More scanning, but not zooming, is associated with diagnostic accuracy in evaluating digital breast pathology slides

Trafton Drew Department of Psychology, University of Utah, Salt Lake City, UT, USA ✉

Mark Lavelle Department of Psychology, University of Utah, Salt Lake City, UT, USA ✉

Kathleen F. Kerr Department of Biostatistics, University of Washington, Seattle, WA, USA ✉

Hannah Shucard Department of Biostatistics, University of Washington, Seattle, WA, USA ✉

Tad T. Brunyé Department of Psychology, Tufts University, Medford, MA, USA ✉

Donald L. Weaver Department of Pathology & Laboratory Medicine, University of Vermont, Burlington, VT, USA ✉

Joann G. Elmore Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA ✉

Diagnoses of medical images can invite strikingly diverse strategies for image navigation and visual search. In computed tomography screening for lung nodules, distinct strategies, termed scanning and drilling, relate to both radiologists' clinical experience and accuracy in lesion detection. Here, we examined associations between search patterns and accuracy for pathologists ($N = 92$) interpreting a diverse set of breast biopsy images. While changes in depth in volumetric images reveal new structures through movement in the z-plane, in digital pathology changes in depth are associated with increased magnification. Thus, "drilling" in radiology may be more appropriately termed "zooming" in pathology. We monitored eye-movements and navigation through digital pathology slides to derive metrics of how quickly the pathologists moved through XY (scanning) and Z (zooming) space. Prior research on eye-movements in depth has categorized clinicians as either "scanners" or "drillers." In contrast, we found that there was no reliable association between a clinician's tendency to scan or zoom while examining digital pathology slides. Thus, in the current work we treated scanning and zooming as continuous predictors rather than categorizing as either a "scanner" or "zoomer." In contrast to prior work in volumetric chest images, we found significant associations between accuracy and scanning rate but not zooming rate. These findings suggest fundamental differences in the relative value of information types and review behaviors across two image formats. Our data suggest that pathologists gather critical information by scanning on a given plane of depth, whereas radiologists drill through depth to interrogate critical features.

Introduction

More than one million women undergo a breast biopsy each year in the United States (Dahabreh et al., 2014). Pathologists are responsible for examining these tissue samples to render a clinical diagnosis, which informs risk assessment and treatment decisions for the patient. There is room for improvement in diagnostic agreement between pathologists, especially for lesions of intermediate severity (Elmore, Longton, Carney, Geller, Onega, Tosteson, Nelson, Pepe, Allison, Schnitt, O'Malley, & Weaver, 2015; Elmore, Nelson, Pepe, Longton, Tosteson, Geller, Onega, Carney,

Citation: Drew, T., Lavelle, M., Kerr, K. F., Shucard, H., Brunyé, T. T., Weaver, D. L., & Elmore, J. G. (2021). More scanning, but not zooming, is associated with diagnostic accuracy in evaluating digital breast pathology slides. *Journal of Vision*, *21*(11):7, 1–17, https://doi.org/10.1167/jov.21.11.7.

https://doi.org/10.1167/jov.21.11.7 Received March 3, 2021; published October 12, 2021 ISSN 1534-7362

Jackson, Allison, & Weaver, 2016). Pathology diagnosis is a complex task requiring expertise in visual search through large, complex series of histologic images and the interpretation of challenging and diverse medical evidence. The recent approval by the Food and Drug Administration (FDA) (Evans, Bauer, Bui, Cornish, Duncan, Glassy, Hipp, McGee, Murphy, Myers, O'Neill, Parwani, Rampy, Salama, & Pantanowitz, 2018) and growing popularity of digital whole-slide imaging provides researchers with opportunities to study the diagnostic viewing process and help improve diagnostic accuracy in ways that would be impossible with glass slide viewing on a microscope. In radiology, digital imaging technology has led to flourishing research on how clinicians search through different types of medical images (for reviews see, Brunyé, Drew, Weaver, & Elmore, 2019; Wu & Wolfe, 2019). Broadly, the goal of this growing area of research is to identify the best methods to search through complex medical images to help clinicians improve efficiency, diagnostic accuracy, or both. Insights about optimal strategy can be particularly helpful in instances where best practices are unknown.

Much of the prior research on clinicians' search behavior through medical images has been done in radiology. Prior research from our group found that radiologists engaged in a lung nodule detection task tended to adopt one of two strategies in searching through these volumetric images. These strategies, termed scanning and drilling, were associated with radiologists' clinical experience and accuracy in lesion detection (Drew, Evans, Võ, Jacobson, & Wolfe, 2013). Drilling involves focusing one's gaze on a small portion of the image while paging through the "stack" of computed tomography (CT) images, whereas scanning involves viewing multiple regions of the image before paging to a new layer in the stack. Drilling was not possible before the onset of digital imaging techniques in radiology. One of the first articles to evaluate the promise of stacking images into a volume found that this novel method for displaying CT images led to improved lung nodule detection (Seltzer, Judy, Adams, Jacobson, & Stark, 1995). The authors proposed that this was because mostly spherical nodules tended to "pop" in and out of view when rapidly scrolling (or drilling) through depth, facilitating their detection. Consistent with this idea, Wen, Aizenman, Drew, Wolfe, Haygood, and Markey (2016) found that radiologists who used a drilling strategy when searching through chest CT scans were more likely to focus on dynamic saliency, which takes depth into account, rather than two-dimensional (2D) saliency, which does not.

Although there is now evidence that drilling is associated with higher accuracy than scanning when engaged in a lung cancer screening task with chest CT images (Drew, Võ, Olwal, Jacobson, Seltzer, & Wolfe, 2013; Williams, Carrigan, Mills, Auffermann, Rich, & Drew, 2021), it is unclear whether this result generalizes to different image screening tasks or different imaging modalities. Ba, Shams, Schmidt, Eckstein, Verdun, and Bochud (2020) found that the drilling strategy was associated with higher rates of detection for a subtle simulated liver lesion while evaluating abdomen CT scans. In contrast, Aizenman, Drew, Ehinger, Georgian-Smith, and Wolfe (2017) found no relationship between search strategy and diagnostic accuracy when they investigated diagnostic accuracy in breast tomosynthesis image evaluation. They also found that breast radiologists tended to move through depth much more quickly than the radiologists evaluating chest CT in the study by Drew et al. (2013). This observation may have been driven by the fact that 10 of 11 participants in this study reported being taught to adopt a drilling strategy for tomosynthesis. Although both the Aizenman and Ba articles adopted the Driller/Scanner dichotomy suggested by Drew and colleagues (2013), both noted that search strategy may be more accurately described as a bipolar continuum anchored at driller and scanner, where not every radiologist neatly falls into one of two dichotomous categories.

Eye-tracking data provide rich information about what regions of medical images receive focal attention, but greatly complicate data collection, especially with expert populations of clinicians who are very busy. Although modern eye-trackers are much less cumbersome than their predecessors, which frequently required rigid posture via chin-rests or a bite bar to attain reliable tracking data, even modern eye-trackers require careful attention from the experimenter to ensure that the clinician does not move too close or too far from the tracker (See Figure 1). Eye tracking also requires in-person data collection, which can be challenging to schedule during normal times and restrictive in the context of the current pandemic. Thus data collection would be greatly simplified if we could remotely extract equivalent search strategy data from clinicians without tracking eye movements.

Two recent studies have examined whether meaningful information about scanning and drilling strategies can be extracted without eye-tracking data (Mercan, Shapiro, Brunyé, Weaver, & Elmore, 2018; van Montfort, Kok, Vincken, van der Schaaf, van der Gijp, Ravesloot, & Rutgers, 2020). A recent longitudinal examination of Dutch medical residents tracked depth changes as a variety of different volumetric images were evaluated as part of a bi-annual mandatory progress examination (van Montfort et al., 2020). The results were mixed. Although they found that the tendency to move through depth quickly (suggestive of a drilling strategy) increased with experience, this strategy was not significantly associated with higher diagnostic accuracy. One possible explanation is that movement through depth was not associated with diagnostic

Figure 1. A participating pathologist reviewing a digital whole slide image with eye tracking; face obscured for privacy.

performance because the depth measure does not include any information about XY eye-movements. It is therefore possible that clinicians who moved more quickly through depth were also moving their eyes quickly in XY space, thereby reducing the presumed reason for the previously observed benefits associated with drilling (Seltzer et al., 1995; Wen et al., 2016). The drilling metric may also not have been associated with diagnostic accuracy due to the heterogenous nature of the cases included in this sample. In contrast to the work previously outlined where researchers focused on a single clinical task (e.g., lung nodule detection, mammography screening for breast cancer), radiology residents in the van Montfort study (2020) examined a variety of CT scans including six different radiology subdomains and structures (e.g., head CT, pelvic CT, chest CT). Although their driller/scanner analyses excluded cases with diffuse areas of interest in depth (e.g., pneumonia), this wide range of case type and task sets it apart from the previously discussed research in this area.

Digital pathology provides a unique intermediary position between logging only depth changes and combining eye-tracking with depth change information. The vast images in digital pathology allow the user to pan through an XY plane within a given magnification level. To characterize search through digital whole slide images, Mercan and colleagues (2018) logged pathologists' movements as they examined breast pathology biopsies. Although biopsy slides are 2D, pathologists utilize zoom (magnification) to reveal features in higher detail for further evaluation (e.g., from 1x to 40x). The researchers used this third dimension of image search (combined with panning across the XY plane) to study pathologist behavior in terms of scanning and drilling. They defined drilling behavior as zooming in and out frequently and rarely panning at high magnification. Scanning behavior was referred to as panning at intermediate magnification and seldom zooming in or out. Mercan and colleagues created a scanning versus drilling (SvD) metric that quantified scanning and drilling on a bipolar continuum. They did not find evidence that the SvD metric was associated with diagnostic accuracy. Notably, unlike Drew and colleagues (2013) and similar to von Montford (2020), Mercan and colleagues (2018) did not use eye tracking, meaning that their scanning metric was restricted to collecting movement of the viewing port in XY space.

Similarly, Mello-Thoms, Mello, Medvedeva, Castine, Legowski, Gardner, Tseytlin, and Crowley (2012) examined movement of the viewing window as pathologists examined dermatopathology slides. This research was conducted before the popularization of the driller/scanner terminology. However, they found that pathologists who moved more through XY space than Z space (a strategy they referred to as a "fishing expedition") were less accurate than colleagues who tended to move more frequently through depth. However, despite this prior evidence from both chest CT and dermatopathology that seems to suggest that drilling/zooming may be a superior strategy in breast pathology, it is important to note that the images in the breast pathology are quite distinct from chest CT and dermatopathology. As we outline below, movement through depth in the volumetric CT images is very different than the magnification in digital pathology images. Moreover, dermatopathology slides typically contain multiple slices of the same region of skin

that are arranged on the pathology slide, often with redundant information across segments of tissue. This allows the pathologist to assess different depths of the structure in one plane of XY space. In contrast breast pathology slides are typically a single slice of breast tissue. These differences in the nature of the pathology slides likely influence the optimal method of searching through them.

In the present study, we tracked pathologists' eyes as they interpreted digital breast biopsy slides. Combining eye-tracking with logged pan and zoom data, we sought to more precisely quantify scanning and drilling behaviors and examine associations with diagnostic accuracy. Based on prior data suggesting that drilling is associated with improved lung nodule detection on CT imaging and our informal discussions with expert pathologists, we hypothesized a positive association between tendency to use a drilling/zooming strategy and diagnostic accuracy in a challenging set of diverse digital breast pathology cases.

"Drilling" or "zooming" in pathology?

The terminology of "scanning" and "drilling" has been borrowed from work that began with radiologists examining volumetric images (Drew, Võ, et al., 2013). Movement through unique anatomical slices also allows the viewer to assess changes in a structure's size through depth (Seltzer et al., 1995). This is not possible in digital pathology because scrolling through depth in this domain reveals new levels of detail through magnification rather than maintaining magnification levels while moving to novel slices, or levels of the structure. Movement through depth is an important part of digital pathology because many abnormalities cannot be resolved and accurately diagnosed at low magnification levels. Whereas movement in depth in volumetric images such as CT allows the clinician to perform volume exploration through anatomical structures, the primary value of movement in depth for pathology appears to be an increased ability to discover more details about the fine structure of the case that are not visible at low magnification. Radiology images explore three-dimensional space (e.g., CT scans) or are a summation of spatial features on a 2D plane (e.g., chest films). Pathology whole slide images are created from a high-resolution base image viewed, using zoom, at various levels of detail. This is similar to viewing Google Earth maps where at low power the viewer can identify global characteristics (e.g., a desert) but at the highest power the viewer can see more granular features (e.g., houses and vehicles). Pathology uses the zoom space to improve discovery by unmasking detail. With CT images, zoom space is used to discover new territory, like swimming in a coral reef where changes in depth reveal new features that were previously not visible, whereas 2D chest radiograph's zoom space magnifies without adding new detail, unlike pathology images.

Based on these differences in what movement through depth means for these different domains, in our opinion, magnification, or "Zooming" is a more accurate descriptor for digital pathology. Though it might be tempting to treat movement through depth as similar across different imaging modalities, in comparing across modalities it is crucial to consider the purpose of the movement through depth. From this perspective, "zooming" seems to be a reasonable description of movement through different levels of magnification in digital pathology and large aerial surveillance images (e.g., Božić-Štulić, Marušić, & Gotovac, 2019). On the other hand, "drilling" seems more appropriate for modalities where depth changes are accompanied by previously invisible layers as in chest CT and breast tomosynthesis.

Finally, although the current examination focused upon pathologists zooming into higher magnification to make diagnoses, radiologists also use magnification to resolve subtle findings in 2D images such as chest radiographs. We are not aware of any research on the usage of magnification as it relates to diagnostic accuracy in radiology but it would be interesting to evaluate how clinicians use zooming in these distinct imaging modalities. As future researchers consider optimal methods for searching through complex images that involve depth, it will be important to recognize that different image formats can lead to changes in depth that reveal different types of information (e.g., higher levels of detail or new levels to view).

Materials & methods

Participants

We collected data from 92 pathologists (*Attendings*, n=20; *Residents*, n=72) recruited from nine academic medical centers across the United States (states represented included CA, KY, MA, UT, VT, VA, and WA) as part of a larger longitudinal study (Elmore et al., 2020). The *Residents* group included 69 pathology residents, two pathology fellows, and one post sophomore pathology student fellow. A contact person at each site introduced the study to their attending physicians and trainees and provided contact information for potential participants, but 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, and be available

during the one- or two-day site visits arranged for data collection. Attending pathologists had to be available during the site visit dates and were willing to interpret breast biopsy cases. Approval was obtained from the appropriate Institutional Review Boards, with the University of California at Los Angeles acting as the Institutional Review Board of record. All participants provided informed consent and received a $50 gift card for their involvement. The study was conducted in accordance with the tenets of Declaration of Helsinki.

Stimulus creation/case selection

Whole slide digital images from each case in 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 elsewhere (Elmore et al., 2015). Each of the 240 B-Path cases has standardized clinical data, a high-quality whole slide digital image and a comprehensive consensus-defined reference standard diagnosis. Each glass slide used for the digital whole slide image was cut from a source paraffin embedded tissue block then routinely stained with hematoxylin and eosin (H&E) in a single histology laboratory to maintain consistency within the resulting digital images. The cases were divided into test sets and interpreted by >200 practicing U.S. pathologists; each case in the test sets has data from a minimum of 27 pathologists.

Each case in the current study were selected from the 240 B-Path cases based on histology form data 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 can be classified into five diagnostic categories of increasing severity: benign, atypia, low- and high-grade ductal carcinoma in situ (DCIS), and invasive. These categories are associated with different histopathological features, treatment and surveillance options, and prognosis. We identified 32 cases to include in the current study: four benign, 10 Atypia, 10 low-grade ductal carcinoma in situ (LGDCIS), four high-grade ductal carcinoma in situ (HGDCIS), four invasive breast carcinoma. One additional invasive carcinoma case with high diagnostic concordance (93%) when interpreted by prior pathologists in the gold-standard glass slide format was selected for use as a practice case before each participant evaluated the test sets (Oster, Carney, Allison, Weaver, Reisch, Longton, Onega, Pepe, Geller, Nelson, Ross, Tosteson, & Elmore, 2013). We divided these cases into three sets of 14 cases each (five 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 (see Table 1). Cases were divided into three sets because this study was part of a larger longitudinal study where residents are asked to evaluate one of the three sets per year during three residency training years.

A reference diagnosis for each case was carefully defined by a panel of three expert pathologists who are internationally recognized for research and continuing medical education on diagnostic breast pathology. The three expert pathologists independently reviewed all 240 breast biopsy cases, blinded to previous interpretations of each specimen and to each other's interpretation (Allison, Reisch, Carney, Weaver, Schnitt, O'Malley, Geller, & Elmore, 2014; Feng, Weaver, Carney, Reisch, Geller, Goodwin, Rendi, Onega, Allison, Tosteson, Nelson, Longton, Pepe, & Elmore, 2014; Oster et al., 2013). Cases without unanimous independent agreement were resolved with consensus discussion during in-person meetings using a modified Delphi approach. For this study we operationalized diagnostic accuracy relative to the expert consensus reference diagnosis. Each breast biopsy case was classified into one of five diagnostic classes reflecting increasing disease severity: benign, atypia, low-grade DCIS, intermediate to high-grade DCIS, and invasive cancer. Correct diagnoses are those where the diagnostic category determined by participants agree with the reference diagnosis.

Images were shown on a digital slide viewer developed for prior work from our group (Elmore, Longton, Pepe, Carney, Nelson, Allison, Geller, Onega, Tosteson, Mercan, Shapiro, Brunyé, Morgan, & Weaver,

	Test set A	Test set B	Test set C	Included in all test sets
Benign	1	1	1	1
Atypia	3	3	3	1
Low grade DCIS	3	3	3	1
High grade DCIS	1	1	1	1
Invasive	1	1	1	1
Total	9	9	9	5

Table 1. Number of cases in each class that were unique to each test set, and the number of cases in each class that were common to all three test sets.

2017), using Microsoft Silverlight and Deep Zoom tools (Microsoft, Inc., Redmond, WA, USA). The viewer displayed images in a navigable viewport that allows zooming (range 1x to 60x) and panning while maintaining full image resolution. Participant behaviors (e.g., current view position and zoom level) were logged by the viewport at 5 Hz, and eye-movements were subsequently overlaid by co-registering the viewport and eye-tracking data.

Eye-tracking

We used the mobile remote eye-tracking device (RED-m) system. This is a noninvasive and portable eye tracking system manufactured by SensoMotoric Instruments (SMI, Boston, MA, USA). The system uses an array of infrared lights and cameras to track eye position at 250 Hz with high gaze position accuracy (0.4°) using a nine-point calibration process. For data collection, we mounted the RED system to the bottom of a color-calibrated 22″ Dell liquid crystal display (LCD, 1920 × 1080 resolution) computer monitor. Participants were seated approximately 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. We used SMI's default settings to categorize saccades (peak velocity exceeding 40°/sec occurring between the first 20% and last 20% of velocities exceeding the fixation velocity threshold, lasting at least 20 ms) and fixations (not a blink or a saccade, lasting at least 50 ms).

Procedures

Before reviewing the cases, participants completed an online consent form and online baseline survey of demographics, clinical experience, and attitudes toward breast pathology and digital whole slide imaging. 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 a practice case to ensure they were comfortable with the procedure. Each participant then viewed one of the three subsets of 14 cases, at full screen. The 14 cases within each slide set were shown in a different random order for all study participants viewing that particular set of cases. 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, whether they would want a second opinion from another pathologist, and then 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 us to ensure that the viewing environment (e.g., room lighting, participant positioning) was as similar across the nine collection sites as possible. All participants used identical computer, monitor, keyboard, mouse and eye-tracker models.

Results

Data preparation and analysis plan

Data from one attending and three trainees were not included for analysis because of poor eye-tracking data quality or failure of the system to save data. Data from the remaining 88 participants were available for analysis (19 attendings, 69 residents). Diagnoses provided by participants were considered accurate if they matched the expert consensus diagnosis and inaccurate otherwise (for more detail, see "Stimulus Creation/Case Selection" above). Consistent with prior work with our group (Elmore et al., 2015), diagnostic accuracy varied widely across our diverse set of cases (mean accuracy across pathologists: 43.5%, standard deviation [*SD*]: 26.6%). Trials with poor eye-tracking quality (11% of all trials) were excluded, as were four cases with ceiling (>98% correct) and floor performance (<2% correct), leaving 1036 trials for analysis. Eye tracking quality was considered inadequate if more than 50% of samples in the raw output were unusable. Samples were unusable if pupil measurements equaled zero or were blank or the estimated gaze position was at or below the coordinate (0,0) in monitor space. Five additional outlying trials were identified in our independent variables (described below) using extremely conservative univariate criteria based on the median absolute deviation (MAD) (Leys, Ley, Klein, Bernard, & Licata, 2013). Trials for which the measurement on any independent variable deviated from the median by more than 5.05*MAD were considered outliers and excluded from analysis.

Based on prior work in this area (Drew, Võ, et al., 2013; Mercan et al., 2018), our initial assumption was to treat zooming and scanning as search strategies in opposition of one another. The SvZ metric quantifies which behavior a pathologist favors: scanning or zooming. This metric ranged from −100% for a trial without gaze movement (no scanning) to 100% for a trial without zoom change (no zooming, see Figure 2). For each trial, zooming was quantified as the number of times participants doubled or halved the magnification level. After co-registering monitor-based gaze data with image navigation, scanning was measured as the cumulative distance in pixels of virtual "eye-movements" divided by 1920—the pixel-width of

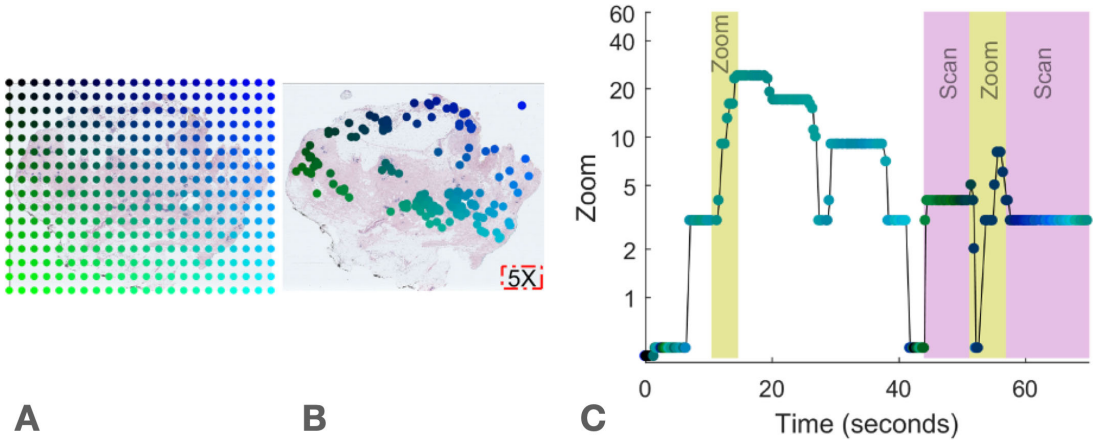

Figure 2. Spatial and temporal representations of fixations from a pathologist interpreting a digital whole slide image. (A) Color conventions for this representation. We created a colormap for each case to enable visualization of where the pathologists looked in the case as a function of time as in panel C. (B) Digital whole slide image at minimum magnification. For reference, the area of viewable tissue at magnification ×5 is represented by the red box in the lower right corner. Green-blue dots represent eye fixations co-registered to the biopsy coordinates. (C) Timeseries of the same fixations in B. Time of each fixation is on the horizontal axis, and zoom (magnification) level of each fixation is on the vertical axis. Scanning is measured cumulatively as distance between fixations, adjusted for the width of viewable tissue at the time of each fixation (determined by zoom level at each fixation). Zooming is measured cumulatively as the difference in zoom level of each fixation. Each doubling or halving of zoom level increments the zooming metric by 1. Relatively high periods of zooming and scanning are indicated by the yellow and pink bars, respectively.

the monitor used for case review. Using this definition, a pathologist panning around the slide without moving their eyes would nonetheless view different tissue regions, thereby increasing the scanning measurement. Zooming tracked additive changes in zoom level after $\log_2$ transforming zoom. In the original zoom units, the zooming metric would have equally weighed changes in zoom level that have little perceptual or diagnostic significance (e.g., from ×55 to ×60—a 9% increase in magnification) with changes that may substantially impact perception and diagnosis (e.g., from ×1 to ×5—a 500% increase in magnification). Doubling (or halving) the magnification level incremented the zooming metric by one, regardless of the starting value, e.g., ×2 to ×4 (or ×10 to ×5).

Our data did not support treating scanning and zooming as strategies in opposition. In contrast to previous research with radiologists (e.g., Drew et al., 2013), in pathology movement through the z-plane of the image does not involve moving to new, previously invisible levels of depth. Thus the unit of measurement for the zooming metric in the present study differ from the units for the scanning metric. Given the lack of a common scale for scanning and zooming, the zero point of the SvZ metric (numerically equal scanning and zooming for the same case) is arbitrary and thus binary or continuous classification of pathologists as scanners or zooming is arbitrary. Critically, careful analysis of the data suggests that tendencies for scanning and zooming may be orthogonal, not oppositional. Although there are instances of pathologists who were high on our zooming metric and low on our scanning metric and vice versa, there were also many instances of pathologists who could not neatly be categorized as a scanner or zooming (see Figures 3 and 4). A Pearson correlation revealed that an individual's average rate of scanning across cases was not significantly correlated with average rate of zooming across cases in our data ($r_=$ 0.13, 95% confidence interval [CI] = [−0.08, 0.33], $p = 0.22$, see Figure 5). This echoes the findings of Aizenman and colleagues (2017) in a study of radiologists interpreting breast tomosynthesis cases where they found no evidence that their scanning metric was negatively correlated with their drilling metric.

To assess associations of interpretive strategies with diagnostic accuracy, we used conditional logistic regression with diagnostic accuracy as the outcome and the cases serving as the strata. Participants were treated as clusters for computing standard errors to account for non-independent data. For reasons given above and additional reasons described below, we focus our analyses on distinct zooming and scanning metrics rather than attempting to combine the two into a single continuum. Thus the predictors of interest were scanning and zooming rates measured separately, rather than a combined SvZ metric. We further investigated associations with accuracy between scanning rate and zooming rate while controlling for each other, as well as interpretive duration. We defined interpretive duration at the time from when the case image was first visible to the moment when the pathologist clicked the "done" button indicating they were ready to move

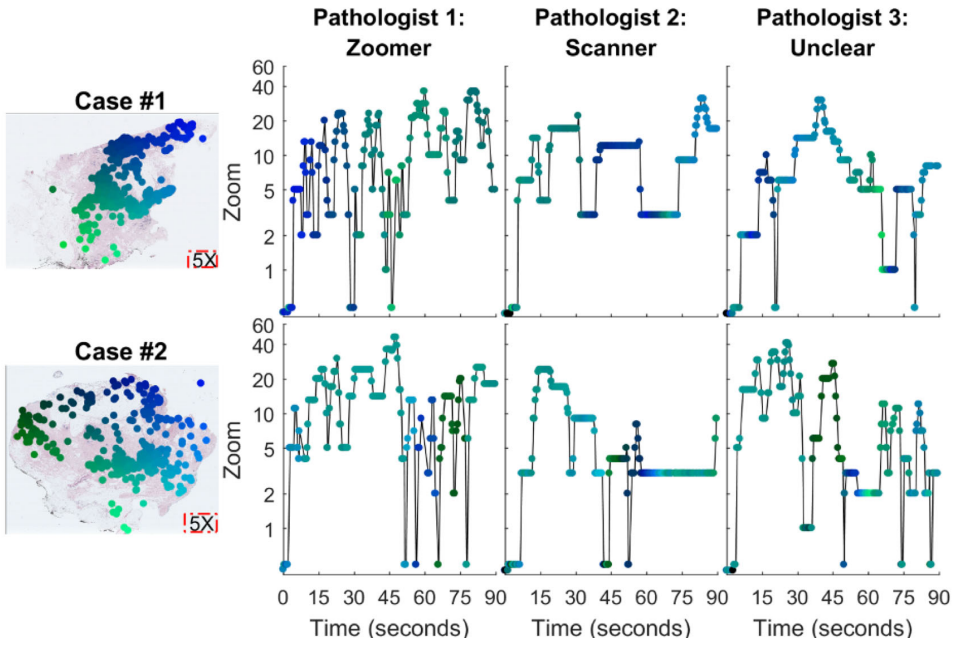

Figure 3. Fixation data for three pathologists viewing the same two cases. Refer to Figure 2 for conventions. Overlaid on the digital whole slide images (left) are fixation data from all three pathologists. On the right, Pathologist 1 showed high drilling (zooming in and out) across both case 1 and case 2. Pathologist 2 displayed less drilling behavior and scanned at a higher rate across the two cases. The third pathologist shows a combination of high drilling and scanning making the categorization uncertain. Note the apparent similarity in search technique across these two cases (see Figure 5 for reliability analyses).

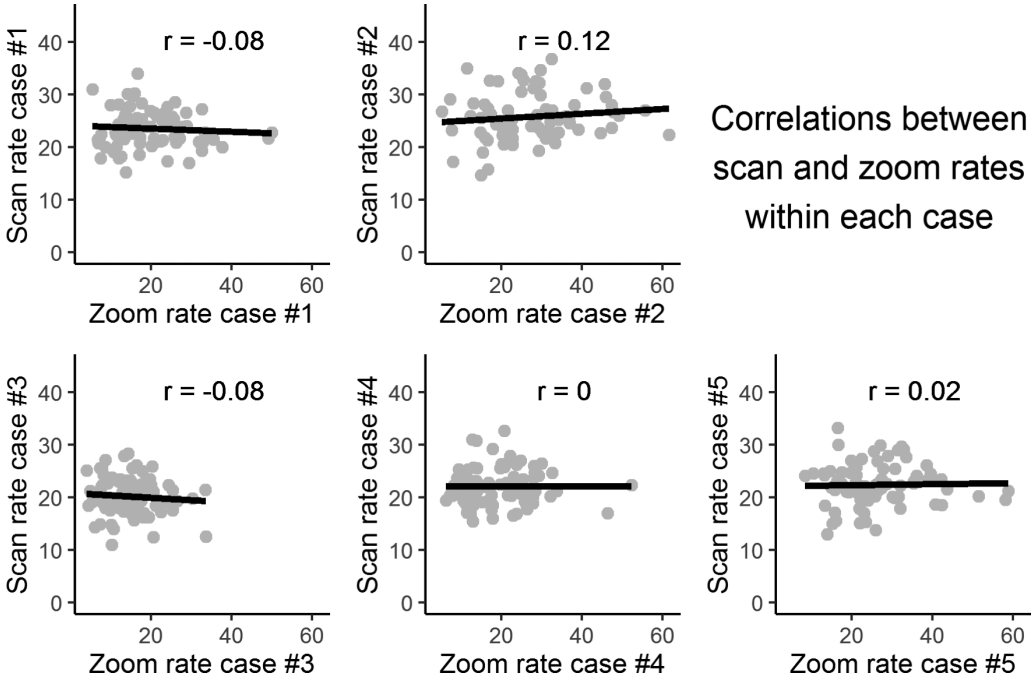

Figure 4. Scan rate plotted against zoom rate across pathologists interpreting the same case, for each of five distinct breast biopsy cases. Each dot represents a single pathologist. These five cases were viewed by every pathologist in our study. Due to data quality, sample sizes for correlations range from $n = 77$ to $n = 85$. None of the correlations was statistically significant: p values for all correlations are >0.3.

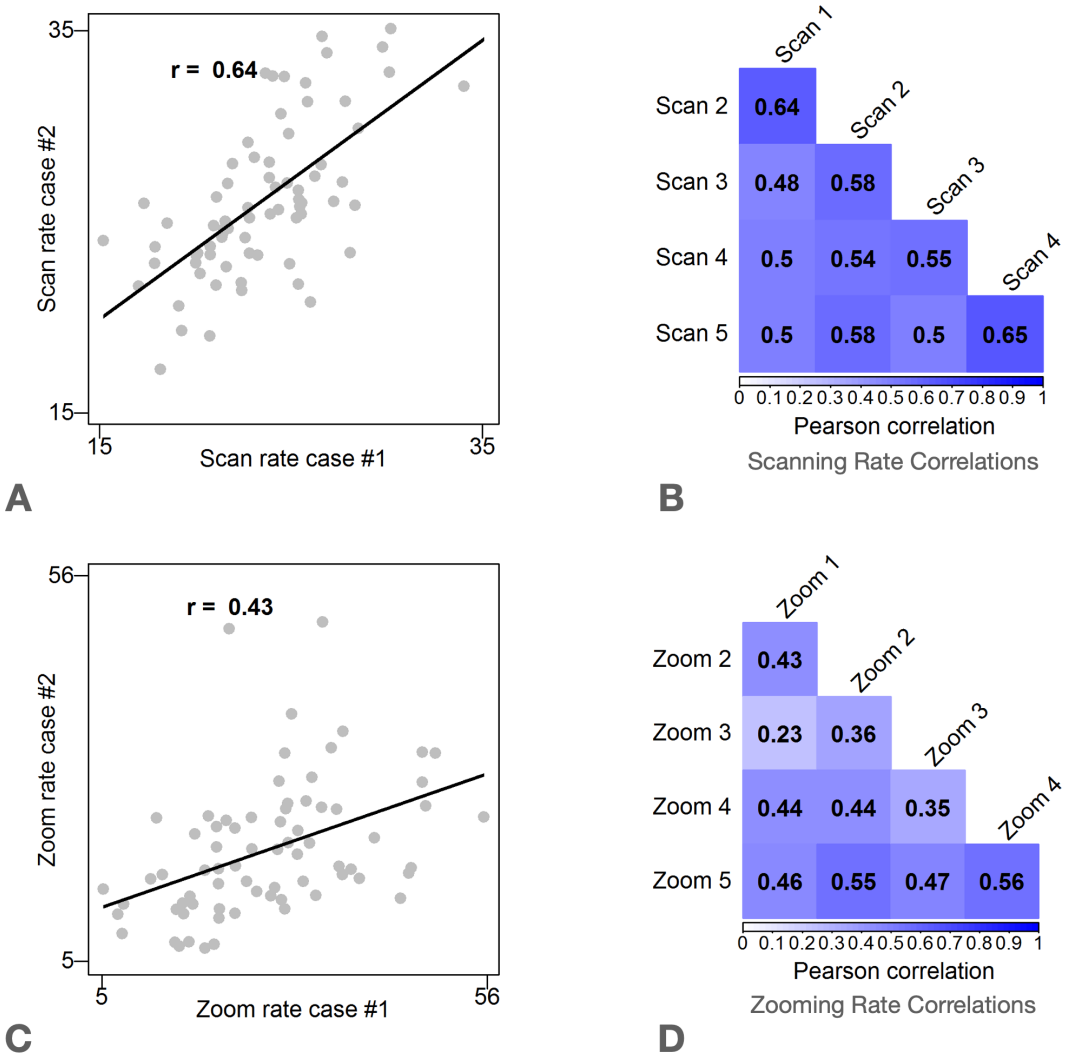

Figure 5. Scan and Zoom rate correlations. Results show that participants who had a high scanning rate on one case tended to have a high scanning rate on other cases. Due to exclusions from poor data quality, sample sizes for correlations range from $n = 69$ to $n = 79$. (A) Scatterplot of scan rate measured between Case 1 and Case 2. Each dot represents one pathologist who interpreted Cases 1 and 2. (B) Scan rate correlations for the five cases viewed by all participants. All p values <0.044. (C) Scatterplot of zoom rate measured between Case 1 and Case 2. Each dot represents one pathologist. D: Zoom rate correlations for the five cases viewed by all study participants. All p values <0.001.

on to the follow questions on the case. Notably, this includes the time it took for the pathologists to draw a single region of interest on the region they deemed most representative of their highest (most severe) diagnosis. Interpretive duration varied widely (mean = 110 seconds, $SD = 54$ seconds) across our diverse sample (first residents to attendings physician) and case selection (benign to invasive carcinoma cases).

We also included saccadic amplitude and average zoom as independent variables in conditional logistic regression to replicate previous analyses from medical image perception and evaluate commonly reported notions of the relationship between use of magnification and pathologists' skills. Saccadic amplitude, defined as the distance (in degrees of visual angle) between two successive saccades, has been shown to be associated with both expertise and experience in medical image perception (Brunyé et al., 2019; Gegenfurtner, Lehtinen, & Säljö, 2011; Williams & Drew, 2019). Longer saccadic amplitude is generally associated with more experience and better performance, though see Williams et al. (2021), for an exception to this general rule in radiologists examining chest CT images.

Scanning and zooming metrics

Figure 2 provides an illustration of how scanning and zooming metrics were calculated. Scanning and zooming were measured per second, i.e., scanning rate

and zooming rate. This served two purposes. First, this isolated these viewing behaviors from interpretive duration, which was highly variable across cases and participants (Mean, *SD*). In prior medical image perception research, shorter interpretive durations have been associated with greater professional experience and higher diagnostic accuracy (e.g., Brams Ziv, Levin, Spitz, Wagemans, Williams, & Helsen, 2019; Gegenfurtner et al., 2011). Therefore, in evaluating the effectiveness of scanning or zooming strategies within breast pathology, we did not want the metrics to be conflated with time. That is, we were most interested in how pathologists were using their time. Second, this distinguished scanning and zooming from each other, removing the covariation due to the mutual dependence on interpretive duration. This allowed us to evaluate whether both, one, or neither strategy is associated with diagnostic accuracy in pathology.

To evaluate whether pathologists' use consistent degrees of scanning and zooming across cases, we estimated the reliability of scanning and drilling measured for each observer and case. Here, we focused on the 5 cases that all participants viewed in order to avoid spurious findings due to examining correlations with a small number of observations. We found that Cronbach's alpha was high for both drilling rate (α= 0.73, 95% CI = [0.63, 0.82]) and scanning rate (α= 0.85, 95% CI = [0.79, 0.91]). The analysis was based on 62 pathologists with available data for the five cases that all participants viewed. In accordance with these analyses, the correlation between each individual's scanning rates for any pair of cases is reliably moderate to strong. The relationship is weak to moderate for any pair of cases for zooming rates, as shown in Figure 5. The Cronbach's α estimates reiterate what can be observed in Figures 3 and 5: zooming and scanning search strategies appear to be consistent within an individual. In contrast, the two strategies are not strongly related to one another (Figure 4).

To summarize, although some of our prior work has attempted to dissociate these two strategies and categorize each clinician as either a "zooming" or a "scanner," this framework appears to be inappropriate for pathologists viewing breast tissue images. Therefore, within pathology it appears more appropriate to measure scanning and zooming separately rather than on a single bipolar continuum. It remains to be seen whether this observation of the apparent independence of the zooming and scanning strategies is unique to this particular task (breast tissue pathology evaluation) or generalizes more broadly to other tissue types or stains other than H&E.

Diagnostic accuracy

Besides the novel scanning and zooming rate measurements, we also investigated associations between diagnostic accuracy and other viewing behavior metrics examined in previous medical image perception studies. For this purpose, we measured average saccadic amplitude and interpretive duration for each trial. Because the distributions of individual saccades on each trial were highly positively skewed, we $\log_2$ transformed them before averaging to obtain a representative measure of central tendency. We also measured the weighted average zoom level per trial that considered the proportional duration spent at each zoom level. As with the zooming rate, zoom levels were $\log_2$ transformed prior to weighted averaging so that perceptually irrelevant changes in zoom (e.g., $\times$55 to $\times$60) were compressed relative to impactful changes ($\times$1 to $\times$5).

We report odds ratios (ORs) comparing the odds of an accurate diagnosis for pathologists interpreting the same case with a one standard deviation difference in the Independent Variable (see Table 2). This was achieved by scaling each variable by its respective standard deviation. First, we fit a logistic regression model with accuracy as the response variable and scanning rate, zooming rate, and interpretive duration as the predictors. In this model (Model 1), scanning rate was significantly and positively associated with accuracy (OR = 1.12, p = 0.01) but drilling rate (OR = 0.95, p = 0.23) and interpretive duration (OR = 0.93, p = 0.16) were not. In a second model (Model 2), we fit a logistic regression model with accuracy as the response variable and saccadic amplitude, average zoom level and interpretive duration as the predictors. Saccadic amplitude was significantly associated with diagnostic accuracy (OR = 1.10, p = 0.04) but average zoom (OR = 1.02, p = 0.70) and interpretive duration (OR = 0.93, p = 0.20) were not.

Model	Independent variable	B (SE)	Odds ratio	*p* value
1	Scan Rate	0.11(0.046)	1.12	**0.01**
	Drill Rate	−0.05(0.044)	0.95	0.23
	Interpretive Duration	−0.07(0.051)	0.93	0.16
2	Saccadic Amplitude[a]	0.09(0.044)	1.10	**0.04**
	Average Zoom[a]	0.02(0.053)	1.02	0.70
	Interpretive Duration	−0.07(0.057)	0.93	0.20
3	Average Zoom[a]	−0.03(0.050)	0.97	0.52

Table 2. Results of conditional logistic regression of single-trial accuracy onto various models, each including independent variable(s) related to eye-movement or image navigation (n = 1031). *Note:* [a]$\log_2$ transformed before averaging.

Discussion

Recent work from our group using data from this study examined early image viewing behaviors, defined

as behaviors observed before the first magnification increase (Brunyé, Drew, Kerr, Shucard, Weaver, & Elmore, 2020). Although we found that more experienced pathologists tended to spend more time in this early period fixating critical regions of interest associated with case diagnosis, we did not find evidence that any viewing behavior during the early period was associated with diagnostic accuracy. This result appears to be contrary to the holistic processing theory of medical image perception (Kundel, Nodine, Conant, & Weinstein, 2007; Kundel, Nodine, Thickman, & Toto, 1987), which suggests that experience allows clinical experts to quickly extract diagnostic information from medical images (e.g., Drew, Evans, et al., 2013; Reingold & Sheridan, 2011). It is important to note that Kundel's holistic model was based on research with focal lesions in 2D images (Kundel et al., 2007; Kundel et al., 1987). Although the breast pathology findings in the current study are focal, rather than diffuse, diagnosis of these findings is extremely uncommon at low magnification. It may therefore not be surprising that this model does not account for our findings. Although the previous work from our group (Brunyé et al., 2020) focused exclusively on initial viewing before any zooming, the current investigation considered the entire visual examination period and found that specific search behaviors were indeed associated with diagnostic accuracy.

Research in radiology has suggested that examining volumetric medical images can be approached in one of two broad strategies: drilling or scanning (Aizenman et al., 2017; Ba et al., 2020; Drew, Võ, et al., 2013). According to this model, clinicians predominantly adopt either a drilling strategy with quick movement in depth and less movement in XY space, *or* scanning with slower movement in depth and more movement in XY space. This prior work has associated the drilling strategy with higher diagnostic accuracy for performing a lung nodule detection task in chest CT scans and when searching for liver lesions in abdominal CTs, but not when evaluating breast tomosynthesis images for signs of breast cancer. This pattern of findings may be due to the fact that although lung nodule detection in chest CT is aided by motion detection through depth (Seltzer et al., 1995; Wen et al., 2016), breast cancer evaluation in volumetric images necessitates both motion detection for masses and focal attention for microcalcifications (Gandomkar & Mello-Thoms, 2019). The purpose of the current investigation was to examine the following:

1. Do pathologists evaluating digital breast pathology slides use similar strategies to those observed in radiologists examining volumetric images?
2. Is search strategy associated with diagnostic accuracy in digital breast pathology?

We address each question below in turn.

Search strategies in digital pathology

Evaluating digital breast pathology slides is fundamentally different from evaluating volumetric medical images. Most importantly, prior investigations with radiologist observers are based on volumetric images, where movement in depth (drilling) reveals new information that was not previously visible. In contrast, movement in depth from low to higher magnification (zooming) within digital pathology results in changes in resolution and narrowing or widening of the field of view on the image. Therefore scrolling through zoom levels in pathology may reveal new histologic features, but it does not reveal previously unseen layers of structure as in radiology volumetric images. However, given the massive size of the images in the digital pathology, it is necessary to magnify to resolve the image in sufficient detail to make crucial clinical judgements. The scale of image size in digital pathology often necessitates panning the viewing window while holding depth constant; this is related to maintaining global orientation on the image. This movement in XY space while holding depth constant is not common technique for examining most radiological images and, to our knowledge, has not yet been systematically examined. Interestingly, panning the viewing window through XY space while holding zoom level constant closely matches the definition of the scanning originally offered by Drew and colleagues (2013). It is also notable that a prior examination of pathologist viewing behavior without eye-tracking focused on these panning movements in defining the scanning strategy, and documented an extreme case of this strategy (informally referred to as "lawn-mowing" by our pathologist colleagues) in Figure 1a of prior work from our group (see Mercan, et al., 2018).

Broadly speaking, whereas radiologists scroll through depth to ***explore*** through the volume, pathologists tend to zoom in order to ***interrogate*** suspicious regions in more detail. The conceptualization of radiologist strategy being *either* scanning *or* drilling through volumetric images grew out of discussions with radiologists who tended to describe one of these two broad strategies. Although we initially used this oppositional framework to guide our conceptualization of zooming and scanning for the current work in pathology, we discovered that our metrics for these two strategies were not negatively associated. For example, pathologists who were high on our zooming scale had no tendency to be low on our scanning scale. Based on this observation, for experts examining this type of image we posit that it is not useful to combine scanning and zooming metrics into a composite metric. Accordingly, we focused subsequent analyses on assessing these metrics separately.

In future work, it will be interesting to examine whether this approach generalizes beyond H&E stained slides in breast pathology, both narrowly within digital

pathology evaluation and more broadly across fields that evaluate images that involve a depth dimension (e.g., CT scans in radiology, aerial surveillance). The notion that drilling/zooming may be a superior strategy grew out of discussions and observations focused on a single task (lung cancer screening) with a relatively small sample (Drew, Võ, et al., 2013). Drilling may be particularly suited to this specific task and imaging modality. Subsequent follow-up work, which has broadened from this starting point to different tasks while searching through different images, has yielded mixed results with respect to whether clinicians' behavior fits neatly into a scanner or driller category (Aizenman et al., 2017; Ba et al., 2020), and the previously observed connections between drilling and diagnostic accuracy (van Montfort et al., 2020). We believe that our current data are compelling in showing that scanning and drilling/zooming are not mutually exclusive strategies in digital breast pathology. Digital pathology can be considered exploration of 2D space at varied resolution where zooming is primarily required when resolution is diagnostically insufficient. Thus zooming may be more critical in complex pathology images where detection of focal findings, like cellular morphology, is more critical to successful diagnosis. In addition to domain-specific differences, there may also be task-specific ideal combinations of scanning and zooming. For example, scanning may be ideal for excluding or identifying an invasive breast cancer on an image; zooming may be more critical for distinguishing proliferative lesions from pre-invasive (in situ) cancer. Future work in this area may benefit from adopting our strategy of examining scanning and zooming separately.

Associating search strategy with diagnostic accuracy

Consistent with prior work (Drew et al., 2013), we found clinicians tended to maintain a consistent strategy across cases (see Figures 3 and 4). However, although there were some individuals who consistently zoomed at a high rate with relatively low scanning scores, there were also individuals who were consistently high on both the scanning and zooming metrics (see Figure 2). Logistic regression analyses with scanning rate and zooming rate as independent variables found evidence that scanning rate is associated with diagnostic accuracy, but not zooming rate. Although the observed association between scanning rate and diagnostic accuracy is interesting, as with any analysis of observational data, we cannot conclude that more scanning will result in higher accuracy. This investigation is part of a larger project where we will observe the changes in behavior and search strategy among pathology residents over their three years of training. These longitudinal analyses may provide converging evidence as to the value of the scanning strategy in diagnostic pathology. Ultimately though, to make strong causal claims about the value of the scanning strategy, we would need to perform an interventional study where pathologists are trained to increase their scanning rate while evaluating cases or placed in a control group. This would be an interesting and challenging area for future investigation.

Based on the prior work in radiology (Drew, Võ, et al., 2013) and dermatopathology (Mello-Thoms et al., 2012), we were surprised that we did not find evidence that zooming was associated with diagnostic accuracy in breast pathology. The simplest explanation for this result is that searching through digital pathology slides engages a diagnostic evaluation that is fundamentally different than searching through chest CT scans while engaged in a screening task. The detection of lung nodules may be particularly advantaged by quick drilling because differentiating lung nodules from lung vessels is aided by rapid changes in the diameter of the structure in question. To our knowledge, there is no equivalent benefit where diagnosis of important characteristics of breast pathology is aided by quickly zooming in or out. As noted in the introduction, prior work has suggested that a strategy similar to what we have defined as "scanning" was associated with poorer outcomes in a dermatopathology task (Mello-Thoms et al., 2012). We suspect that benefit associated with scanning in our study are due to important differences in how pathologists examine breast pathology and dermatopathology images. Dermatopathology slides typically contain multiple tissue segments of the same structure which often contain redundant information. Therefore scanning with these images may be associated with unnecessary rescanning of information that can be perceived on other parts of the image. Clearly, more work, ideally with high-resolution eye-tracking in addition to movement of the viewing window, is necessary to test this interpretation.

More broadly, the idea that quick movement through medical images is a marker of expertise pervades much of the medical image perception literature (for reviews see Brunyé et al. 2019; Reingold and Sheridan 2011; Williams and Drew 2019). Most of this growing literature is devoted to 2D medical images, the most popular examples being chest radiographs and mammograms. If we broaden to the larger field of visual expertise, one of the most consistent correlates of expert visual behavior is saccadic amplitude (Gegenfurtner et al., 2011) where experts tend to make large eye-movements when examining images in their domain of expertise. Consistent with these observations, we found diagnostic accuracy was positively associated with saccadic amplitude even when controlling for time viewing the case and average viewing depth.

The same meta-analysis (Gegenfurtner et al., 2011) that suggested saccadic amplitude as a reliable metric of visual expertise found that response time was nearly as strongly correlated with expertise, such that experts tend to spend less time on the task. In contrast, we did not find a relationship between time devoted to a case and diagnostic accuracy. It is possible that this apparent lack of association is driven by the large differences in experience in our sample. This sample included 20 attending pathologists and 72 trainees of varied experience. Of course, a more direct test of this prediction would be to examine the relationship between experience and time spent per case. Because this is an ongoing study, we hope to address the question of the relationships between interpretive duration, expertise, and diagnostic accuracy once we have a larger sample of attending pathologists.

Before we initiated this study, a common adage we heard from pathologists was "High-power pathologist, low-power microscopist." This well-known adage suggests expert attending pathologists could often discern trainees' competence by observing a reduced tendency of the trainees to examine biopsies at high magnification. This would be a very difficult prediction to evaluate in the glass slide era, but with digital slides and a system that monitors zooming behavior, we were able to test this prediction. Specifically, if the adage were to hold true, one would expect to see higher average zoom level associated with lower accuracy. However, we did not find an association between diagnostic accuracy and average zoom level when controlling for evaluation time and saccadic amplitude. To more directly address this idea, we also computed a logistic regression (Model 3, see Table 2) between accuracy and zoom level, similar to Model 1 in Table 2 above but leaving out saccadic amplitude and interpretive duration. We found no evidence supporting an association (OR = 0.97, p = 0.51) in our large sample of practicing pathologists. This contrasts with prior work that found that experts spent less time at high magnification than novice pathologists (Jaarsma Jarodzka, Nap, van Merriënboer, & Boset al., 2015). Importantly, the "novice" group in this prior work was medical students, whereas all participants in the current work were physicians with the least-experienced participants residents and fellows who were part of an anatomic or anatomic and clinical training program. The lack of experience in the prior work's novice medical student group compared to the current study of physicians might explain the discrepancy in results. As with the prior discussion of time per case, in future work it will be interesting to examine the relationship between zoom level and experience. However, if zoom level does not relate to diagnostic accuracy, we would argue that this would render any relationship with experience less interesting.

The observed association between the tendency to have larger eye movements and diagnostic accuracy may relate to growing evidence that saccadic amplitude is related to the nature of the information being gleaned from the image being evaluated (Tatler, Baddeley, & Vincent, 2006). In picture viewing, longer saccades are associated with low-frequency information, which may relate to architectural and relational features in pathology. In contrast, shorter saccades are associated with high-frequency information, which may correspond to smaller details in pathology, such as variation in contrast. From this perspective, high frequency information may be more likely to contain highly salient information that may not be clinically relevant. This could help explain why saccadic amplitude is one of the strongest correlates of visual expertise across a variety of domains (Gegenfurtner et al., 2011). As mentioned previously, it is interesting to note that this relationship appears to break down in volumetric CT images where it may be difficult to quickly extract a global gist of the case (Williams et al., 2021). This apparent interaction between performance, nature of the stimulus and saccadic amplitude would be consistent with recent advances in our understanding of gaze in nature vision (Tatler, Hayhoe, Land, & Ballard, 2011). According to this theory, gaze allocation is driven by reward maximization and uncertainty reduction. Data from the current investigation suggests that longer saccades and generally more eye-movement in XY space leads to better outcomes in breast pathology, but prior work suggests that this may not be the case in dermatopathology (Mello-Thoms et al., 2012), or volumetric chest CT images (Williams et al., 2021). The basic science literature suggests that this apparent discrepancy is likely due to important differences in the nature of the images being evaluated by clinicians. Although there has been recent modeling work to adapt ideal observer models from 2D medical images to the volumetric stack images like chest CT (Lago, Abbey, & Eckstein, 2021), we are not aware of work to adapt this work to zoomable high-resolution images that are found in digital pathology.

Future directions

A central unanswered question for this investigation is why scanning but not drilling is associated with accuracy in this sample of pathologists. We propose two possible explanations. Perhaps the key to a good search strategy is finding a method that efficiently and systematically moves through the search space. Chest CT is a fundamentally different search environment than digital pathology slides, so it should not be surprising that the same strategy is not advantageous in both venues. Although the images that are being evaluated are very different, medical image perception

researchers argue that there are fundamental factors, such as trade-offs between speed and accuracy, that are consistent across the many different types of search (Wolfe et al., 2016). Despite these broad commonalities across different types of search, it is likely that the optimal search strategy varies with the type of search that is being performed. As we have outlined, drilling quickly through depth appears to be an effective strategy to search through chest CT scans during a lung cancer screening task. However, it is much less clear whether drilling is an effective strategy when engaged in different tasks within radiology. Perhaps most relevant to the current investigation, van Montfort and colleagues (2020) found no association between drilling behavior and diagnostic accuracy when radiologists evaluated a variety of different types of volumetric CT images. In a different domain, when searching for one of several potential target objects in a 2D search task with novices, it appears that a "passive" search strategy, where subjects were encouraged to let the targets pop out while evaluating the scene, yields better performance than a more "active" strategy (Madrid & Hout, 2019). Along similar lines, research with expert Transportation Security Administration baggage screeners has found that the strongest correlate of an individual's search ability is search consistency, defined as the variance in how long they evaluated each image (Biggs, Cain, Clark, Darling, & Mitroff, 2013; Biggs & Mitroff, 2014).

To understand the best search strategies for digital pathology slides, in future work we hope to more carefully consider search coverage and search consistency. One of the common sources of error in medical image perception is "search errors" where the clinician never looks in the lesion's location (Kundel, Nodine, & Carmody, 1978). Drew and colleagues (2013) found that coverage, or the proportion of lung tissue evaluated in chest CT images, was reliably higher for driller than scanners (see also Williams et al., 2021). This increase in coverage may be the underlying reason why the drillers ultimately found more nodules than the scanners in prior studies. Based on the differences in the nature of the images being examined and our current finding that scanning is positively associated with diagnostic accuracy, perhaps increased scanning in digital pathology is associated with improved search coverage. However, it is important to note that it not currently known how often errors in the digital pathology are due to "search errors" where the lesion is not fixated. One challenge to this proposed future direction is that it is not clear how to calculate the Useful Field of View (UFOV) as a function of zoom. Historically, medical image perception research has assumed that expert clinicians have a UFOV of 5° of visual angle (Kundel et al., 2007). Given that a 5° window at the lowest zoom typically covers ~90 times the amount of tissue at the highest zoom, it would seem that a fixation centered on the same structure at different levels of magnification must denote different levels of recognition of the clinical significance of the structure in question. It therefore seems clear that magnification level should be taken into account when determining whether a structure was "covered" or not. We hope to address this conundrum in future work.

Conclusions

With the recent FDA approval of digitized glass slides for primary diagnosis, pathology is in the midst of a revolution in how pathologists are trained and how cases are evaluated (Elmore et al., 2020). Although the degree to which digital whole slide images will replace glass slides is not yet clear, this new method of viewing pathology slides will allow researchers to evaluate search techniques and strategies in a manner that is not possible with glass slides. By tracking eye movements as well as clinician-guided movements through digital pathology slides, the current study documents substantial differences across practicing pathologists evaluating a set of challenging breast pathology whole slide images. We found that scanning rate was positively associated with diagnostic accuracy. This is an important first step in understanding what types of search behaviors and strategies may be associated with improved diagnostic accuracy in breast pathology. More work is necessary to assess whether these results generalize across different tissue types (e.g., skin biopsy) within pathology and how these strategies may be influenced by external factors such as time pressure, prevalence of finding, and patient history.

Keywords: medical image perception, eye-tracking, visual search, pathology

Acknowledgments

The authors thank Ventana Medical Systems, Inc., a member of the Roche Group, for use of iScan Coreo AuTM 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 thank the staff and residents at the various pathology training programs across the U.S. for their participation and assistance in this study.

Supported by the National Cancer Institute of the National Institutes of Health under award number R01 CA225585.

Commercial relationships: none.
Corresponding author: Trafton Drew.
Email: trafton.drew@psych.utah.edu.
Address: Department of Psychology, University of Utah, 380 S 1530 E Beh S 1003, Salt Lake City, UT 84112, USA.

References

Aizenman, A., Drew, T., Ehinger, K. A., Georgian-Smith, D., & Wolfe, J. M. (2017). Comparing search patterns in digital breast tomosynthesis and full-field digital mammography: an eye tracking study. *Journal of Medical Imaging, 4*(4), 045501.

Allison, K. H., Reisch, L. M., Carney, P. A., Weaver, D. L., Schnitt, S. J., O'Malley, F. P., Geller, B. M., . . . Elmore, J. G. (2014). Understanding diagnostic variability in breast pathology: Lessons learned from an expert consensus review panel. *Histopathology, 65*(2), 240–251, https://doi.org/10.1111/his.12387.

Ba, A., Shams, M., Schmidt, S., Eckstein, M. P., Verdun, F. R., & Bochud, F. O. (2020). Search of low-contrast liver lesions in abdominal CT: the importance of scrolling behavior. *Journal of Medical Imaging, 7*(04), 1–12.

Biggs, A. T., Cain, M. S., Clark, K., Darling, E. F., & Mitroff, S. R. (2013). Assessing visual search performance differences between Transportation Security Administration Officers and nonprofessional visual searchers. *Visual Cognition, 21*(3), 330–352, https://doi.org/10.1080/13506285.2013.790329.

Biggs, A. T., & Mitroff, S. R. (2014). Different Predictors of Multiple-Target Search Accuracy between Nonprofessional and Professional Visual Searchers. *Quarterly Journal of Experimental Psychology, 67*(7), 1335–1348, https://doi.org/10.1080/17470218.2013.859715.

Božić-Štulić, D., Marušić, Ž., & Gotovac, S. (2019). Deep Learning Approach in Aerial Imagery for Supporting Land Search and Rescue Missions. *International Journal of Computer Vision, 127*(9), 1256–1278, https://doi.org/10.1007/s11263-019-01177-1.

Brams, S., Ziv, G., Levin, O., Spitz, J., Wagemans, J., Williams, A. M., . . . Helsen, W. F. (2019). The relationship between gaze behavior, expertise, and performance: A systematic review. *Psychological Bulletin, 145*(10), 980–1027.

Brunyé, T. T., Drew, T., Kerr, K. F., Shucard, H., Weaver, D. L., & Elmore, J. G. (2020). Eye tracking reveals expertise-related differences in the time-course of medical image inspection and diagnosis. *Journal of Medical Imaging, 7*(05), 051203, https://doi.org/10.1117/1.JMI.7.5.051203.

Brunyé, T. T., Drew, T., Weaver, D. L., & Elmore, J. G. (2019). A review of eye tracking for understanding and improving diagnostic interpretation. *Cognitive Research: Principles and Implications, 4*(1), 1–16.

Dahabreh, I. J., Wieland, L. S., Adam, G. P., Halladay, C., Lau, J., & Trikalinos, T. A. (2014). Core needle and open surgical biopsy for diagnosis of breast lesions: an update to the 2009 report.

Drew, T., Evans, K., Võ, M. L.-H., Jacobson, F. L., & Wolfe, J. M. (2013). Informatics in radiology: What can you see in a single glance and how might this guide visual search in medical images? *Radiographics, 33*(1), 263–274.

Drew, T., Võ, M. L.-H., Olwal, A., Jacobson, F., Seltzer, S. E., & Wolfe, J. M. (2013). Scanners and drillers: Characterizing expert visual search through volumetric images. *Journal of Vision, 13*(10), 3–3.

Elmore, J. G., Longton, G. M., Carney, P. A., Geller, B. M., Onega, T., Tosteson, A. N. A., Nelson, H. D., Pepe, M. S., Allison, K. H., Schnitt, S. J., O'Malley, F. P., . . . Weaver, D. L. (2015). Diagnostic Concordance Among Pathologists Interpreting Breast Biopsy Specimens. *JAMA, 313*(11), 1122–11.

Elmore, J. G., Longton, G. M., Pepe, M. S., Carney, P. A., Nelson, H. D., Allison, K. H., Geller, B. M., Onega, T., Tosteson, A. N. A., Mercan, E., Shapiro, L. G., Brunyé, T. T., Morgan, T. R., . . . Weaver, D. L. (2017). A Randomized Study Comparing Digital Imaging to Traditional Glass Slide Microscopy for Breast Biopsy and Cancer Diagnosis. *Journal of Pathology Informatics, 8*, 12, https://doi.org/10.4103/2153-3539.201920.

Elmore, J. G., Nelson, H. D., Pepe, M. S., Longton, G. M., Tosteson, A. N. A., Geller, B., Onega, T., Carney, P. A., Jackson, S. L., Allison, K. H., . . . Weaver, D. L. (2016). Variability in Pathologists' Interpretations of Individual Breast Biopsy Slides: A Population Perspective. *Annals of Internal Medicine, 164*(10), 649–655, https://doi.org/10.7326/M15-0964.

Elmore, J. G., Shucard, H., Lee, A. C., Wang, P.-C., Kerr, K. F., Carney, P. A., Drew, T., Brunyé, T. T., . . . Weaver, D. L. (2020). Pathology Trainees' Experience and Attitudes on Use of Digital Whole Slide Images. *Academic Pathology, 7*, 237428952095192, https://doi.org/10.1177/2374289520951922.

Evans, A. J., Bauer, T. W., Bui, M. M., Cornish, T. C., Duncan, H., Glassy, E. F., Hipp, J., McGee, R. S., Murphy, D., Myers, C., O'Neill, D. G., Parwani, A. V., Rampy, B. A., Salama, M. E., . . . Pantanowitz, L. (2018). US Food and Drug Administration Approval of Whole Slide Imaging for Primary Diagnosis: A Key Milestone Is Reached and New Questions Are Raised. *Archives of Pathology & Laboratory Medicine, 142*(11), 1383–1387, https://doi.org/10.5858/arpa.2017-0496-CP.

Feng, S., Weaver, D. L., Carney, P. A., Reisch, M., L., M. Geller, B., Goodwin, A., Rendi, M., H., Onega, T., Allison, K. H., Tosteson, A. N. A., Nelson, H. D., Longton, G., Pepe, M., . . . Elmore, J. G. (2014). A Framework for Evaluating Diagnostic Discordance in Pathology Discovered During Research Studies. *Archives of Pathology & Laboratory Medicine, 138*(7), 955–961, https://doi.org/10.5858/arpa.2013-0263-OA.

Gandomkar, Z., & Mello-Thoms, C. (2019). Visual search in breast imaging. *The British Journal of Radiology, 92*(1102), 20190057, https://doi.org/10.1259/bjr.20190057.

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise Differences in the Comprehension of Visualizations: A Meta-Analysis of Eye-Tracking Research in Professional Domains. *Educational Psychology Review, 23*(4), 523–552.

Jaarsma, T., Jarodzka, H., Nap, M., van Merriënboer, J. J. G., & Boshuizen, H. P. A. (2015). Expertise in clinical pathology: Combining the visual and cognitive perspective. *Advances in Health Sciences Education, 20*(4), 1089–1106, https://doi.org/10.1007/s10459-015-9589-x.

Kundel, H. L., Nodine, C. F., & Carmody, D. (1978). Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Investigative Radiology, 13*(3), 175.

Kundel, H. L., Nodine, C. F., Thickman, D., & Toto, L. C. (1987). Searching for Lung Nodules A Comparison of Human Performance with Random and Systematic Scanning Models. *Investigative Radiology, 22*(5), 417.

Kundel, H., Nodine, C., Conant, E., & Weinstein, S. (2007). Holistic Component of Image Perception in Mammogram Interpretation: Gaze-tracking Study. *Radiology, 242*(2), 396.

Lago, M. A., Abbey, C. K., & Eckstein, M. P. (2021). Foveated Model Observers for Visual Search in 3D Medical Images. *IEEE Transactions on Medical Imaging, 40*(3), 1021–1031, https://doi.org/10.1109/TMI.2020.3044530.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764–766, https://doi.org/10.1016/j.jesp.2013.03.013.

Madrid, J., & Hout, M. C. (2019). Examining the effects of passive and active strategies on behavior during hybrid visual memory search: Evidence from eye tracking. *Cognitive Research: Principles and Implications, 4*(1), 39, https://doi.org/10.1186/s41235-019-0191-2.

Mello-Thoms, C., Mello, C. A., Medvedeva, O., Castine, M., Legowski, E., Gardner, G., Tseytlin, E., . . . Crowley, R. (2012). Perceptual Analysis of the Reading of Dermatopathology Virtual Slides by Pathology Residents. *Archives of Pathology & Laboratory Medicine, 136*(5), 551–562, https://doi.org/10.5858/arpa.2010-0697-OA.

Mercan, E., Shapiro, L. G., Brunyé, T. T., Weaver, D. L., & Elmore, J. G. (2018). Characterizing Diagnostic Search Patterns in Digital Breast Pathology: Scanners and Drillers. *Journal of Digital Imaging, 31*(1), 32–41, https://doi.org/10.1007/s10278-017-9990-5.

Oster, N. V., Carney, P. A., Allison, K. H., Weaver, D. L., Reisch, L. M., Longton, G., Onega, T., Pepe, M., Geller, B. M., Nelson, H. D., Ross, T. R., Tosteson, A. N. A., . . . Elmore, J. G. (2013). 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), 3, https://doi.org/10.1186/1472-6874-13-3.

Reingold, E. M., & Sheridan, H. (2011). Eye movements and visual expertise in chess and medicine. In: S. P. Liversedge, I. D. Gilchrist, & S. Everling (Eds.), *Oxford Handbook on Eye Movements* (pp. 523–550). Oxford, UK: Oxford University Press.

Seltzer, S. E., Judy, P. F., Adams, D. F., Jacobson, F. L., & Stark, P. (1995). Spiral CT of the chest: Comparison of cine and film-based viewing. *Radiology, 197*(1), 73–78.

Tatler, B. W., Baddeley, R. J., & Vincent, B. T. (2006). The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision Research, 46*(12), 1857–1862, https://doi.org/10.1016/j.visres.2005.12.005.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision, 11*(5), 5–5, https://doi.org/10.1167/11.5.5.

van Montfort, D., Kok, E., Vincken, K., van der Schaaf, M., van der Gijp, A., Ravesloot, C., . . . Rutgers, D. (2021). Expertise development in volumetric image interpretation of radiology residents:

What do longitudinal scroll data reveal? *Advances in Health Sciences Education, 26*(2), 437–466, https://doi.org/10.1007/s10459-020-09995-6.

Wen, G., Aizenman, A., Drew, T., Wolfe, J. M., Haygood, T. M., & Markey, M. K. (2016). Computational assessment of visual search strategies in volumetric medical images. *Journal of Medical Imaging, 3*(1), 015501.

Williams, L. H., Carrigan, A. J., Mills, M., Auffermann, W. F., Rich, A. N., & Drew, T. (2021). Characteristics of expert search behavior in volumetric medical image interpretation. *Journal of Medical Imaging, 8*(04), 041208, https://doi.org/10.1117/1.JMI.8.4.041208.

Williams, L. H., & Drew, T. (2019). What do we know about volumetric medical image interpretation?: A review of the basic science and medical image perception literatures. *Cognitive Research: Principles and Implications, 4*(1), 21.

Wolfe, J. M., Evans, K. K., Drew, T., Aizenman, A., & Josephs, E. (2016). How do radiologists use the human search engine? *Radiation Protection Dosimetry, 169*(1–4), 24–31.

Wu, C.-C., & Wolfe, J. M. (2019). Eye Movements in Medical Image Perception: A Selective Review of Past, Present and Future. *Vision, 3*(2), 32, https://doi.org/10.3390/vision3020032.