents and attending pathologists showed very similar saccade amplitudes and breadth of coverage during early image viewing. There are at least two reasons why this might be the case. First, the hematoxylin and eosin (H&E) stains used in the current biopsies produce colorful high contrast regions that might serve to quickly orient attention toward relevant information, reducing differences that novices and experts may exhibit in other medical imaging domains. Second, eye tracking is only able to index foveated visual attention, and theories of holistic image perception suggest that at least a portion of early image scanning might be done parafoveally.[23,42,43] In other words, the eyes might not always follow a transient shift of attention to peripheral image regions. Thus, experts may indeed be conducting a more holistic initial scan of an image, but if this is being done covertly, we may not detect the behavior. This possibility may be exacerbated by high contrast image features that can be perceived without an overt shift of attention, allowing pathologists to prioritize image regions without directly viewing them. Images used in other medical imaging domains, such as radiology, are usually more homogeneous in color and contrast, perhaps necessitating a relatively thorough scan of the image to identify areas for subsequent review.

Beyond the breadth of early viewing, we found a few additional patterns that distinguished early viewing behavior among residents versus faculty pathologists. Interestingly, residents and faculty were similarly likely to move their eyes to a critical dROI and took similar time to first fixate on the dROI. However, once they fixated in the dROI their behaviors began to diverge, with faculty showing generally more and longer duration fixations in the dROI relative to residents. Longer fixation durations are thought to reflect several perceptual and cognitive processes, including higher reward value of information in a region, a more challenging interpretive process, and/or successful recognition of perceived features.[1,38,44–46] In other words, these data suggest that more experienced pathologists are more likely to move from successful detection of a critical region to successful processing and recognition of its features. Interestingly, this process begins relatively quickly during early image viewing and before the pathologists have zoomed to examine specific features more closely.

This special series for the *Journal of Medical Imaging* concerns medical image interpretation in 2D and three-dimensional (3D) formats; while this study used 2D, it is worth considering how our results might relate to the interpretive process with 3D images. With WSI, surgical pathology is confined to relatively flat tissue sections with minimal topographical variation. While the glass slides technically have some variation in focal depth (~0.005 mm), digital WSI can be considered a 2D representation with zoom capabilities designed to represent various objective magnifications. On a traditional microscope with glass slides, a pathologist would move on the horizontal plane (x, y) for target identification at low magnification, change the objective magnification of the microscope (several fixed vertical Z-planes), and place identified features in central vision for higher resolution examination. With digital WSI, variable (continuous) magnification is accomplished by scanning the glass slide at a high objective magnification and digitally recreating lower objective magnification along the vertical (z) axis, producing multiple image pseudoscans at various magnification (Z-planes).[47,48] A magnification plane can be considered as a specific physical height above the glass slide (in microns),[49] which in WSI represents a series of magnification levels (e.g., 1×, 5×, 10×, 20×, 40×). The collection of scanned focal plane images is then combined into a master pyramidal image, making it possible to zoom in and out on the image similar in principle to changing the objective magnification on the microscope.

The stacked representation of WSI can give the imaginary impression of depth[49] and possibly introduce some of the mental demands seen when navigating a 3D volumetric image.[50] For example, maintaining a catalog in working memory of areas already viewed on not only the horizontal axes but also in depth (vertical axis).[13] Thus, WSI introduces an interesting case for visual search residing somewhere between the demands of a fixed-depth 2D image and shifting between slices in a 3D volumetric scan. The difference being that in a 3D volumetric scan, the slices (Z-axis) are at the same magnification, whereas in the WSI the Z-axis adds additional feature detail (magnification plus new information). In some cases, such as invasive carcinoma where diagnostic features are pervasive and salient, successful categorical diagnosis (e.g., invasive cancer present versus absent) may be possible at low magnification when the entire case is visible on the display. At this fully zoomed-out level, fine-grained cytologic features will not be perceived, but a skilled pathologist may be able to make a coarse, but accurate, categorical determination based on the presence of extreme architectural features (e.g., invasion). However, verification of that hypothesis would require examining features at higher zoom levels, achieving detailed focus on specific features and deriving accurate estimates of diagnostically relevant markers (e.g., nuclear grade, mitotic activity, and/or tubule formation).

We also conducted an exploratory analysis of later viewing behavior, after pathologists made their initial zoom. This follow-up analysis demonstrated two additional patterns. First, faculty pathologists showed lower overall viewing durations than residents, an efficiency advantage typically seen with increasing expertise in a domain.[27] Second, a more interesting result demonstrated that faculty tended to show higher saccade amplitudes than residents during later viewing, an association not found during early viewing. This was unexpected as the holistic model of image perception suggests that faculty pathologists would show larger saccade amplitudes during the early global analysis of the image.[23,42] To our knowledge, no other study has separately examined early versus later viewing periods; with our approach, we found evidence that expertise-related saccade amplitude differences may emerge later during viewing. It could be the case that more experienced pathologists show a relatively directed visual search during early viewing, quickly identifying diagnostically relevant regions; indeed, according to Lesgold's model,[51] experts may engage in perceptual and cognitive processes early in image inspection, whereas novices tend to front-load perceptual processes. While faculty tend to show highly directed search to relevant features during early viewing, later in their search they may exhibit a broad sweep to exclude other possible regions of interest or to catalog additional diagnostic findings that are not normal but less clinically significant. This is not to say that faculty are not performing a relatively covert (parafoveal) early scan of the image but that evidence for increased scanning breadth may be more robust during more extended image viewing. Again, this pattern may be specific to pathologists viewing images with highly contrasting regions that can quickly capture and guide attention during early viewing (irrelevant but visual salient regions are also shown to distract novices more than experts[15,52]) and may not be generalizable to medical imaging formats other than WSI.

Our second cross-sectional analyses examined associations between year of residency and early viewing behavior. While most of the early viewing behaviors were not significantly associated with year of residency, residents did show an ∼10% higher odds of fixating the dROI with each additional year of residency training. This finding suggests that residency training is laying a knowledge foundation that affords efficient detection of critical image regions and that this knowledge progresses across years of training. When contrasted with the longer fixation durations in the dROI among faculty members, however, there is a suggestion that residents may not have the requisite knowledge to effectively recognize histopathological features in the dROI and map them to appropriate diagnoses. An exploratory analysis of later viewing behavior showed a similar but less pronounced association, with residents showing ∼2% higher odds of fixating the dROI with each additional year of residency training.

Analysis of diagnostic accuracy by year of residency estimated that each year of additional training was associated with ∼16% higher odds of correct diagnosis. However, even third- and fourth-year residents showed about 43% lower accuracy than faculty pathologists when interpreting atypia or DCIS cases. Altogether, results suggest that residents are increasingly likely to detect critical image features as they move through residency training, and their accuracy improves year over year, but relative to more experienced faculty they are still less likely to recognize critical features and map them to an appropriate diagnosis. This pattern suggests that advancements in resident training may be found in focusing medical education efforts on a few phases of the diagnostic process: recognizing histopathological features, developing and testing diagnostic hypotheses, and arriving at diagnostic decisions. Indeed, residents detect and move their eyes toward abnormalities like faculty but are less adept at recognizing and understanding how features map to diagnoses.

In motivating this study, we considered two specific theories borne out of cognitive science and medical decision-making research: the information-reduction hypothesis and the holistic model of image perception. Our results provide limited support for both theories. One prediction of the information-reduction hypothesis is that experts should be able to quickly reduce complex information to a limited set of crucial features that will guide interpretation.[24] While both residents and more experienced pathologists took a similar amount of time to find a critical image region, experienced pathologists showed longer fixation durations in the region once it was found (and a trend toward more fixations in the region). According to previous research, these prolonged fixation durations may indicate successful attendance to and recognition of critical features, providing some limited support for the information-reduction theory.[38,53] Second, the holistic model of image perception predicts that experts should show a broader and more efficient early search of an image relative to novices, as typically demonstrated with larger saccade amplitudes.[22] We found no evidence that the initial breadth of a viewing behavior, using saccade amplitude or a measure of image coverage, differed between novices and experienced pathologists. However, an exploratory analysis of later viewing behavior did find evidence of increased saccade amplitudes among more experienced pathologists, suggesting strategy-based search differences among residents versus faculty pathologists. We acknowledge that eye tracking may not be the best technique for assessing holistic image perception that may proceed both overtly (foveal) and covertly (parafoveal). Methods for inferring the distribution of covert attention may prove valuable for future research in this area, including attention probes,[54] examining the spatial orientation of microsaccades,[55] measuring electromyography in neck muscles,[56] or examining the breadth of expert useful field of view.[57] Consistent support for a holistic model of image perception may require these innovative approaches to gain insights into the time course and spatial distribution of covert attention during early versus later image viewing.

Eye tracking has revealed several important features of the visual interpretive process underlying medical decision-making, making this technology potentially useful to demonstrate, train, and assess visual performance during medical education.[1] Quantitative performance metrics derived from this research can assist the medical community's desire to adopt meaningful, relevant, and repeatable outcomes-based assessments during medical training (e.g., education, residency, fellowship).[58–60] One method for training visual search strategies is eye-movement modeling examples (EMMEs), which provide video demonstrations and narrations of expert eye movements, allowing trainees to observe, learn from, and emulate specific search strategies.[26] Some research suggests that EMMEs can be valuable for accelerating the transition from novice to expert in both medical and nonmedical domains, including detecting seizures, inspecting aircraft, debugging programs, and learning how to read.[61–64] However, a recent review of this literature notes limited effect sizes and generalizability for the EMME approach,[65] suggesting that EMMEs may help novices develop search strategies but may not be effective at promoting accurate feature recognition or diagnosis. Of course, there is still a critical step between detecting a critical region and recognizing and accurately diagnosing features. Exposing novices to high numbers of normal and abnormal examples is likely more beneficial than guiding their looking behavior,[65] promoting the development of robust and highly differentiated target templates in memory. These templates can then be used to efficiently guide attention to visual features and more effectively map those features to templates representing diagnostic categories.[66]

Presently, we found no compelling evidence that visual behavior can help dissociate situations in which a resident arrived at an accurate versus inaccurate diagnosis. Indeed, recognizing features and mapping to templates stored in long-term memory may be better captured by think-aloud protocols and annotations of recognized features, rather than eye tracking. Our continuing research will explore this possibility, allowing more effective parsing of detection, recognition, and diagnostic phases and the points at which errors may emerge in this process.

4.1 *Limitations*

In interpreting our results and motivating continuing research, some limitations are worth discussing. First, breast pathology is an inherently challenging medical discipline, with diagnostic disagreement arising even among highly experienced practicing pathologists interpreting atypia and DCIS cases.[31,34,67,68] For this reason, it is challenging to reliably associate viewing behavior to diagnostic outcomes that are highly variable both within and across physicians. Second, H&E staining of biopsies produces high contrast regions that likely attract attention early in the viewing process and many of these high contrast regions are relevant to diagnosis. Thus, these highly salient regions may capture the visual attention of even relatively inexperienced viewers, making it difficult to determine how expertise modulates the initial feature detection phase of image inspection. Third, while we used the largest sample size to date in a study examining eye tracking with pathologists, and sampled from a diverse set of university medical centers, training and cultural differences at nonacademic institutes may lead to differences in feature reliance and diagnostic interpretation. Finally, while we used zoom level as an objective method for parsing early versus later viewing episodes, the duration of early viewing varied considerably across participants, likely reflecting differences in search strategies (e.g., driller versus scanner).[12,13] For example, a search strategy involving quickly "drilling" into a specific image region after only a few seconds of low power viewing will reduce the likelihood that early image viewing will prove informative in our early viewing analyses. Each of these limitations can be used to motivate continuing research with more diverse pathology images (e.g., breast pathology, dermatopathology), biopsy staining methods, participant sampling, and techniques for parsing early versus later viewing.

4.2 *Conclusion*

In conclusion, in the largest study of its kind, results support the premise that experience-based progression of specialized medical knowledge manifests in both viewing behavior and diagnostic interpretation. The extent to which both novice and expert diagnosticians can quickly identify and exploit important and relevant regions of a scene is critical to interpretive efficiency and lays a foundation for diagnostic accuracy. Results provide mixed support for extant research and theory and also lay a foundation for further research to better elucidate error sources during medical interpretation, refine extant theories of medical image interpretation, and identify tractable applications of eye tracking technology for training and clinical practice.

5 Appendix

Tables 6–12 provide additional information.

5.1 *Early Viewing Variables and Behavior*

Tables 6 and 7 provide pretransformation mean and standard deviation data for each of the early viewing variables, compared by experience level (Table 6) and separated by year of residency training (Table 7).

Table 6 Pretransformation mean and standard deviation data for each of early viewing variables, compared by experience level (residents versus attending pathologists).

	Resident pathologists ($N = 70$)		Experienced pathologists ($N = 20$)	
Variable	Mean	StDev	Mean	StDev
Viewing duration (s)	2.48	2.03	2.80	1.93
Saccade amplitude (°)	6.54	3.33	6.50	4.32
Coverage (prop.)	0.075	0.041	0.067	0.036
Fixated dROI (binary)	0.729	—	0.718	—
Time to first dROI fixation (s)	1.40	1.49	1.40	1.78
Fixation count on dROI	2.75	3.09	3.37	3.95
Fixation duration dROI (ms)	345.9	183.5	409.4	232.4

Table 7 Pretransformation mean and standard deviation data for each of the early viewing behavior variables, separated by year of residency training. Note that Fixated dROI is a calculated proportion based on binary outcomes.

	First-year residents ($N = 19$)		Second-year residents ($N = 25$)		Third-year residents ($N = 17$)		Fourth-year residents ($N = 9$)	
Variable	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev
Viewing duration (s)	2.42	1.77	2.79	2.31	2.12	1.75	2.44	2.14
Saccade amplitude (°)	6.41	3.19	6.99	3.94	6.68	3.09	5.49	1.66
Coverage (prop.)	0.07	0.04	0.08	0.04	0.07	0.04	0.09	0.05
Fixated dROI (binary)	0.68	—	0.74	—	0.73	—	0.81	—
Time to first dROI fixation (s)	1.40	1.46	1.56	1.71	1.26	1.33	1.20	1.34
Fixation count on dROI	2.51	2.87	3.00	3.37	2.46	2.85	3.09	3.15
Fixation duration dROI (ms)	350.5	161.4	339.4	188.4	350.8	212.9	342.4	152.6

5.2 *Later Viewing Variables and Behaviors*

Table 8 provides pretransformation means and standard deviations for later viewing variables, by experience level. Table 9 details associations between career level (residents versus faculty) and later viewing behavior. Table 10 details trends in viewing behaviors during later viewing by year of residency. Table 11 shows associations between later viewing behaviors and accuracy.

Table 8 Pretransformation means and standard deviations for later viewing variables, by experience level.

Variable	Resident pathologists ($N = 70$)		Experienced pathologists ($N = 20$)	
	Mean	StDev	Mean	StDev
Viewing duration (s)	104.6	57.7	88.6	44.9
Saccade amplitude (°)	5.68	1.91	6.27	1.82
Fixated dROI (binary)	0.950	0.217	0.957	0.202
Time to first dROI fixation (s)	16.1	20.0	12.7	14.4
Fixation count on dROI	136.8	120.0	150.4	130.2
Fixation duration dROI (ms)	306.4	65.1	322.4	67.1

Table 9 Associations between career level (residents versus faculty) and later viewing behavior. For each variable, a positive value of $\hat{\beta}$ indicates higher values observed in faculty pathologists compared to residents, on average. Negative values indicate the reverse.

Variable	Analysis scale	$\hat{\beta}$ for experienced	*StErr*	*P*
Viewing duration (s)	Log_2	−0.235[a]	0.113	0.04
Saccade amplitude (°)	Log_2	0.147[a]	0.072	0.04
Fixated dROI (binary)	NA (categorical)	<0.001[b]	0.085	0.98
Time to first dROI fixation (s)	Log_2	−0.172[a]	0.102	0.09
Fixation count on dROI (n)	Untransformed	13.47[c]	15.99	0.39
Fixation duration dROI (ms)	Log_2	0.077[a]	0.055	0.16

[a] $2^{\hat{\beta}}$ estimates the ratio of geometric means for faculty pathologists compared to residents interpreting the same case.

[b] $\exp(\hat{\beta})$ estimates the odds ratio comparing faculty pathologists to residents interpreting the same case.

[c] $\hat{\beta}$ estimates the difference in means for faculty pathologists compared to residents interpreting the same case.

Table 10 Trends in viewing behaviors during later viewing by year of residency (1 to 4). For each variable, a positive value of $\hat{\beta}$ indicates a trend toward higher values among residents in later years of residency compared to more junior residents, on average. Negative values indicate the reverse.

Variable	Analysis scale	$\hat{\beta}$ per year	*StErr*	*P*
Viewing duration (s)	Log_2	0.042[a]	0.052	0.42
Saccade amplitude (°)	Log_2	−0.018[a]	0.025	0.47
Fixated dROI (binary)	NA (categorical)	0.023[b]	0.039	0.02
Time to first dROI fixation (s)	Log_2	0.041[a]	0.061	0.51
Fixation count on dROI (n)	Untransformed	1.582[c]	5.465	0.77
Fixation duration dROI (ms)	Log_2	0.028[a]	0.019	0.17

[a] $2^{\hat{\beta}}$ estimates the ratio of geometric means for residents 1 year apart in year of residency interpreting the same case.

[b] $\exp(\hat{\beta})$ estimates the odds ratio comparing residents 1 year apart in year of residency interpreting the same case.

[c] $\hat{\beta}$ estimates the difference in means comparing residents 1 year apart in year of residency interpreting the same case.

Table 11 Associations between later viewing behaviors and accuracy. For each variable, a negative value of $\hat{\beta}$ indicates a trend toward lower values associated with higher accuracy. Positive values indicate the reverse.

Variable	Analysis scale	$\hat{\beta}$	*StErr*	*P*
Viewing duration (s)	Log_2	0.047	0.108	0.62
Saccade amplitude (°)	Log_2	−0.020	0.236	0.93
Fixated dROI (binary)	NA (categorical)	1.499	1.135	0.25
Time to first dROI fixation (s)	Log_2	−0.059	0.067	0.42
Fixation count on dROI (n)	Untransformed	<0.001	<0.001	0.83
Fixation duration dROI (ms)	Log_2	−0.209	0.259	0.36

5.3 *Descriptive Data from Residents Versus Faculty*

Table 12 details subjective responses to histology form questions.

Table 12 Descriptive data from residents versus faculty subjective responses to histology form questions regarding case difficulty, confidence in interpretation, whether the case might be borderline between two diagnoses, and whether they would seek a second opinion.

	First-year residents ($N = 19$)		Second-year residents ($N = 25$)		Third-year residents ($N = 17$)		Fourth-year residents ($N = 9$)		Faculty ($N = 20$)	
Variable	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev
Case difficulty rating (1 to 6)	3.34	1.21	3.17	1.25	3.37	1.38	3.53	1.11	2.84	1.57
Confidence rating (1 to 6)	6.41	3.19	6.99	3.94	6.68	3.09	5.49	1.66	4.73	1.32
Proportion deemed borderline	0.29	0.45	0.29	0.45	0.29	0.45	0.31	0.46	0.33	0.47
Proportion second opinion	0.52	0.50	0.53	0.50	0.41	0.49	0.59	0.49	0.44	0.50

Disclosures

The authors declare that they have no competing interests.

Acknowledgments

We 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/. We also wish to thank the staff and residents at the various pathology training programs across the U.S. for their participation and assistance in this study. Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under award number R01 CA225585, R01 CA172343, R01 CA140560. 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. Researchers are independent from funders, and the funding agency played no role in study design, conduct, analysis or interpretation. JGE, DLW, TTB, KFK, and TD conceived and designed this study. TTB and TD collected the data. HS managed

pathologist recruitment and scheduling and provided feedback and suggestions on manuscript drafts. TTB drafted the manuscript and KFK performed the statistical analyses. TD, KFK, DLW, and JGE reviewed the manuscript at several times during preparation, providing feedback and suggestions. All authors read and approved the final manuscript. All participants provide written informed consent in accordance with Institutional Review Board approvals granted by the University of California Los Angeles.

Codes, Data, and Materials Availability

Due to the highly specialized nature of participant expertise and therefore increasing risk of identifiable data, we have decided not to make our data available in a repository. In the interest of minimizing the risk of participant identification, we will distribute study data on a case-by-case basis. Interested parties may contact Hannah Shucard at the University of Washington with data requests: hshucard@uw.edu

References

1. T. Brunyé et al., "A review of eye tracking for understanding and improving diagnostic interpretation," *Cognit. Res.* **4**, 7 (2019).
2. M. O. Al-Moteri et al., "Eye tracking to investigate cue processing in medical decision-making: a scoping review," *Comput. Human Behav.* **66**, 52–66 (2017).
3. H. L. Kundel and C. F. Nodine, "A short history of image perception in medical radiology," in *The Handbook of Medical Image Perception and Techniques*, E. Samei and E. A. Krupinski, Eds., pp. 9–20, Cambridge University Press, London (2010).
4. C. F. Nodine and H. L. Kundel, "The cognitive side of visual search in radiology," in *Eye Movements: From Psychology to Cognition*, J. K. O'Regan and A. Levy-Schoen, Eds., pp. 572–582, Elsevier Science, Amsterdam, Netherlands (1987).
5. H. Ashraf et al., "Eye-tracking technology in medical education: a systematic review," *Med. Teach.* **40**, 62–69 (2018).
6. K. Blondon and C. Lovis, "Use of eye-tracking technology in clinical reasoning: a systematic review," in *Digital Healthcare Empowering Europeans*, R. Cornet et al., Eds., pp. 90–94, European Federation for Medical Informatics, Madrid, Spain (2015).
7. A. Fogarasi et al., "Improving seizure recognition by visual reinforcement," *Neurol. Psychiatr. Brain Res.* **18**, 1–7 (2012).
8. D. J. Manning, S. C. Ethell, and T. Donovan, "Detection or decision errors? Missed lung cancer from the posteroanterior chest radiograph," *Br. J. Radiol.* **77**, 231–235 (2004).
9. M. S. Cain, S. H. Adamo, and S. R. Mitroff, "A taxonomy of errors in multiple-target visual search," *Vision Cognit.* **21**, 899–921 (2013).
10. M. Sibbald et al., "Why verifying diagnostic decisions with a checklist can help: insights from eye tracking," *Adv. Heal. Sci. Educ.* **20**, 1053–1060 (2015).
11. E. S. M. Tiersma et al., "Visualising scanning patterns of pathologists in the grading of cervical intraepithelial neoplasia," *J. Clin. Pathol.* **56**, 677–680 (2003).
12. E. Mercan et al., "Characterizing diagnostic search patterns in digital breast pathology: scanners and drillers," *J. Digital Imaging* **31**, 32–41 (2018).
13. T. Drew et al., "Scanners and drillers: characterizing expert visual search through volumetric images," *J. Vision* **13**, 3 (2013).
14. T. Balslev et al., "Visual expertise in paediatric neurology," *Eur. J. Paediatr. Neurol.* **16**, 161–166 (2012).
15. T. T. Brunyé et al., "Eye movements as an index of pathologist visual expertise: a pilot study," *PLoS One* **9**, e103447 (2014).
16. H. Matsumoto et al., "Where do neurologists look when viewing brain CT images? An eye-tracking study involving stroke cases," *PLoS One* **6**, e28928 (2011).
17. L. Cooper et al., "Radiology image perception and observer performance: how does expertise and clinical information alter interpretation? Stroke detection explored through eye-tracking," *Proc. SPIE* **7263**, 72630K (2009).

18. N. A. Giovinco 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.* **54**, 382–391 (2015).
19. G. Wood et al., "Visual expertise in detecting and diagnosing skeletal fractures," *Skeletal Radiol.* **42**, 165–172 (2013).
20. D. Manning et al., "How do radiologists do it? The influence of experience and training on searching for chest nodules," *Radiography* **12**, 134–142 (2006).
21. C. F. Nodine et al., "Time course of perception and decision making during mammographic interpretation," *Am. J. Roentgenol.* **179**, 917–923 (2002).
22. H. L. Kundel et al., "Using gaze-tracking data and mixture distribution analysis to support a holistic model for the detection of cancers on mammograms," *Acad. Radiol.* **15**, 881–886 (2008).
23. H. L. Kundel et al., "Holistic component of image perception in mammogram interpretation: gaze-tracking study," *Radiology* **242**, 396–402 (2007).
24. H. Haider and P. A. Frensch, "Eye movement during skill acquisition: more evidence for the information-reduction hypothesis," *J. Exp. Psychol.* **25**, 172–190 (1999).
25. D. Litchfield et al., "Viewing another person's eye movements improves identification of pulmonary nodules in chest x-ray inspection," *J. Exp. Psychol. Appl.* **16**, 251–262 (2010).
26. H. Jarodzka et al., "Conveying clinical reasoning based on visual observation via eye-movement modelling examples," *Instr. Sci.* **40**, 815–827 (2012).
27. A. Gegenfurtner, E. Lehtinen, and R. Säljö, "Expertise differences in the comprehension of visualizations: a meta-analysis of eye-tracking research in professional domains," *Educ. Psychol. Rev.* **23**, 523–552 (2011).
28. E. M. Kok et al., "Looking in the same manner but seeing it differently: bottom-up and expertise effects in radiology," *Appl. Cognit. Psychol.* **26**, 854–862 (2012).
29. T. Tien et al., "Eye tracking for skills assessment and training: a systematic review," *J. Surg. Res.* **191**, 169–178 (2014).
30. N. Oster 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 Women's Heal.* **13**, 3 (2013).
31. J. G. Elmore et al., "Diagnostic concordance among pathologists interpreting breast biopsy specimens," *JAMA* **313**, 1122 (2015).
32. Ventana Medical Systems Inc., "iScan Coreo Au Product Page," 2012.
33. T. Onega et al., "The diagnostic challenge of low-grade ductal carcinoma *in situ*," *Eur. J. Cancer* **80**, 39–47 (2017).
34. K. H. Allison et al., "Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel," *Histopathology* **65**, 240–251 (2014).
35. I. T. C. Hooge et al., "Gaze tracking accuracy in humans: one eye is sometimes better than two," *Behav. Res. Methods* **51**, 2712–2721 (2019).
36. J. V. Guttag, *Introduction to Computation and Programming Using Python with Application to Understanding Data*, MIT Press, Cambridge, Massachusetts (2016).
37. E. A. Krupinski et al., "Eye-movement study and human performance using telepathology virtual slides. Implications for medical education and differences with experience," *Hum. Pathol.* **37**, 1543–1556 (2006).
38. C. Mello-Thoms et al., "Effects of lesion conspicuity on visual search in mammogram reading," *Acad. Radiol.* **12**, 830–840 (2005).
39. G. Aston-Jones and J. D. Cohen, "An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance," *Annu. Rev. Neurosci.* **28**, 403–450 (2005).
40. T. Drew and L. H. Williams, "Simple eye-movement feedback during visual search is not helpful," *Cognit. Res.* **2**, 44 (2017).
41. T. Drew, S. E. P. Boettcher, and J. M. Wolfe, "One visual search, many memory searches: an eye-tracking investigation of hybrid search," *J. Vision* **17**, 5 (2017).
42. E. M. Reingold et al., "Visual span in expert chess players: evidence from eye movements," *Psychol. Sci.* **12**, 48–55 (2001).
43. A. K. Ericsson et al., *The Cambridge Handbook of Expertise and Expert Performance*, Cambridge University Press, Cambridge (2006).

44. J. M. Findlay and I. D. Gilchrist, *Active Vision: The Psychology of Looking and Seeing*, Oxford University Press, Oxford, England (2008).
45. K. Rayner, "Eye movements in reading and information processing: 20 years of research," *Psychol. Bull.* **124**, 372–422 (1998).
46. H. L. Kundel and C. F. Nodine, "Studies of eye movements and visual search in radiology," in *Eye Movements and the Higher Psychological Processes*, J. W. Senders, D. F. Fisher, and R. A. Monty, Eds., pp. 317–327, Lawrence Erlbaum Associates, Hillsdale, New Jersey (1978).
47. L. Pantanowitz, A. V. Parwani, and W. E. Khalbuss, "Digital imaging for cytopathology: are we there yet?" *Cytopathology* **22**, 73–74 (2011).
48. A. Donnelly et al., "Optimal *z*-axis scanning parameters for gynecologic cytology specimens," *J. Pathol. Inf.* **4**, 38 (2013).
49. R. Singh et al., "Standardization in digital pathology: supplement 145 of the DICOM standards," *J. Pathol. Inf.* **2**, 23 (2011).
50. S. H. Adamo et al., "Mammography to tomosynthesis: examining the differences between two-dimensional and segmented-three-dimensional visual search," *Cognit. Res.* **3**, 17 (2018).
51. A. Lesgold et al., "Expertise in a complex skill: diagnosing x-ray pictures," in *The Nature of Expertise*, M. T. H. Chi, R. Glaser, and M. J. Farr, Eds., pp. 311–342, Lawrence Erlbaum Associates, Hillsdale, NJ (1988).
52. H. Jarodzka et al., "In the eyes of the beholder: how experts and novices interpret dynamic stimuli," *Learn. Instr.* **20**, 146–154 (2010).
53. K. S. Berbaum et al., "Gaze dwell times on acute trauma injuries missed because of satisfaction of search," *Acad. Radiol.* **8**, 304–314 (2001).
54. P. A. Frewen et al., "Selective attention to threat versus reward: meta-analysis and neural-network modeling of the dot-probe task," *Clin. Psychol. Rev.* **28**, 307–337 (2008).
55. R. Engbert and R. Kliegl, "Microsaccades uncover the orientation of covert attention," *Vision Res.* **43**, 1035–1045 (2003).
56. B. D. Corneil et al., "Neuromuscular consequences of reflexive covert orienting," *Nat. Neurosci.* **11**, 13–15 (2008).
57. J. D. Edwards et al., "Systematic review and meta-analyses of useful field of view cognitive training," *Neurosci. Biobehav. Rev.* **84**, 72–91 (2018).
58. J. R. Kogan et al., "Opening the black box of clinical skills assessment via observation: a conceptual model," *Med. Educ.* **45**, 1048–1060 (2011).
59. E. S. Holmboe, L. Edgar, and S. Hamstra, "The milestones guidebook," 2016, https://www.acgme.org/Portals/0/MilestonesGuidebook.pdf.
60. R. Aggarwal and A. Darzi, "Technical-skills training in the 21st century," *N. Engl. J. Med.* **355**, 2695–2696 (2006).
61. S. Sadasivan et al., "Use of eye movements as feedforward training for a synthetic aircraft inspection task," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, pp. 289–300 (2005).
62. D. Nalanagula, J. S. Greenstein, and A. K. Gramopadhye, "Evaluation of the effect of feed-forward training displays of search strategy on visual search performance," *Int. J. Ind. Ergon.* **36**, 289–300 (2006).
63. R. Stein and S. E. Brennan, "Another person's eye gaze as a cue in solving programming problems," in *Proc. 6th Int. Conf. Multimodal Interfaces*, pp. 9–15 (2004).
64. L. Mason, P. Pluchino, and M. C. Tornatora, "Eye-movement modeling of integrative reading of an illustrated text: effects on processing and learning," *Contemp. Educ. Psychol.* **41**, 172–187 (2015).
65. L. H. Williams and T. Drew, "What do we know about volumetric medical image interpretation? A review of the basic science and medical image perception literatures," *Cognit. Res.* **4**, 21 (2019).
66. M. C Hout and S. D. Goldinger, "Target templates: the precision of mental representations affects attentional guidance and decision-making in visual search," *Attention Perception Psychophys.* **77**, 128–149 (2015).
67. B. M. Geller et al., "Second opinion in breast pathology: policy, practice and perception," *J. Clin. Pathol.* **67**, 955–960 (2014).
68. D. B. Nagarkar et al., "Region of interest identification and diagnostic agreement in breast pathology," *Mod. Pathol.* **29**, 1004 (2016).

Tad T. Brunye, PhD, is a visiting associate professor at Tufts University, scientific manager at the Center for Applied Brain and Cognitive Sciences, and a senior cognitive scientist at the U.S. Army Combat Capabilities Development Command Soldier Center (CCDC SC). He received his doctorate in experimental cognitive psychology from Tufts University in 2007. He is an expert in observer performance, including perception, comprehension, and decision-making, and the cognitive and neuroscientific mechanisms underlying success and failure in medical interpretation.

Trafton Drew, PhD, is an assistant professor at the University of Utah and director of the Applied Visual Attention Laboratory. He received his doctorate in cognitive psychology at the University of Oregon in 2009. He is an expert in visual search and attention, and uses a variety of neuroscientific and psychophysical methods to characterize and understand observer performance.

Kathleen F. Kerr, PhD, is a professor of biostatistics at the University of Washington, and director of the Summer Institute in Statistics for Clinical and Epidemiological Research. She received her doctorate in statistics from the University of California Los Angeles in 1999, and is an expert in risk prediction models, biomarker evaluation, statistical genetics and genomics, and the design and analysis of experiments in the health sciences.

Hannah Shucard, MS, is a research coordinator in the Department of Biostatistics at the University of Washington. She is engaged in developing and coordinating research projects across multidisciplinary and multicenter teams.

Donald L. Weaver, MD, is a professor in the Department of Pathology and Laboratory Medicine at the University of Vermont Larner College of Medicine, attending pathologist at UVM Medical Center, and the medical director of the University of Vermont Cancer Center Biobank. He received his MD degree from the University of Vermont in 1984 and completed residency and fellowship training in surgery, anatomic and clinical pathology, and analytical cytometry. He is a leading expert in breast pathology.

Joann G. Elmore, MD, is a professor of medicine in the Division of General Internal Medicine, in the Department of Medicine at the David Geffen School of Medicine at UCLA, holds the Rosalinde and Arthur Gilbert Foundation Endowed Chair in Health Care Delivery and serves as Director of the UCLA National Clinician Scholars Program. She received her MD from Stanford University, MPH from Yale University, and completed residency and fellowships in internal medicine. She is a national leader in academic general internal medicine and has a distinguished career as an investigator, mentor, administrator and educator.