Title

Pathologist pupil dilation reflects experience level and difficulty in diagnosing medical images.

Journal

Journal of Medical Imaging, 10(2)

ISSN

2329-4302

Authors

Drew, Trafton
Konold, Catherine
Lavelle, Mark
et al.

Publication Date

2023-03-01

DOI

10.1117/1.JMI.10.2.025503

Pathologist pupil dilation reflects experience level and difficulty in diagnosing medical images

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

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

Abstract **Purpose**: Digital whole slide imaging allows pathologists to view slides on a computer screen instead of under a microscope. Digital viewing allows for real-time monitoring of pathologists' search behavior and neurophysiological responses during the diagnostic process. One particular neurophysiological measure, pupil diameter, could provide a basis for evaluating clinical competence during training or developing tools that support the diagnostic process. Prior research shows that pupil diameter is sensitive to cognitive load and arousal, and it switches between exploration and exploitation of a visual image. Different categories of lesions in pathology pose different levels of challenge, as indicated by diagnostic disagreement among pathologists. If pupil diameter is sensitive to the perceived difficulty in diagnosing biopsies, eye-tracking could potentially be used to identify biopsies that may benefit from a second opinion.

Approach: We measured case onset baseline-corrected (phasic) and uncorrected (tonic) pupil diameter in 90 pathologists who each viewed and diagnosed 14 digital breast biopsy cases that cover the diagnostic spectrum from benign to invasive breast cancer. Pupil data were extracted from the beginning of viewing and interpreting of each individual case. After removing 122 trials (<10%) with poor eye-tracking quality, 1138 trials remained. We used multiple linear regression with robust standard error estimates to account for dependent observations within pathologists.

Results: We found a positive association between the magnitude of phasic dilation and subject-centered difficulty ratings and between the magnitude of tonic dilation and untransformed difficulty ratings. When controlling for case diagnostic category, only the tonic-difficulty relationship persisted.

Conclusions: Results suggest that tonic pupil dilation may indicate overall arousal differences between pathologists as they interpret biopsy cases and could signal a need for additional training, experience, or automated decision aids. Phasic dilation is sensitive to characteristics of biopsies that tend to elicit higher difficulty ratings and could indicate a need for a second opinion. © *2023 Society of Photo-Optical Instrumentation Engineers (SPIE)* [DOI: 10.1117/1.JMI.10.2.025503]

Keywords: pathology; pupillometry; expertise; diagnostic accuracy; medical image perception; decision-making.

Paper 22333GR received Nov. 27, 2022; revised manuscript received Mar. 26, 2023; accepted for publication Apr. 10, 2023; published online Apr. 22, 2023.

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

1 Introduction

Cancer diagnosis relies on visual inspection of tissue sections by a pathologist; this process is prone to diagnostic ambiguity that can have significant adverse consequences for patients, including medical errors. Although pathologic criteria for cancer and its precursors are defined, the subjective assessment of these criteria results in high degrees of discordance among pathologists and, in some cases, diagnostic errors.[1–3] Our group has documented significant discordance in the diagnosis of pre-invasive and invasive breast carcinoma and melanoma.[1,2] These errors can influence major treatment decisions (e.g., surgery, radiation, chemotherapy), which adversely impacts patient care and outcomes. Thus, there is great value in identifying even subtle signs associated with incorrect diagnoses.

The advent of digital pathology whole slide imaging (WSI) has enabled pathologists to begin transitioning from microscope-based viewing to computer monitor viewing. This Food and Drug Administration-approved method of viewing has ushered in a revolution in our understanding of how pathologists examine histology by allowing researchers to precisely measure how pathologists navigate through the enormous search space (generally $<1,00,000 \times 50,000$ pixels) for each case.[4] Our group and others are taking advantage of this new technology to perform some of the first eye-tracking investigations of pathologists as they examine digital pathology slides that can be dynamically zoomed and panned.[5–9] Although there is a large amount of literature devoted to understanding search techniques in radiology,[10,11] it is increasingly clear that what has been learned for radiology may not be valid for pathology. Although performance in the two different domains overlaps in important ways (i.e., searching for clinically-relevant, difficult-to-find targets in a complex search environment), these are fundamentally distinct tasks that are associated with different search strategies. For example, our group has recently shown in a large sample of pathologists that scanning within a given magnification level of the pathology image was associated with higher diagnostic accuracy but zooming quickly through different magnification levels was not.[12] This result is in stark contrast to radiology, in which moving quickly through depth was associated with higher accuracy in lung nodule detection than scanning.[13,14] This pattern of results highlights the gap in our understanding of how visual search processes adapt in a domain-specific manner and how they are accomplished in pathology.

In this paper, we are expanding upon the growing understanding of how pathologists search digital pathology slides by shifting our focus from eye movements that occur during clinical examination to the pupil response during the initial view of the slide. Pupil diameter is sensitive to both environmental (especially lighting conditions) and psychological processes. A large and growing body of work demonstrates that the pupil is sensitive to mental effort or engagement, a particular interest in the context of pathology image diagnosis. Below, we outline some of the relevant evidence for the connection between these mental constructs and pupil diameter. Several excellent recent review articles provide a more in-depth treatment of this topic.[15–17]

For over 50 years, it has been understood that pupil dilation is a meaningful and reliable correlate of cognitive processes.[18–21] Changes in pupil dilation reflect cognitive function starting in infancy.[22] Prior work has shown that pupil diameter tends to increase in response to interesting stimuli, unexpected stimuli, sexually arousing stimuli, and tasks that are cognitively difficult.[20–23] More recent work has distinguished between ongoing (tonic) activity and event-related (phasic) activity as relating to distinct cognitive processes.[16] Increased tonic pupil diameter throughout a task is thought to relate to increased mental effort or cognitive load associated with the task. Tonic differences in pupil dilation are thought to reflect the broad attentional state associated with a task.[24] This is in contrast to phasic amplitude, which is associated with task-related activity and more sensitive to moment-to-moment fluctuations that occur in response to specific stimuli.[25] For instance, studies have shown that effortful decisions are associated with a transient increase in pupil size that is sensitive to the difficulty of the decision.[21,26] In a difficult target detection task, de Gee et al.[26] found that phasic changes in pupil diameter predicted whether the target was detected, suggesting that the pupil may be a valuable window into what reaches conscious awareness. In this paper, we examine the association between changes in pupil diameter for pathologists viewing WSI in combination with self-reported diagnosis confidence ratings and diagnostic accuracy.

Our group previously conducted a pilot feasibility study that examined changes in tonic and phasic pupil amplitude as pathologists examined digitized breast biopsies.[27] We found that a larger tonic pupil diameter was associated with cases that were rated as more subjectively difficult. In addition, we found that the phasic response occurring immediately after fixating a key region of interest (ROI) containing important diagnostic information varied as a function of the difficulty of the case and whether pathologist ultimately made a diagnosis that was concordant or discordant with the consensus diagnosis for that case. These results could have important implications for potential training protocols for future pathologists. If phasic changes are reliably indicative of whether a critical ROI is detected and properly diagnosed, a training system might be able to use this signal to adaptively provide real-time feedback to help improve didactic training protocols.

This work is extending that prior work with a new and much larger sample and a more varied sample of participating pathologists ranging from first-year residents training in pathology to attending physicians. This study represents the first year of data collection for a larger longitudinal study that will examine changes in expertise development as residents progress through their pathology training. As such, although the current sample includes a broad range of experience, in this first investigation, we focus on a cross-sectional examination of pathology residents in different years of training, compared with more senior attending physicians. Our analyses focused on time-locking our pupil analyses to the first onset of the case image. This approach allows for examining phasic responses to case types as well as tonic responses that are more task general. Following prior pilot work from our group, we examined pupillary responses time-locked to the first moment that the eyes fixated within ROIs containing important diagnostic information. However, we did not find evidence of associations between this response and expertise, diagnostic accuracy, or subjective difficulty ratings. We discuss the methodological differences that may have contributed to this lack of an effect in the discussion.

2 Methods

2.1 *Participants*

We collected data from 92 pathologists ($n = 20$ attendings; $n = 72$ pathology trainees) recruited from 9 academic medical centers across the United States as part of a larger longitudinal study. The mean ages of junior residents, senior residents, and faculty are as follows: 32.6 years, 33.4 years, and 47.1 years. A contact person at each site introduced the study to their attending physicians and trainees and provided contact information for potential participants, but they were not otherwise involved in data collection. Participants were invited via email (maximum of four attempts). To be eligible, trainees had to be in an anatomic or combined anatomic and clinical pathology residency training program or related fellowship, available during the one- or two-day site visits arranged for data collection, and willing to interpret breast biopsy cases. Approval was obtained from the appropriate Institutional Review Boards (IRBs), with University of California, Los Angeles (UCLA) acting as the IRB of record. All participants provided informed consent and received a $50 gift card for their involvement.

2.2 *Stimulus Creation/Case Selection*

Cases for this study were identified from a larger breast pathology study (B-Path) aimed at understanding diagnostic variability in interpreting breast biopsies, which has been described in detail.[2] Each of the 240 B-Path cases has standardized clinical data, comprehensive consensus-defined reference standard diagnosis, a high-quality whole side digital image, and data on previous interpretations by >200 practicing US pathologists.

Cases for this study were selected from the B-Path cases based on diagnoses gathered from prior interpretations by experienced pathologists (those who were fellowship-trained in breast pathology and/or considered by their peers to be an expert; $n = 54$ pathologists). Breast pathology cases were classified into five diagnostic categories of increasing severity: benign, atypia, low- and high-grade ductal carcinoma *in situ* (DCIS), and invasive. These categories were associated with different histopathological features, treatment and surveillance options, and prognoses.[28] We identified 32 cases to include in this study: 4 benign, 10 atypia, 10 low-grade DCIS (LGDCIS), 4 high-grade DCIS (HGDCIS), and 4 invasive breast carcinoma. We divided these cases into 3 sets of 14 cases each (5 of the cases were the same across all sets and the remaining nine cases were unique to each test set). Each set included two benign cases, four atypia, four LGDCIS, two HGDCIS, and two invasive cases. Cases were divided into three sets because this study was part of a larger longitudinal study in which residents are asked to evaluate one of the three sets per year during three residency training years. To provide a practice case for all pathologists, one invasive carcinoma case with high diagnostic concordance of pathologists with the expert consensus reference diagnosis (93%) was selected from the larger set of B-Path cases. The diagnostic concordance of the practice case was established when interpreted in the B-Path study using the standard glass slide format.[29]

Images were shown on a digital slide viewer developed for prior work from our group,[4] using Microsoft Silverlight and Deep Zoom tools (Microsoft, Inc., Redmond, Washington, United States). The viewer displayed images in a navigable viewport that allows for zooming (range 1× to 60× magnification) and panning while maintaining full image resolution. Participant behaviors (e.g., current view position and zoom level) were logged by the viewport at 10 Hz and eye movements at 250 Hz were subsequently overlaid by co-registering the viewport and eye-tracking data. Both data streams contain precise time logs, allowing for alignment in a common time axis using a next-closest data point approach to account for varied sampling frequency. Then, we equated two temporally aligned viewports in vector length by up-sampling viewport data and repeating the last-known value of the stream until we found a new value.

2.3 *Eye-Tracking*

We used the mobile remote eye-tracking device (RED-m) system. This was a noninvasive and portable eye-tracking system manufactured by SensoMotoric Instruments (SMI, Boston, Massachusetts, United States). The system used an array of infrared lights and cameras to track eye position at 250 Hz with high gaze position accuracy (0.4 deg) and a nine-point calibration. For data collection, we mounted the RED system to the bottom of a color-calibrated 24″ Dell U2417H Ultrasharp liquid crystal display (1920×1080 resolution) computer monitor. Participants were seated ∼65 cm from the monitor and received feedback from the experimenter if they moved closer than 54 or farther than 77 cm from the eye-tracker. Eye-tracking data were lost due to technical errors for 2 of the 92 participants.

2.4 *Pupillometry*

In addition to gaze position, the eye-tracker records pupil diameter in millimeters at 0.1 mm resolution. All signal-processing and statistical analyses of pupil data were performed in R version 1.3.1056.[30] Pupil diameter was extracted for each case in a 35-s window encompassing the appearance of the biopsy on the screen, as indicated by the viewport. A 5-s baseline preceding the appearance of the biopsy was selected. Baseline correction, described below, used data between −5 and −4 s relative to biopsy appearance. The screen luminance was stable during this period, and as can be seen from the average pupil traces (e.g., Fig. 1), the pupil diameter was also stable. In the 4 s preceding the biopsy onset (from −4 to 0 s), a new tab opened on screen with a blank white background, followed by a black background, leading to constriction and dilation, which precluded the use of this period for baseline correction. This is why we used the period from −5 to −4 s preceding biopsy onset instead as our baseline correction period for pupil diameter.

Pupil samples with corresponding gaze data suggestive of eye-tracker malfunction and samples with a diameter reported as 0 mm were treated as missing. The eye-tracking software itself also detected eye-blinks and momentary failures in locating the eyes. Samples from both events were removed. Due to slight irregularities in inter-sample-intervals, data were linearly interpolated within the 35-s epochs. This ensured that each pupil epoch contained 8750 evenly spaced samples. In the case in which a sample was to be interpolated between a missing sample and a valid sample, the interpolated sample was also treated as missing. Following this first

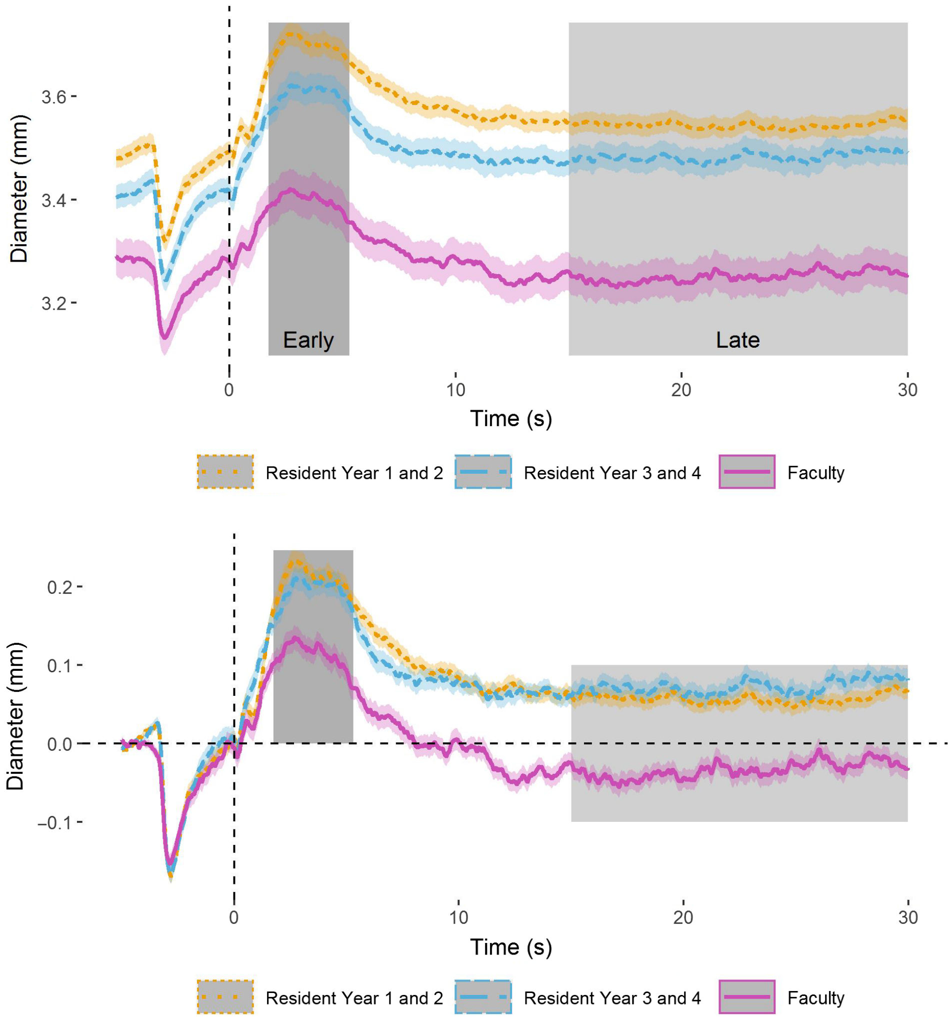

Fig. 1 Uncorrected (raw, upper) and baseline-corrected (lower) pupil waveforms separated by experience level. The dark gray box (1.75 to 5.32 s) represents the "early" window, and the light gray box (15 to 30 s) represents the "late" time window over which mean amplitude was computed. Lighter shading around waveforms represents the 95% confidence interval (CI).

interpolation step, a dilation-acceleration algorithm also marked samples as missing when the change in the change of pupil diameter was too fast. The acceleration outlier threshold was defined as 4.5 × the median absolute deviation of all acceleration measurements for an epoch. An additional blink-detection and removal algorithm was then applied, which identified segments of missing data exceeding 76 ms and then removed an additional 52 ms before and after the segment. Any epoch that contained more than 50% missing samples at this stage was discarded from further analyses, leading to the loss of 78 (6.2%) epochs. The maximum stretch of missing samples for each epoch was brief, overall. The longest stretch was shorter than 1 s for 75% of cases. Only 13 cases had stretches longer than 5 s, and the results (below) did not depend on whether these cases were excluded. Missing samples were replaced via linear interpolation between valid samples. In the case in which the first or last sample of an epoch was missing, the closest valid adjacent sample was extrapolated to the beginning or end of the epoch, respectively. Following replacement of missing samples, the data were low-pass filtered at 10 Hz with a second-order Butterworth filter to attenuate rapid variations in pupil diameter that were likely artificial. In addition to epochs rejected due to suboptimal eye-tracking, an additional 44 (3.5%) trials were lost due to malfunctions of the viewport. A total of 1138 epochs were retained and submitted to subsequent analyses.

Four pupil diameter measurements were taken from each epoch: a corrected and uncorrected measurement from both an "early" and a "late" window. The two measurement windows were defined based on inspection of the grand-average uncorrected pupil waveform. A prominent dilation was visible roughly 3 s after biopsy appearance. We defined the window surrounding this peak by first recording the maximum and minimum diameter following biopsy onset. The leading edge of this early measurement window was defined as the first moment when the diameter was at least 75% as large as the maximum diameter, where 0% represents the minimum diameter. The tail end of the window was similarly defined as the last moment when the diameter was at least 75% as large as the maximum diameter. For each epoch, the early diameter submitted to analyses represents the mean within this window, from 1.75 to 5.32 s after biopsy appearance. The late measurement window was intended to capture the apparent stable-point of pupil diameter during biopsy review, when the average waveform appeared to level out. The mean diameter in the period spanning 15 to 30 s following biopsy onset was extracted for each epoch and submitted to analyses. Analyses of the late diameter include 1131 trials because pathologists finished reviewing the cases faster than 30 s in 7 instances. Baseline correction simply entailed subtracting the mean from the baseline period (from −5 to −4 s) from the early and late diameter measurements. Luminance was measured using the co-registered viewport and eye-tracking data. By combining the coordinates of fixation points and the zoom level of the viewer, we generated luminance values at 4 different levels of visual angle (1 deg, 3 deg, 5 deg, and 10 deg) using the hue, saturation, value representation of digital color.[31] This enabled us to account for luminance as a factor in pupil size.

2.5 *Procedures*

Prior to reviewing the cases, participants completed an online consent form and online baseline survey that gathered information on their level of experience interpreting breast biopsies and their perceptions of breast pathology and digital WSI. Participants reviewed cases while seated in a private room with the experimenter during a scheduled one-hour appointment at the participating site. Participants completed eye-tracker calibration and were then shown how to use the image viewer and completed the standard practice case to ensure that they were comfortable with the procedure. Each participant then viewed 1 of the 3 subsets of 14 cases, at full screen. Case viewing order was randomized across participants. After viewing each case, they completed a diagnostic histology form in which they indicated their final diagnosis, whether the case was borderline between two diagnoses, and whether they would want a second opinion from another pathologist, and then they rated their perceived diagnostic difficulty of the case and confidence in their assessment. To assist with instructions, eye-tracker calibration, and troubleshooting, at least one author (T.D. or T.B.) was present for all data collections. This also allowed for ensuring that the viewing environment (e.g., room lighting, participant positioning) was as similar across the nine collection sites as possible. All participants used one of two exact replicas of the experimental apparatus, including the same computer, monitor, keyboard, mouse, and eye-tracker models; the two Dell monitors were factory color calibrated with highly similar color calibration reports.

3 Results

3.1 *Experience*

Our overarching aim was to assess the predictive capacity of pupillometry for objective measures of diagnostic performance and competence. We tested the association of pupil diameter with two behavioral measures: diagnostic accuracy and interpretive duration. Before directly addressing these questions, it is necessary to characterize the influence of experience level on pupil responses and diagnostic performance. For a simple check of the effect of experience on accuracy, we first obtained a measure of mean accuracy for each pathologist, which is a continuous measure theoretically ranging from 0 to 1. We dummy-coded the experience level such that the reference group was first- and second-year junior resident pathologists (n cases = 541, n pathologists = 44), and the comparison groups were third- and fourth-year senior residents (n cases = 353, n pathologists = 26) and faculty members (n cases = 244, n pathologists = 20). Using linear regression, we regressed each participant's mean accuracy (M) onto their experience group and described the slope of the regression line as Bf Faculty ($M = 64.2\% \pm 3.4\%$) were significantly more accurate than junior residents (years 1 and 2, $M = 41.3\% \pm 2.0\%$, $B = 23.0\%$, $p < 0.001$) but senior residents (years 3 and 4) were not significantly more accurate than junior residents ($M = 47.2\% \pm 3.1\%$, $B = 5.9\%$, $p = 0.062$).

All pupil analyses were adjusted for unique cases. A single case is the unit of observation for which pupillometry may provide the most clinical or educational benefit. We regressed pupil diameter from each separate case onto the three-level, dummy-coded experience variable, using general estimating equations (GEE[32]) to adjust standard errors for nonindependent observations within pathologists. The contrast of faculty versus junior residents was significant for all pupil measurements except early uncorrected (early-corrected: $B = -0.10$ mm, $p < 0.001$, late-corrected: $B = -0.09$ mm, $p < 0.001$, late-uncorrected: $B = -0.28$ mm, $p = 0.042$, and early-uncorrected: $B = -0.27$, $p = 0.069$ (see Table 1). Compared with first- and second-year pathologists' average early, late-corrected, and late-uncorrected diameters of 0.21 mm, 0.06 mm, and 3.54 mm, attending pathologists' pupil diameters were reduced by 0.10 , 0.09 , and 0.28 mm, respectively (Fig. 1). The contrast of third- and fourth-year to first- and second-year pathologists was not significant for any of the four pupil diameter measures: all ps > 0.42. To simplify subsequent analyses, we collapsed junior and senior residents into one group. In subsequent analyses, the experience factor represents a contrast of all residents versus attending pathologists.

3.2 *Subjective Case Difficulty Ratings*

In addition to the mediating relationship of experience level with diagnostic accuracy and pupil diameter, we sought to characterize the association between subjective difficulty of each case as perceived by the interpreting pathologist and pupil diameter. After submitting their diagnosis, pathologists were asked "please rate on the following scale your opinion of the level of diagnostic difficulty of this case" using a Likert scale rating from 1 to 6 ("very easy" to "very challenging"). In general, resident pathologists' average difficulty ratings were significantly higher compared with faculty ratings ($B = 0.49$, $p = 0.008$, $M_{\text{res}} = 3.35 \pm 0.63$, $M_{\text{fac}} = 2.89 \pm 0.81$).

3.3 *Difficulty*

Because individual pathologists in our study overwhelmingly tended to use variably restricted ranges on the 1 to 6 difficulty scale, we mean-centered difficulty ratings separately for each person before including them in GEE analyses. When controlling for experience group, the associations between the corrected and uncorrected early phasic response with centered-difficulty were significant (early corrected: $B = 0.02$ mm, $p = 0.002$, and early uncorrected:

Table 1 Pupil measures regressed onto individually centered difficulty ratings, individually centered luminance values, and experience group using GEE to account for nonindependent observations. Significant *p*-values are indicated with boldface, robust standard error (SE) reported.

		Estimate	Robust SE	Robust *Z*	*p*-value
Early corrected	Experience	−0.09	0.03	−3.74	<**0.001**
	Difficulty	0.02	0.01	3.04	**0.002**
Early uncorrected	Experience	−0.27	0.15	−1.81	0.069
	Difficulty	0.02	0.01	2.65	**0.008**
Late corrected	Experience	−0.10	0.02	−4.48	<**0.001**
	Difficulty	0.00	0.00	−0.37	0.710
Late uncorrected	Experience	−0.28	0.14	−2.03	**0.042**
	Difficulty	0.00	0.01	0.00	0.997

$B = 0.02$ mm, $p = 0.008$; Table 1). This suggests that, for two pathologists in the same experience group, the corrected and uncorrected early pupil responses were 0.02 mm greater for a pathologist who rated a case as one point above average difficulty compared with a pathologist who rated a case as average difficulty (Fig. 2). This difficulty pupil association was neither significant for late-corrected nor late-uncorrected pupil diameter, $ps > 0.34$. For corrected pupil measures, the contrast between faculty and resident pathologists was significant when controlling for centered difficulty (early: $B = -0.09$ mm, $p < 0.001$, and late: $B = -0.09$ mm, $p < 0.001$). The faculty-resident contrast did not reach significance for early uncorrected pupil measures when controlling for difficulty (early: $B = -0.27$ mm, $p = 0.069$). For late uncorrected pupil measures, faculty-resident contrast was significant (late: $B = -0.28$ mm, $p = 0.042$).

3.4 *Diagnostic Accuracy*

The above difficulty-pupil analyses were important for understanding the association between experience group and pupil diameter. However, we chose to first determine whether pupil diameter was associated with diagnostic accuracy without controlling for subjective difficulty. As mentioned above, we controlled for experience group when regressing pupil diameter onto diagnostic accuracy dummy coded with "incorrect" as the reference. Correctly diagnosing a case

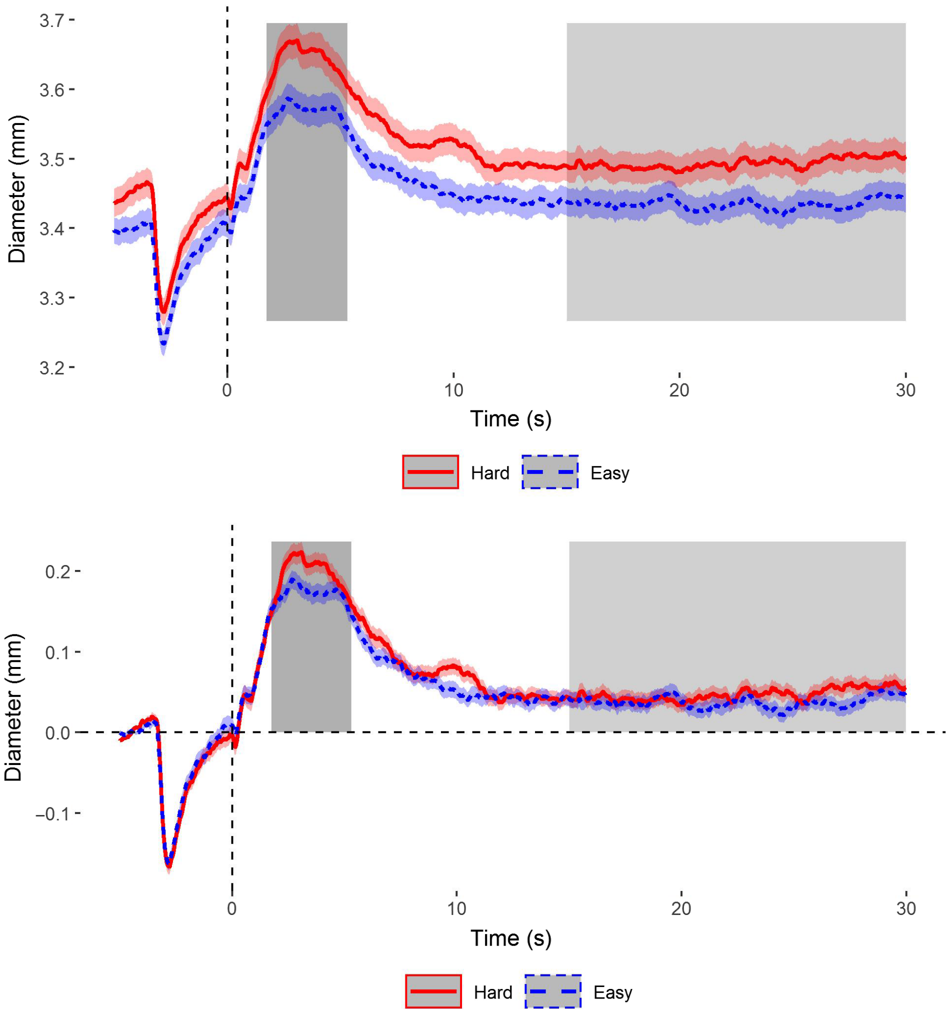

Fig. 2 Uncorrected (raw, upper) and baseline-corrected (lower) pupil waveforms separated by subject-centered difficulty ratings by quartile (easiest to hardest). The dark gray box (1.75 to 5.32 s) represents the "early" window, and the light gray box (15 to 30 s) represents the "late" time window over which mean amplitude was computed. Lighter shading around waveforms represents the 95% CI.

was associated with a significant change in pupil diameter for early-corrected pupil measurement alone (early-corrected: $B = -0.03$ mm, $p = 0.039$, early-uncorrected: $B = -0.01$ mm, $p = 0.68$, late-corrected: $B = -0.003$ mm, $p = 0.75$, and late-uncorrected: $B = 0.02$ mm, $p = 0.40$; Fig. 3). However, the significant effects of experience group persisted for all pupil measures except early uncorrected (early-corrected: $B = -0.09$ mm, $p =< 0.001$, late-corrected: $B = -0.10$ mm, $p < 0.001$, late-uncorrected: $B = -0.28$ mm, $p = 0.038$, and early-uncorrected: $B = -0.27$, $p = 0.07$). This result was surprising as the previous pilot study suggested that pupil diameter was associated with diagnostic accuracy, albeit in the context of an interaction with difficulty ratings and time-locked to fixation on a diagnostically-relevant ROI.[27] This pattern of results was consistent with pupil regressed simultaneously on difficulty and experience and added diagnostic accuracy as a third covariate.

3.5 *Diagnostic Efficiency*

Although our results suggest that pupil diameter during the initial 30 s of biopsy review may not be a useful predictor of diagnostic errors on a case-by-case basis, its utility as an indicator of diagnostic efficiency, which we operationalize as the speed with which a diagnosis was reached, is still undetermined. We measured interpretive duration on a case-by-case basis as the time elapsed between the appearance of the biopsy on screen and the moment when a pathologist

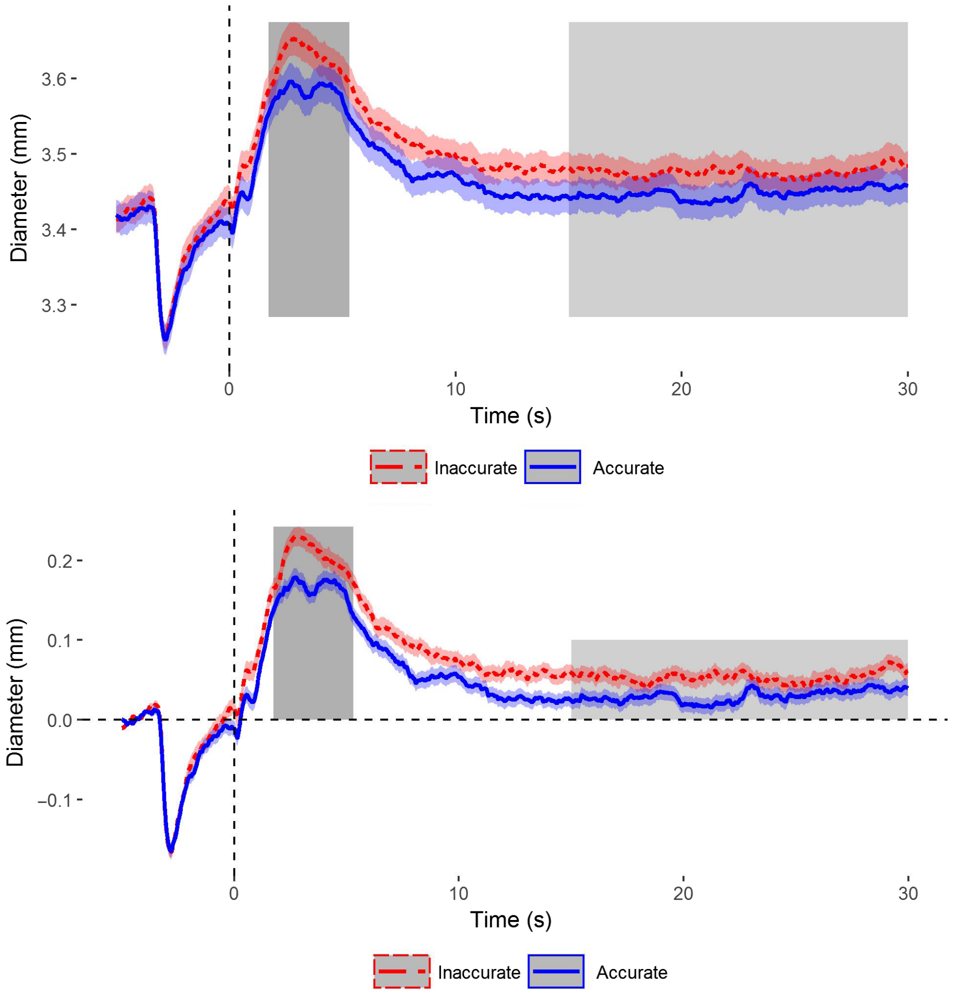

Fig. 3 Uncorrected (raw, upper) and baseline-corrected (lower) pupil waveforms separated by consensus accuracy. The dark gray box (1.75 to 5.32 s) represents the “early” window, and the light gray box (15 to 30 s) represents the “late” time window over which mean amplitude was computed. Lighter shading around waveforms represents the 95% CI.

clicked the button to close the digital slide viewer and begin the histology report. Fourteen additional cases were removed from the present analyses due to irregular interactions with the viewport (e.g., closing the case, then opening it again) that led to missing duration data, bringing the total to 1124 case interpretations. Pathologists spent an average of 113 ± 55 s reviewing each case. Considerable between-subject variability in both the average and range of interpretive duration prompted us to z-score the measure separately for each individual. We controlled for experience when assessing the pupil-interpretive duration relationships on a case-by-case basis, adjusting standard errors using GEE as before. None of the four pupil measurements were significantly associated with standardized interpretive duration when controlling for experience (early corrected: $B = 0.00$, $p = 0.44$, late corrected: $B = -0.00$, $p = 0.99$, early uncorrected: $B = 0.00$, $p = 0.77$, and late uncorrected: $B = -0.01$, $p = 0.64$; Fig. 4). The baseline-corrected and late uncorrected pupil diameters were still significantly smaller for faculty compared with residents when controlling for interpretive duration (early: $B = -0.09$, $p < 0.001$ and late: $B = -0.10$, $p < 0.001$), but the uncorrected early measure did vary reliably (early: $B = -0.27$, $p = 0.07$ and late: $B = -0.28$, $p = 0.042$). As with diagnostic accuracy, the results of regressing pupil measures onto difficulty and experience did not change when interpretive duration was added as a third covariate.

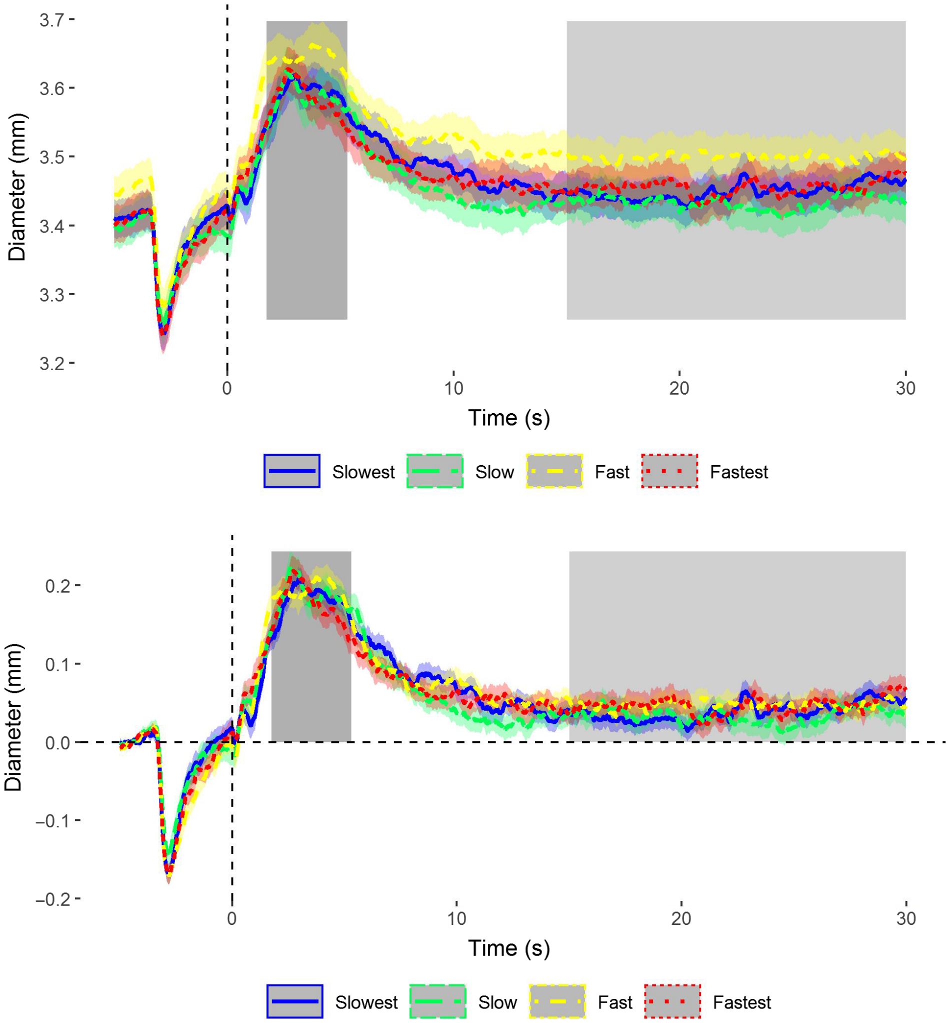

Fig. 4 Uncorrected and baseline-corrected pupil waveforms separated by subject-normalized interpretive duration. Duration was split into quartiles to allow for averaging and graphing. The dark gray box (1.75 to 5.32 s) represents the “early” window, and the light gray box (15 to 30 s) represents the “late” time window over which mean amplitude was computed. Lighter shading around waveforms represents the 95% CI.

4 Discussion

The purpose of this study was to evaluate whether associations between diagnostic difficulty, experience, and pupil dilations exist. If such associations were present, it would indicate that pupil dilation has the potential to be a meaningful predictor of performance when clinicians examine digital pathology slides. In a large sample of pathologists with varying experience levels recruited from nine medical centers across the United States, we found that pupil dilation time-locked to case onset was not significantly associated with diagnostic accuracy or efficiency. However, when controlling for pathologist experience level (or age), pupil dilation was significantly associated with individually-centered difficulty ratings, but only in the early time window. During these early moments, as the biopsy appears on the screen, the pupil reaches peak dilation, and this peak is modulated by pathologists' subjective impression of the difficulty in diagnosing this case relative to other cases. Because the review process is time consuming and highly variable across pathologists, time-locking to the early onset of the case may hinder detection of pupillary signals associated with the moment when a particular pathology was first detected. Thus, by focusing on the initial onset of the case, this study represents an in-depth examination of the pupillary response to the first impression of the case. Our data suggest that the initial exposure to the case evokes cognitive processing that scales with the perceived challenge in diagnosing the biopsy.

Following prior work in the pupillometry literature,[10] after cleaning, we separated our task-evoked pupil responses into early and late activity, which appear to correspond to different aspects of phasic activity. The pupil waveforms clearly exhibit a transient increase in pupil diameter 1 to 2 s after the onset of the case that we believe maps most clearly to a phasic response to processing new information.[23,24] Interestingly, pupil diameter does not immediately return back to baseline activity after this early peak. Rather, we found that pupil diameter remained elevated for at least 30 s after the initial onset of the case. Indeed, an exploratory follow-up analysis that time-locked to the completion of the case (rather than the onset of the case) confirmed that pupil diameter remained larger than during the baseline period. In addition to this "early" and "late" phasic activity found in baseline-corrected waveforms, we also observed ongoing differences in activity that were clearest in the uncorrected waveforms that were not baseline-corrected. We interpret this tonic activity to more task-general processes rather than specifically related to the case currently being examined. Although we examined uncorrected pupil diameter in the same "early" and "late" time windows used in the baseline-corrected waveforms, given that there is a clear deflection 1 to 2 s after the onset of the case, we focus our analyses of the uncorrected pupil amplitude on the "late" time window, which is therefore likely a better representation of task general tonic activity.

Overall, the largest effect in this examination is that experience was associated with large differences in both the uncorrected tonic and baseline-corrected late phasic pupil diameter. Less experienced pathologists generally had larger pupil diameter through these intervals, consistent with the task being associated with a higher cognitive workload. However, subjective difficulty ratings that we administered after each case challenged this interpretation; there was no compelling statistical evidence of a difference between subjective difficulty ratings from junior and senior residents ($p = 0.11$). It is certainly not surprising that the residents, who often had less than a year of experience reading digital pathology slides in a clinical setting, found this task to be difficult. Here, pupil diameter appears to function as biometric measure of difference in perceived overall difficulty of the task. This general trend extends to perceived difficulty of specific cases, even when controlling for experience level. Thus, our results suggest that the early phasic pupil response is indicative of perceived difficulty provided we know how difficult the pathologist in question finds the task more generally.

All of the major eye-tracking hardware companies automatically measure pupil diameter while tracking eye movements. As eye-tracking becomes less expensive and more readily available, easier to calibrate, and thus more likely to be installed in clinical reading rooms, they present a promising opportunity to examine this rich source of data that has been linked to the cognitive effort,[19,24] the mental workload,[33,34] and the locus coeruleus/norepinephrine[35,36] pathway in the brain. This potential application is particularly appealing in applied fields, such as pathology and radiology, in which errors can have dire consequences for patients. Pupil dilation may therefore serve as a marker of important cognitive constructs that help predict performance. Crucially, this signal is continuous and could theoretically serve as an online index of current cognitive workload. If this signal is robust, it might be a useful target for future work that could alert the user or a supervisor if the observer finds something about the case particularly challenging. One could imagine a system in which large increases in pupil diameter while evaluating a particular case could then be a trigger for a second read by a clinician. Having now worked with this large, rich pupil and eye-tracking dataset, we believe there are several reasons why this vision of using pupillometry in this manner is currently unrealistic. We delineate these concerns not to discourage future researchers from working toward this promising goal but to articulate challenges that need to be surmounted for ideas of this nature to live up to their promise.

The most significant challenge associated with interpreting pupil diameter is that variation in diameter is only minimally driven by cognitive processes in the context of a large and continuous adaptation to changing ambient and focal lighting conditions. Because of this, most early pupil work focused on paradigms in which the visual stimulation was held constant, whereas cognitive processing was manipulated using a different sensory modality, most often auditory cues (e.g., Refs. 19 and 24). More recently, researchers have begun to use visual paradigms, such as visual search[37] or multiple object tracking,[35] but researchers who have examined these paradigms often go to great lengths to ensure that luminance is matched between conditions of interest. A related challenge is that the apparent size of the pupil diameter varies with the location on the screen that is being fixated.[38] In a laboratory setting, a simple solution to this challenge is to not allow eye movements and focus on covert attentional shifts during the task. Clearly, it is not possible to manipulate the luminance of digital pathology slides to ensure a cleaner signal or to ask clinicians to avoid moving their eyes while interpreting a case. Thus, the pupil signal in these less controlled and more naturalistic settings will necessarily have a lower signal-to-noise ratio than in more tightly controlled settings. This is a persistent challenge for the translational value of laboratory findings using pupil diameter, though there is some promising analytical work to mitigate this challenge in the future.[39,40]

An additional challenge to applying traditional pupillometry techniques in this more applied setting is determining what event to time-lock the pupil waveforms to. In traditional pupillometry in a laboratory setting, there is a clear time course for when important events should occur. For instance, in a visual search paradigm, the cognitive processing associated with the task can be time-locked to the onset of the trial information and the processing associated with a given trial's decision (whether a target object is present or absent), or it can be time-locked to the moment when that decision is rendered by the observer. In a complicated task, such as evaluating a digital pathology slide, precise timing of the crucial decision (patient diagnosis) is much less clear. Indeed, the interpretive process likely follows a dynamic and cyclical medical decision-making process of inspection, hypothesis development, information accumulation, reasoning, hypothesis checking and modification, and mapping perceived features to potential diagnostic categories.[41–43] The complexity of the interpretive process and the cases examined means that it is very unusual for a clinician to be able to render a diagnosis immediately after fixating a particular region. In laboratory visual search tasks, the target is typically drawn from a discrete and clearly defined target set such that there is generally little doubt once a target has been found. Conversely, it is difficult to define what serves as a "target" in digital pathology slides. In this study, we used cases from a well-described dataset in which each positive case included a consensus ROI that a panel of expert pathologists agreed contained crucial diagnostic information. A simple model of decision making in pathology would predict that the first moment when this ROI was fixated would be correlated with important case-level cognitive processing. Indeed, by co-registering the eye-movement data with the ROI, our initial analyses focused on time-locking to this moment in time.

Unfortunately, we did not find reliable associations between early or late pupil dilations and diagnostic accuracy or case difficulty ratings, even when we parsed cases by type (e.g., invasive carcinoma cases). There are several reasons why this might be the case. The simplest explanation is that, under the aforementioned conditions of examining digital pathology slides, our measure of pupil dilation is simply too noisy to detect the presumed signal associated with seeing important signs of cancer for the first time. Another possibility is that the uncertain time course of the realization that a given region contains important diagnostic information decreased our ability to detect a reliable deflection of pupil amplitude. This is a common problem in signal processing that is difficult to surmount. Temporal variability of this nature may render even large signals difficult to detect.

It is notable that, although the experimental conditions in this study are less controlled than a more traditional laboratory pupillometry search study, we took great care to optimize these settings in ways that would not be possible in a clinical practice. First, our cases were drawn from our database of digital pathology slides. Prior to the study, we convened an expert panel of pathologists who worked together to reach a consensus diagnosis on each case and agree upon an ROI that was most representative of this diagnosis.[44] In a clinical setting, none of this information would exist, and it would therefore be impossible to time-lock analyses to fixating on a specific ROI. In addition, the sample size for this study ($n = 92$) is one of the largest studies that we are aware of in the domain of medical image perception. Furthermore, to ensure high eye-tracking data quality, this study was conducted with an experimenter sitting next to the pathologist. Thus, these data are in many ways an idealized situation that is quite different from the quality of data that installing an eye-tracker into a pathology reading room would likely yield. It is notable that, despite all of these efforts, our data suggest that pupil dilation is not significantly associated with diagnostic accuracy. Despite this somewhat disappointing overall conclusion in terms of the feasibility of using pupillometry in the reading room, we think there are still many promising avenues for future research in this domain.

4.1 *Future Directions*

Training pathology students how to evaluate cases is a challenging task in which there are few universally accepted best practices and huge variability in terms of training protocols across different institutions. In addition, we are in the midst of what may become a sea change as some clinics adopt more digital evaluations of pathology slides, whereas others do not currently have any formal training on digital pathology slides.[45] In this climate, there is value in empirically validated techniques that could help instructors identify trainees that may be struggling with a specific case or class of cases. If training sets are taken from something similar to those employed in this study in which the level of discordance and subject difficulty rating on a given case are known, the early phasic responses in baseline-correct pupil amplitude could serve as evidence to corroborate the need for additional training.

Although this study focused on early responses to the onset of the case, pupil amplitude is a rich, continuous data source that may lend itself to more advanced machine learning (ML) techniques. Using an ML approach would be particularly promising in the aforementioned training situation in which there is an established diagnostic "truth" for a given case and a relatively simple ML algorithm should be able to detect the degree to which a pathologist's evaluation of a given case is likely to result in a correct or incorrect diagnosis. Recently, similar logic was used to detect readers with dyslexia using ML and simple eye-tracking methods.[46] Generalizing eye-tracking and ML to detecting accuracy with pupillometry in an expert population would be a powerful extension of this work.

As a first exploration of the promise of pupillometry in a large sample of pathologists, we focused this work on two moments in the process of diagnostic evaluation that we believed would yield the largest changes in pupil diameter: the first view of the case and the moment when the critical diagnostic region is first fixated. Although we found a reliable effect of expertise and subject-centered difficulty for the first view, none of our measures were strongly associated with first fixation in the ROI. There are a number of reasons why this might be the case, which we have outlined above. Most importantly, to extract a reliable change in pupil diameter, timing is crucial. In retrospect, the moment that pathologists first fixate on an ROI is likely not tightly locked in time to the moment when they reached a diagnosis on the case. Following this logic, one promising avenue for future work might be to time-lock to the moment when the pathologist decides to zoom in or zoom out. Theoretically, these moments should be associated with the moment when the pathologist decides that a given region deserves a closer look when zooming in, or when they have finished closely examining a region when zooming out. It is possible that these decision points could correspond to a shift in strategy from exploration to exploitation (when zooming in) or vice versa (when zooming out).[47,48]

5 Conclusions

Prior work has shown that pupil diameter is sensitive to ongoing task demands and cognitive effort.[19,24,37] This rich, continuous signal therefore presents a promising potential opportunity for identifying medical image cases that may require additional scrutiny in a clinical setting in which decisions have life-altering consequences. This study examined pupil diameter changes in which pathology experts and trainees examined a case set that included the full spectrum from benign to invasive carcinoma. This study took place in what may be considered idealized conditions: the cases were drawn from a uniquely well-studied set such that the diagnostic difficulty and a consensus location for a ROI were known for each case. In addition, the study took place in a large, geographically diverse population of pathologists in the presence of an experimenter who ensured that any difficulties with the eye-tracking system were addressed immediately to ensure data quality and standardization. Nonetheless, the results for the study were mixed. We observed large differences in mean pupil diameter associated with expertise, which may have been driven by the relative difficulty of the task for these two groups. In addition, we found that larger pupil diameter was associated with subject-centered subjective difficulty ratings across cases. However, despite these promising findings, we did not find that pupil diameter in response to the first view of the case was sensitive to diagnostic accuracy or viewing efficiency (how quickly a diagnosis was reached on a case). Finally, although pupil diameter appears to be a reliable metric for overall subject engagement, it is less clear whether it could be used in a clinical setting to improve diagnostic outcomes without additional technical innovations in how pupil data are processed.

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 the iScan Coreo Au™ WSI 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 United States for their participation and assistance in this study. Research reported in this publication was supported by the National Institutes of Health (Grant No. R01 CA225585-05). The content is solely the responsibility of the authors and does not necessarily represent the views of the National Institutes of Health. Researchers are independent from funders, and the funding agency played no role in study design, conduct, analysis, or interpretation. J.G.E., D.L.W., T.T.B., K.F.K., and T.D. 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. T.D., M.L., C.E.K., and T.T.B. drafted the manuscript, and M.L. and C.E.K. performed the statistical analyses. T.D., K.F.K., D.L.W., H.S., and J.G.E. reviewed the manuscript at several times during preparation, providing feedback and suggestions. All authors read and approved the final manuscript. All participants provided written informed consent in accordance with IRB 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. J. G. Elmore et al., "Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study," *BMJ* **357**, j2813 (2017).
2. J. G. Elmore et al., "Diagnostic concordance among pathologists interpreting breast biopsy specimens," *JAMA* **313**(11), 1122–1132 (2015).
3. K. C. Veenhuizen et al., "Quality assessment by expert opinion in melanoma pathology: experience of the pathology panel of the Dutch Melanoma Working Party," *J. Pathol.* **182**(3), 266–272 (1997).
4. J. G. Elmore et al., "A randomized study comparing digital imaging to traditional glass slide microscopy for breast biopsy and cancer diagnosis," *J. Pathol. Inform.* **8**, 12 (2017).
5. C. Mello-Thoms et al., "Perceptual analysis of the reading of dermatopathology virtual slides by pathology residents," *Arch. Pathol. Lab. Med.* **136**(5), 551–562 (2012).
6. T. Jaarsma et al., "Expertise in clinical pathology: combining the visual and cognitive perspective," *Adv. Health Sci. Educ. Theory Pract.* **20**(4), 1089–1106 (2015).
7. 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**(12), 1543–1556 (2006).
8. T. T. Brunye et al., "Eye tracking reveals expertise-related differences in the time-course of medical image inspection and diagnosis," *J. Med. Imaging* **7**(5), 051203 (2020).
9. T. T. Brunye et al., "Eye movements as an index of pathologist visual expertise: a pilot study," *PLoS One* **9**(8), e103447 (2014).
10. T. T. Brunyé et al., "A review of eye tracking for understanding and improving diagnostic interpretation," *Cogn. Res. Princ. Implic.* **4**(1), 7 (2019).
11. 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," *Cogn. Res. Princ. Implic.* **4**(1), 21 (2019).
12. T. Drew et al., "More scanning, but not zooming, is associated with diagnostic accuracy in evaluating digital breast pathology slides," *J. Vis.* **21**(11), 7 (2021).
13. T. Drew et al., "Scanners and drillers: characterizing expert visual search through volumetric images," *J. Vis.* **13**(10), 3 (2013).
14. L. H. Williams et al., "Characteristics of expert search behavior in volumetric medical image interpretation," *J. Med. Imaging* **8**(4), 041208 (2021).
15. M. B. Winn et al., "Best practices and advice for using pupillometry to measure listening effort: an introduction for those who want to get started," *Trends Hear.* **22**, 2331216518800869 (2018).
16. G. Aston-Jones and J. D. Cohen, "An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance," *Annu. Rev. Neurosci.* **28**, 403–550 (2005).
17. B. Laeng, S. Sirois, and G. Gredebäck, "Pupillometry: a window to the preconscious?" *Perspect. Psychol. Sci.* **7**(1), 18–27 (2012).
18. D. Kahneman, *Attention and Effort*, Prentice Hall, Englewood Cliffs, New Jersey, United States (1973).
19. D. Kahneman and J. Beatty, "Pupil diameter and load on memory," *Science* **154**(3756), 1583–1585 (1966).
20. E. H. Hess and J. M. Polt, "Pupil size as related to interest value of visual stimuli," *Science* **132**(3423), 349–350 (1960).
21. E. H. Hess and J. M. Polt, "Pupil size in relation to mental activity during simple problem-solving," *Science* **143**(3611), 1190–1192 (1964).
22. S. Sirois and J. Brisson, "Pupillometry," *Wiley Interdiscip. Rev. Cogn. Sci.* **5**(6), 679–692 (2014).
23. P. van der Wel and H. van Steenbergen, "Pupil dilation as an index of effort in cognitive control tasks: a review," *Psychon. Bull. Rev.* **25**(6), 2005–2015 (2018).
24. J. Beatty, "Task-evoked pupillary responses, processing load, and the structure of processing resources," *Psychol. Bull.* **91**(2), 276–292 (1982).
25. C. M. Privitera et al., "Pupil dilation during visual target detection," *J. Vis.* **10**(10), 3 (2010).

26. J. W. de Gee, T. Knapen, and T. H. Donner, "Decision-related pupil dilation reflects upcoming choice and individual bias," *Proc. Natl. Acad. Sci. U S A* **111**(5), E618–E625 (2014).
27. T. Brunyé et al., "Pupil diameter changes reflect difficulty and diagnostic accuracy during medical image interpretation," *BMC Med. Inform. Decis.* **16**, 77 (2016).
28. N. V. 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). Evaluation Studies Research Support, N.I.H., Extramural," *BMC Womens Health* **13**(1), 3 (2013).
29. N. V. 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)," *BMC Women's Health* **13**, 1–8 (2013).
30. R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2021). https://www.R-project.org/.
31. N. A. Ibraheem et al., "Understanding color models: a review," *ARPN J. Sci. Technol.* **2**(3), 265–275 (2012).
32. S. L. Zeger, K. Y. Liang, and P. S. Albert, "Models for longitudinal data: a generalized estimating equation approach," *Biometrics* **44**(4), 1049–1060 (1988).
33. T. Piquado, D. Isaacowitz, and A. Wingfield, "Pupillometry as a measure of cognitive effort in younger and older adults," *Psychophysiology* **47**(3), 560–569 (2010).
34. T. J. Wright, W. R. Boot, and C. S. Morgan, "Pupillary response predicts multiple object tracking load, error rate, and conscientiousness, but not inattentional blindness," *Acta Psychol.* **144**(1), 6–11 (2013).
35. D. Alnæs et al., "Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus," *J. Vis.* **14**(4), 1 (2014).
36. S. Joshi et al., "Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex," *Neuron* **89**(1), 221–234 (2016).
37. G. Porter, T. Troscianko, and I. D. Gilchrist, "Effort during visual search and counting: insights from pupillometry," *Q. J. Exp. Psychol.* **60**(2), 211–229 (2007).
38. T. R. Hayes and A. A. Petrov, "Mapping and correcting the influence of gaze position on pupil size measurements," *Behav. Res. Methods* **48**(2), 510–527 (2016).
39. V. Peysakhovich et al., "Frequency analysis of a task-evoked pupillary response: Luminance-independent measure of mental effort," *Int. J. Psychophysiol.* **97**(1), 30–37 (2015).
40. B. Pfleging et al., "A model relating pupil diameter to mental workload and lighting conditions," in *Proc. 2016 CHI Conf. Human Factors in Comput. Syst.*, pp. 5776–5788 (2016).
41. M. Li and G. B. Chapman, "Medical decision making," in *The Wiley Encyclopedia of Health Psychology*, K. Sweeny, M. L. Robbins, and L. M. Cohen, Eds., pp. 347–353, John Wiley & Sons Ltd. (2020).
42. G. M. Joseph and V. L. Patel, "Domain knowledge and hypothesis generation in diagnostic reasoning," *Med. Decis. Making* **10**(1), 31–46 (1990).
43. J. F. Arocha, D. Wang, and V. L. Patel, "Identifying reasoning strategies in medical decision making: a methodological guide," *J. Biomed. Inform.* **38**(2), 154–171 (2005).
44. K. H. Allison et al., "Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel," *Histopathology* **65**(2), 240–251 (2014).
45. J. G. Elmore et al., "Pathology trainees' experience and attitudes on use of digital whole slide images," *Acad. Pathol.* **7**, 2374289520951922 (2020).
46. L. Rello and M. Ballesteros, "Detecting readers with dyslexia using machine learning with eye tracking measures," in *Proc. 12th Int. Web for All Conf.*, pp. 1–8 (2015).
47. T. T. Hills et al., "Exploration versus exploitation in space, mind, and society," *Trends Cogn. Sci.* **19**(1), 46–54 (2015).
48. F. Regnath and S. Mathôt, "Pupil size reflects exploration and exploitation in visual search (and it's like object-based attention)," bioRxiv (2021).

Trafton Drew received his doctorate in cognitive psychology from the University of Oregon in 2009. He is an assistant professor at the University of Utah, director of the Applied Visual

Attention Laboratory, and an expert in visual search and attention. He uses a variety of neuroscientific and psychophysical methods to characterize and understand observer performance.

Catherine E. Konold is a research associate in the Applied Visual Attention Laboratory at the University of Utah.

Mark Lavelle is a graduate student at the University of New Mexico and is motivated to understand and evaluate a unified framework for understanding perception and behavior as forms of decision making.

Tad T. Brunyé received his doctorate in experimental cognitive psychology from Tufts University in 2007. He is a visiting associate professor at Tufts University and scientific manager at the Center for Applied Brain and Cognitive Sciences. 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.

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

Hannah Shucard is a research coordinator in the Department of Biostatistics at the University of Washington and is engaged in developing and coordinating research projects across multidisciplinary and multi-center teams.

Donald L. Weaver received his MD degree from the University of Vermont in 1984 and completed his residency and fellowship training in analytical cytometry and anatomic pathology. He is a professor of medicine in the Department of Pathology and Laboratory Medicine at the University of Vermont and Vermont Medical Center, the director of the Surgical Pathology Fellowship Program, and the medical director of the University of Vermont Cancer Center Biobank. He is a leading expert in breast and cervical pathology.

Joann G. Elmore received her MD from Stanford University and her MPH from Yale University, and she completed her residency and fellowships in internal medicine and infectious disease. She 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 University of California, Los Angeles (UCLA) and is the Director of the UCLA National Clinician Scholars Program. She is a national leader in academic general internal medicine and has a distinguished career as an investigator, mentor, administrator, and educator.