

UCLA

UCLA Previously Published Works

Title

From Image to Diagnosis: Characterizing Sources of Error in Histopathologic Interpretation.

Permalink

<https://escholarship.org/uc/item/1q73121r>

Journal

Modern Pathology, 36(7)

Authors

Brunyé, Tad

Balla, Agnes

Drew, Trafton

et al.

Publication Date

2023-07-01

DOI

10.1016/j.modpat.2023.100162

Peer reviewed



HHS Public Access

Author manuscript

Mod Pathol. Author manuscript; available in PMC 2024 September 10.

Published in final edited form as:

Mod Pathol. 2023 July ; 36(7): 100162. doi:10.1016/j.modpat.2023.100162.

From Image to Diagnosis: Characterizing Sources of Error in Histopathologic Interpretation

Tad T. Brunyé^{a,b,*}, Agnes Balla^e, Trafton Drew^c, Joann G. Elmore^f, Kathleen F. Kerr^d, Hannah Shucard^d, Donald L. Weaver^e

^aCenter for Applied Brain and Cognitive Sciences, Tufts University, 177 College Ave., Suite 090, Medford, MA 02155

^bDepartment of Psychology, Tufts University, 490 Boston Ave., Medford, MA 02155

^cDepartment of Psychology, University of Utah, 380 S 1530 E Beh S 502, Salt Lake City, UT 84112

^dDepartment of Biostatistics, University of Washington, 1705 NE Pacific Street, Seattle, WA 98195

^eDepartment of Pathology, University of Vermont and Vermont Cancer Center, 89 Beaumont Ave., Burlington, VT 05405

^fDavid Geffen School of Medicine, Department of Medicine, University of California, Los Angeles, 885 Tiverton Drive, Los Angeles, CA 90095

Abstract

An accurate histopathologic diagnosis on surgical biopsy material is necessary for the clinical management of patients and has important implications for research, clinical trial design/enrollment, and public health education. This study used a mixed methods approach to isolate sources of diagnostic error while residents and attending pathologists interpreted digitized breast biopsy slides. Ninety participants including pathology residents and attendings at major United States medical centers reviewed a set of 14 digitized whole slide images of breast biopsies. Each case had a consensus-defined diagnosis and critical region of interest (cROI) representing the most significant pathology on the slide. Participants were asked to view unmarked digitized slides, draw their own participant region of interest (pROI), describe its features, and render a diagnosis. Participants' review behavior was tracked using case viewer software and an eye tracking device. Diagnostic accuracy was calculated in comparison to the consensus diagnosis. We measured the frequency of errors emerging during four interpretive phases: 1) detecting the cROI, 2) recognizing

*Corresponding Author: Tad T. Brunyé, tbruny01@tufts.edu.

Author Contributions

JGE, DLW, TTB, KFK, and TD conceived and designed this study. TTB and TD collected the data. HS managed pathologist recruitment and scheduling, and provided feedback and suggestions on manuscript drafts. DLW and AB provided expert annotation ratings. TTB drafted the manuscript and TTB and KFK performed the statistical analyses. TD, KFK, DLW, AB, and JGE reviewed the manuscript at several times during preparation, providing feedback and suggestions. All authors read and approved the final manuscript.

Ethics Approval and Consent to Participate

All participants provide written informed consent in accordance with Institutional Review Board approvals granted by the University of California Los Angeles.

its relevance, 3) using the correct terminology to describe findings in the pROI, and 4) making a diagnostic decision. According to eye tracking data, both trainees and attending pathologists were very likely (about 94% of the time) to find the cROI when inspecting a slide. However, trainees were less likely to consider the cROI relevant to their diagnosis. Pathology trainees were more likely (41% of cases) to use incorrect terminology to describe pROI features than attending pathologists (21% of cases). Failure to accurately describe features was the only factor strongly associated with an incorrect diagnosis. Identifying where errors emerge in the interpretive and/or descriptive process and working on building organ-specific feature recognition and verbal fluency in describing those features are critical steps for achieving competency in diagnostic decision making.

Keywords

breast cancer; pathology; decision-making; medical image interpretation; eye-tracking; diagnosis; medical education; accuracy

Introduction

Diagnostic errors are a critical challenge in all areas of health care¹, however the sources of error are unique to each specialty/sub-specialty in medicine. While many areas of medicine rely on clinical history, physical exam, and lab values to render a diagnosis, the field of anatomical/surgical pathology is unique in that integrates the aforementioned information with gross, microscopic, and where relevant, molecular findings to render a tissue diagnosis. Like other fields of medicine, pathology requires a medical degree, followed by post-graduate residency training, and then often an organ-specific subspecialty fellowship to acquire career-specific expertise. In their daily work, surgical pathologists are tasked with recognizing and integrating complex visual images present on microscopic slides and assigning clinically actionable organ-specific language/terminology to their diagnostic reports. This process relies on the pathologist to identify the area of most significant pathology in a background of normal but variable histology, recognize the relevance of the pathologic features, then use all available clinical and pathological data to translate the visual finding(s) into a written diagnosis using universally accepted terminology. This entire process of arriving at a final diagnosis requires the brain to engage in a multitude of complex cognitive processes.^{2,3} Isolating these processes and identifying where errors emerge is critical for educating future pathologists and for reducing errors with serious downstream clinical implications and monetary costs to the healthcare system.^{4,5}

To isolate and better understand the processes involved in interpreting biopsies, we recruited pathology trainees and board-certified attending pathologists to view a set of whole slide breast biopsy images while we tracked their eye movements and viewing behavior. Breast biopsies were chosen for this study because they are relatively high volume in pathology residency training, have a well-established and universally accepted lexicon of terminology, and can show a spectrum of pathologic findings (both benign and malignant) that can be challenging for trainees to interpret. We developed a conceptual framework (Fig. 1) that distinguished four interpretive phases and potential sources of diagnostic error: detecting

the critical region of interest, recognizing its relevance in the context of the background, accurately applying terminology to describe salient features, and finally reaching a correct diagnosis. *Detecting a critical region* involves visually scanning an image to find regions that represent the most significant pathology on the slide. In our study, a failure to detect these regions occurs when a participant does not fixate their eyes on a consensus-defined region of interest (cROI) measured by eye tracking software.^{2,6} *Recognizing relevance* is the process of accumulating visual evidence and recognizing patterns that are potentially informative to solving a task.^{7,8} In our study, a failure to recognize relevance occurs when participants fixate their eyes on the cROI but the drawn participant region of interest (pROI) does not overlap with the cROI. *Describing features* is the process of using correct terminology to describe the microscopic pathologic features in the pROI in a written annotation. We consider a failure of description to occur when a participant's written annotations do not correctly describe the features present in the pROI. Finally, *diagnostic decision-making* is the process of developing hypotheses and ultimately deciding to place a case into a diagnostic category.⁹⁻¹¹ A failure of decision-making occurs when a participant's diagnosis does not match the case consensus diagnosis.

Identifying where errors emerge during the interpretive process is critical for not only informing residency curricula and assessments but also to identify diagnostic categories that might have inherently high inter-observer variability and are therefore difficult to teach. The Accreditation Council for Graduate Medical Education (ACGME) emphasizes competency-based training and evaluation, and training programs and their residents stand to benefit from curricula that can accelerate and expand competency development.^{12,13} However, one challenge facing the realization of competency-based training is understanding precisely where competency gaps occur, and therefore which aspects of curriculum and assessment require more attention. Our study examines four possible competency gaps in the interpretive process: detecting critical regions, recognizing relevance of critical regions, describing perceived features, and diagnostic decision-making. Based on our results, we provide concrete recommendations for filling identified gaps with pedagogical approaches derived from the cognitive and learning sciences.

It is known that inter-observer variability exists even among experienced pathologists in certain breast pathology diagnostic categories (e.g., atypia, low-grade ductal carcinoma in situ), and these cases may lead to slightly different diagnostic interpretations.^{14,15} The purpose of the current study was to further explore the genesis of diagnostic errors and disagreement, particularly during the learning phase of competency. We used our novel conceptual framework to examine the nature and frequency of errors in the four interpretive phases. We also compared the frequency of these error types among participants with relatively low (residents, fellows) versus high (attending physicians) experience levels.

Methods

Human Research Participants Protection

All participants provided written informed consent, and all study procedures were approved by the appropriate Institutional Review Boards (IRB), with the University of California, Los Angeles acting as the IRB of record (Protocol #18-000327).

Participating Pathologists

Data were collected from ninety individuals with varying experience in interpreting breast pathology, including 70 pathology trainees (residents and fellows), and 20 general surgical pathology attending physicians, from nine major United States university medical centers (see Supplementary Table 1 for demographic details). Note that 8 of the 9 sites used sub-specialty sign-out models in their training programs (with one site transitioning from general to sub-specialty within the timeframe of our data collection), with the other site using a general sign-out model.

Test Set Development

We used cases from a larger test set of 240 hematoxylin and eosin stained digital whole slide images (WSI) developed in earlier research.^{16,17} Each case was scanned using an iScan Coreo Au digital slide scanner¹⁸ at 40x objective magnification (0.23 um per pixel maximum optical resolution); a single image was created for each case. A subset of 32 cases was selected for this study, each comprised of one WSI that included one or more cROI(s), and a single consensus reference diagnosis previously determined by a panel of three fellowship-trained expert breast pathologists using a modified Delphi technique.^{14,16} In our prior work, the expert panel members independently reviewed each case and then attended four in-person meetings to come to agreement on a consensus reference diagnosis for each case and to ensure that each WSI provided all necessary image detail to render a primary diagnosis.¹⁴ The cROI was agreed upon by the expert panel as the image region(s) best representing the most clinically significant consensus diagnosis. None of the expert panel members was a participant in this study.

The 32 cases selected included a full range of consensus diagnostic categories and were divided into three test sets of 14 cases (one image per case). Five identical cases (one from each diagnostic category) were included in all three test sets, and the remaining 9 cases within each test set were unique. In total, each test set contained two benign cases, four atypia cases, four low-grade ductal carcinoma in situ (DCIS) cases, two high-grade DCIS cases, and two invasive cases. Within the DCIS category, we used accepted research practice¹⁹ to parse cases into low-grade versus high-grade based upon nuclear grade and the presence of necrosis. As a practice case, we selected one invasive carcinoma case with very high rates (93%) of participants agreeing with the consensus diagnosis in prior research.¹⁶

Note that cROIs provided less discriminatory power for cases diagnosed as benign; benign cases are more likely than the other diagnostic categories to have many different image areas that could be interpreted as supporting the benign diagnosis. Thus, eye fixation and pROI overlap are more difficult to interpret in the benign category.

Study Equipment

A Dell Precision M4800 laptop and a color-calibrated 22" Dell liquid crystal display (LCD) running at 1920 × 1080 resolution were used for data collection. A digital viewer software was developed to depict images using the Microsoft, Inc. DeepZoom Silverlight application. The viewer displayed the high resolution digital whole slide images, and users could zoom (up to 60× magnification equivalent) and pan the image while maintaining high

resolution. The viewer software included a rectangle drawing tool (i.e., to draw the pROI), and continuously recorded all image navigation data including zooming, panning, and pROI markings. The position of the cROI(s) was known to the research team but was not visible to participants.

A remote eye tracking device manufactured by SensoMotoric Instruments (Boston, MA; Model RED250) was used to collect binocular eye movement data 250Hz while maintaining high gaze position accuracy ($\pm 0.5^\circ$). The eye tracker was attached to the bottom of the computer monitor, and we used a 9-point calibration process.

To record annotations and diagnostic decisions, we developed a histology form using the Qualtrics web-based platform. The histology form used an open-ended text entry box to collect descriptive text annotations of pROI features (i.e., *What are the critical histopathological features in the ROI you drew? Be specific.*), the case's diagnostic category and a Likert-based confidence rating. When relevant, the form collected additional data such as nuclear grade and the presence or absence of necrosis.

Study Locations and Procedures

Following informed consent and completing a demographic survey, at each of the nine data collection sites, participants met with an experimenter (authors TTB or TD) individually (for up to 1.5 hours) to review cases in a private office or conference room (Fig. 2). Following an eye tracker calibration, all participants practiced navigating the image (zooming, panning), marking a critical region of the image (drawing a pROI), providing a textual annotation of features found in the pROI (like on a pathology report), and then completing the remainder of the histology form using the same practice case. This process was then completed for each of the 14 experimental cases (in random order). Each of the experimental cases included only standardized and basic clinical history (patient age, biopsy type) that was intentionally of limited value for interpreting the histopathological features of a case; this allowed our study to focus on information derived solely from review of the case. Upon completion of the study session, participants were compensated with a \$50USD gift card.

Data Scoring & Analysis

Eye-tracking data were merged with case position information gathered from the case viewer software, allowing us to co-register (temporally and spatially) eye fixations to the case coordinate system. The eye tracker parses continuous eye movements into fixations and saccades; fixations are brief pauses of the eye that exceed a temporal threshold (80 ms), and saccades are rapid ballistic movements of the eye between successive fixations. When examining eye movements, we used the version signal, a measure that reduces variable error by averaging left and right eye positions.²⁰

To assess critical region detection, we assessed whether a participant fixated their eyes at least once (for longer than 80 ms) on the cROI during their review. A detection error was defined as having zero eye fixations on a case's cROI(s). To assess recognizing the relevance of critical regions, we evaluated whether the pROI overlapped (in pixel area) with the cROI: a recognizing relevance error was defined as having zero overlap between the pROI and cROI.

To assess whether participants accurately described features, we asked two board-certified anatomic and clinical pathologists (authors DW and AB, with the former being more senior than the latter) with over 45 years of combined sub-specialty expertise in breast pathology to manually score participants' written annotations. They viewed each case, the drawn pROIs, and the written annotations provided by the participants. They assessed whether the written annotation used language that accurately and sufficiently described the features in the drawn pROI by selecting one of six assessment categories (Table 1). In the first round, each rater independently assessed all interpretations with annotations (1,211 annotations). After the first round, the two raters had an overall agreement of 67.8% in categorizing the written annotations (821 of the 1,211 annotations; see Supplementary Table 2 for details). The two expert raters then re-reviewed the 390 annotations that were categorized differently while also being able to view the other rater's assessments and discussed interpretations among themselves. Most disagreements were due to the rater's interpretations in language when non-standard, non-universal terminology was used by participants, and some variability in how raters interpreted the "ambiguous" category. In instances where the terminology used in descriptors was not universally accepted in the lexicon of breast pathology terms, raters had to make inferences about what the participant may have intended with their descriptors. At the end of the second round, the two raters achieved 94.2% agreement (1,141 of the 1,211 annotations; Table 1). See Supplementary Table 3 for example scored annotations.

For analysis, we excluded 136 feature annotations where raters answered "*Ambiguous, difficult to judge*" or disagreed whether the annotation accurately represented features (Table 1, light gray shading). The remaining 1,075 assessments (distributed across all 90 participants) were divided into two categories: one when the annotation accurately represented the features (Table 1, white shading, 685 ratings), and one when it inaccurately represented the features (Table 1, dark grey shading, 390 ratings).

To assess errors of diagnostic decision making, we compared participant diagnoses to the consensus reference diagnosis for each case; an error was defined as the participant response for the case mismatching the single consensus (of five) diagnostic category.

To explore the contribution of each interpretive phase towards an accurate diagnosis, we fit a generalized estimating equation (GEE) logistic model using SPSS v21 (IBM, Inc., Armonk, NY) with diagnostic accuracy as the binary outcome and three binary explanatory variables: detecting the lesion, recognizing its relevance, and describing features. We included case ID as fixed effects and each participant as a cluster.

Results

All 1,075 interpretations were included in the analyses; Table 2 details the number and proportion of successful interpretations within each of the four phases of the interpretive process.

For *detecting the critical region*, participants fixated on the cROI on 1,012 interpretations (94%). Overall, errors at this phase were low (6%).

For *recognizing the relevance* of a region, participants drew a pROI that overlapped the cROI on 727 interpretations (68%); mean area of pROI:cROI overlap was 35%. Except for invasive carcinoma, where overlap of ROI was high among all participants, attending physicians were 16% more likely to overlap their pROI with the cROI than pathology trainees (76% vs 65%).

For *describing features*, participants successfully described histopathological features in drawn ROIs on 685 interpretations (64%). Overall, errors of feature description occurred on 36% of interpretations; more challenging diagnostic categories (atypia, DCIS) and relatively inexperienced participants tended to show higher error rates at this stage of the interpretive process.

For *diagnostic decision making*, participants successfully selected a diagnosis on a case that matched the expert consensus diagnostic category on 468 interpretations (44%), substantially higher than random chance (20%). Decision making errors were generally higher among pathology trainees and in the more difficult diagnostic categories (atypia, DCIS).

We plotted Sankey diagrams to visualize where errors tend to occur (Fig. 3). Sankey diagrams are used to depict complex flows of information; the width of a flow is proportional to its quantity (i.e., the number of interpretations), branches represent change in flow (e.g., detecting or not detecting a cROI), and colors represent phases (e.g., detection, recognition).²¹ In the upper panel, the overall error rate during the *Detecting the Region* phase is 5.9% (63 of 1,075 cases), and the dependent (sequential) error rate at *Recognizing Relevance* is 28.4% (287 of 1,012 cases), at *Describing Features* is 36.1% (262 of 725 cases), and at the *Diagnostic Decision* phase is 35.4% (164 of 463 cases). Given the challenging categories included in our test set, these overall error rates were expected^{14,19,22}.

Medical decision making is a complicated process. Even without detecting the critical image region or accurately describing features, a pathologist can still arrive at the consensus diagnosis. These possibilities are detailed in Table 3, organized by the accuracy of the diagnosis. In this table, we detail the eight possible paths to a successful diagnosis (rows 1–8), and the eight paths to a diagnostic error (rows 9–16), both sorted in descending order of likelihood.

Of the 63 interpretations when the participants did not detect the critical region (Table 3 rows 4, 6–8, 13–16), only two recognized the relevance of the cROI (row 15); neither resulted in successfully describing features or an accurate diagnosis. Of the 348 interpretations when the importance of the cROI is not recognized (regardless of whether the cROI was detected; rows 2, 4–6, 11–14), participants accurately described features about two-thirds of the time (222 of 348 interpretations; 64%). This is the same rate as when the cROI overlaps the pROI (463 out of 725 = 64%); recall that recognizing features is assessed with respect to each participant's drawn ROI (not with respect to the consensus cROI).

Finally, interpretations wherein the features are not successfully described (Table 3 rows 3, 5–7, 11, 14–15) have a very low chance of ending with an accurate diagnosis (50 of 390; 12.8%). A GEE logistic regression showed that describing features, $\chi^2 = 271.1$, OR =

10.71 (95% CI: 8.1, 14.2), $p < .001$, was about eight-times more strongly associated with diagnostic accuracy than recognizing relevance of features, $\chi^2 = 2.78$, OR = 1.27 (95% CI: 0.96, 1.67), $p = .10$. When including covariates for participant experience level and the consensus diagnostic category of the case, describing features remained as the variable with the strongest association with diagnostic accuracy, $\chi^2 = 261.8$, OR = 10.37 (95% CI: 7.8, 13.8), $p < .001$, about eight-times stronger than recognizing relevance, $\chi^2 = 1.69$, OR = 1.21 (95% CI: 0.91, 1.63), $p = .19$.

A final logistic regression demonstrated that higher experience level was associated with a higher likelihood of accurate feature description, with attending physicians showing nearly 3 times higher odds of accurately describing features compared to trainees, controlling for case, $\chi^2 = 16.9$, OR = 2.96 (95% CI: 1.8, 4.9), $p < .001$. For consensus diagnostic category, feature description showed lowest accuracy for high-grade DCIS (35.5%; OR = 0.14, 95% CI: 0.1–0.2), Atypia (56.3%, OR = 0.35, 95% CI: 0.2–0.5), and low-grade DCIS (62.1%, OR = 0.45, 95% CI: 0.3–0.7) categories, $\chi^2 = 124.9$, $p < 0.001$. These diagnostic categories historically show high interobserver diagnostic variability¹⁴, possibly due to challenges associated with subtleties in their histopathologic features and difficulty in adequately interpreting and describing those subtle differences.

Exploratory Analyses

Early research examining error sources during medical image interpretation was conducted with radiologists examining x-ray films²³. In this research, eye-tracking was used to monitor the position of the eyes during medical image inspection, and the investigators studied whether radiologists were committing search errors versus recognition and/or decision errors. Search errors were operationalized as a failure to fixate the eyes upon a critical lesion (e.g., lung nodule), which occurred in about 30% of all errors. Recognition errors were operationalized as restricted dwell time on the nodule, which occurred in about 25% of all errors. Dwell time is the cumulative amount of time (typically expressed in milliseconds, e.g., 600 ms) that the eyes fixated within a circumscribed region (typically within a few degrees of a target, e.g., 2.8°). Finally, decision errors were noted (about 45% of the time) when a radiologist found (according to eye fixations) and recognized (according to dwell time) the nodule, but then did not appropriately map their recognition to diagnostic criteria.

While many subsequent studies adopted a similar recognition metric based on dwell time^{6,24–26}, some research challenges the appropriateness of this metric as an indicator of successful recognition^{2,27}. Our novel conceptual framework allows us to distinguish between recognizing the relevance of features, and accurately describing histopathological features. However, it is unknown whether our novel measures are related to traditional measures of dwell time, and possibly carry more value than dwell time for predicting diagnostic outcomes. We built two GEE models to examine associations between dwell time, our novel recognition relevance and description measures, and diagnostic accuracy. The first model assessed the association between a binary dwell time measure (i.e., greater than or less than 600 ms dwell time on dROI, as done in prior work^{6,24–26}) with recognizing relevance and describing features. Binary dwell time was positively associated with recognizing the relevance of features, $\chi^2 = 71.1$, OR = 44.55 (95% CI: 18.4, 107.7),

$p < .001$, and negatively associated with describing features, $\chi^2 = 11.6$, OR = 0.39 (95% CI: 0.23, 0.68), $p < .001$. In other words, the traditional method of using dwell time as a measure of recognition appears to be like our recognizing relevance measure (i.e., pROI and cROI overlap). However, dwell time appears to be negatively related to accurately describing features, possibly because increased dwell time can indicate uncertainty or lead to an over-interpretation of features.^{27–31} In other words, prior work may have misinterpreted dwell time as indicative of successful recognition; instead, it appears to more accurately indicate uncertainty.

The second GEE model used diagnostic accuracy as a binary outcome with three independent variables: recognizing relevance, describing features, and a binary measure of dwell time (i.e., greater than or less than 600 ms dwell time). This GEE demonstrated the strong positive association between feature description and diagnostic accuracy, $\chi^2 = 264.1$, OR = 10.6 (95% CI: 7.9, 14.1), $p < .001$, relative to feature relevance, $\chi^2 = 2.9$, OR = 1.4 (95% CI: 0.9, 1.9), $p = .09$, or dwell time, $\chi^2 = 0.6$, OR = 0.8 (95% CI: 0.5, 1.4), $p = .44$. In other words, our feature description measure is about 13-times more strongly related to diagnostic accuracy than the traditional dwell time measure.

Discussion

The present study assessed error sources while trainees and attending pathologists interpreted whole slide images of breast biopsies. Participants were very likely, on about 94% of interpretations, to find critical regions while scanning slides. Likewise, participants recognized the importance of critical regions on about 68% of their interpretations. Interestingly, neither of these measures was associated with diagnostic accuracy. When examining our innovative measure of describing histological features in drawn ROIs, participants applied incorrect descriptors on 36% of their cases. This pattern was especially pronounced in trainees (41% incorrect) relative to relatively experienced attending pathologists (21% incorrect).

Incorrect feature description was the only variable significantly associated with diagnostic accuracy. Unsuccessful feature descriptions were associated with an accurate diagnosis only 13% of the time, whereas successful descriptions were associated with an accurate diagnosis 61% of the time. The ability to correctly describe histologic features is an important determinant of diagnostic accuracy and may be largely determined by the experience level of pathologists and their specific experience in an organ system. Teaching pathology trainees how to accurately recognize *and* describe histologic features may have the highest yield for achieving diagnostic accuracy during training and clinical practice.

Implications for Postgraduate Pathology Training

While the most impactful errors appeared to emerge during the recognition and description of histopathological features, it is important to consider methods for advancing competency throughout the interpretive process. During the initial *detection and relevance assessment* of critical slide regions, pathologists may benefit from support in identifying the morphological features on cases that should attract immediate attention as potentially pathologic. For experienced pathologists, relevant features tend to “pop out” when viewing a case because

they have many years of experience reinforcing value-based (i.e., which features are rewarding and which are not) assessments of visual features.^{2,5,32} Visual attention towards especially salient (e.g., in color, size, contrast, brightness, form) features is automatic, but for residents it can be challenging to recognize and weight visual dimensions that are more relevant to the task.^{33,34} How can trainees be taught to effectively prioritize attention during an initial global image analysis, shielding their visual search from the potentially distracting effects of salient but irrelevant features?³⁵ Research with eye movement modeling examples suggests that when medical trainees view an expert's eye movements over cases, it can help them learn which regions are most informative to their task.^{36,37} This practice is done regularly when trainees view cases with an attending pathologist through a double headed microscope at sign out where they have the opportunity to watch them move to a higher objective and study certain areas on a slide. The more time a trainee can spend at a microscope viewing cases with an attending, the more they can observe this process and hone this skill. When residents are multi-tasking (i.e., typing reports, taking notes) and not looking through the microscope, they may be missing opportunities to engage in this process. Residents benefit from seeing how experienced pathologists navigate a slide, helping them learn statistical regularities associated with the informativeness of case regions. To supplement this time, rather than looking at static images, trainees may also benefit in the future from superimposed heat maps of search patterns on whole slide imaging cases, which can visually highlight probable regions of interest.³⁸

When pathologists were asked to *describe histopathological features* deemed important to the diagnostic task, residents tended to show more challenges than attending physicians. It is not surprising that attending physicians could extract relevant information from cases and more adequately describe it than residents, as this is one hallmark characteristic of emergent expertise.^{38–40} There are several processes involved in this ability, with critical reliance on pattern recognition (perceptual learning) and verbal description.⁴¹ In the cognitive sciences, perceptual learning is regarded as one of the most critical aspects of expertise development in nearly any real-world domain, and extended experience is often the gold standard method for acquiring pattern recognition skills.⁴² There are, however, a few ways to accelerate the transition from slow to fast recognition, high to low attentional load, and serial to parallel processing. Perceptual learning focuses on the commonalities and variations in perceived features and have the goal of making pattern recognition faster, more generalizable to novel images, and operate at the level of implicit (i.e., subconscious) pattern recognition. In general, pattern recognition can be facilitated by asking students to repeatedly view contrasting features with immediate feedback,⁴³ studying structurally similar cases that differ only in superficial features,⁴⁴ and using adaptive learning algorithms that scaffold learning based on ongoing accuracy and response latencies.⁴⁵ The latter is particularly promising in pathology training, with students showing significant improvement in recognizing histopathological features following an adaptive perceptual learning program.⁴⁵

Once pathologists learn to accurately recognize histopathological features, they must also effectively *communicate* their findings to others to facilitate diagnosis, prognosis, and treatment. This is especially challenging in diagnostic medicine when highly specialized nomenclature can vary between individuals, institutions, and regions. Developing expertise

in specialized language is challenging, and various attempts have been made to model and optimize the syntax, semantics, and pragmatics to facilitate communication in specialized domains.⁴⁶ When learning a new language, it is widely accepted that immersing oneself in the language is the fastest way to achieve fluency.⁴⁷ The more verbal a pathologist can be when looking at a case with a trainee, describing salient features out loud (what cognitive psychologists call a *think-aloud* method), repeating terminology, and verbalizing their thought process with organ specific terms, the more likely the trainee will benefit and acquire fluency.^{48–50} In contrast, viewing slides with a trainee silently or with minimal verbalization of thought processes and merely stating the correct diagnosis after silent deliberation is less likely to support organ-specific language development in the trainee. Reading pathology books to reinforce the terminology heard at sign out could further support fluency.

One relevant aspect of *diagnostic decision-making* is what cognitive scientists refer to as category learning, which involves learning key aspects of members of a category, and building mental representations of how categories overlap and diverge.⁵¹ In pathology, category learning helps physicians differentiate between fuzzy diagnostic boundaries, mapping patterns of recognized features (along with clinical context) onto a candidate diagnosis. There are several approaches for facilitating category learning that could be applied to pathologist training. First, *fading* approaches initially expose learners to exaggerated distinctions between across-category exemplars (e.g., clearly benign with non-proliferative features, versus invasive); over time, the difference is systematically reduced until the categories are very challenging to distinguish (even for experts).⁵² Second, *feature highlighting* is a strategy that involves marking and describing the diagnostic features critical for recognizing category membership; this approach, particularly when combined with causal explanations, can be very valuable for learning within- and across-category features.⁵³ Finally, research demonstrates that *interleaved practice* (rather than blocked practice) is valuable for facilitating category learning; this type of practice involves alternating between the study of category exemplars (e.g., benign, invasive, benign, invasive) rather than blocking their study (e.g., benign, benign, invasive, invasive).⁵⁴ The daily case load and sign out of a pathologist reflects the interleaved practice, while study sets of malignant breast pathology cases for example are more reflective of blocked practice. These three learning sciences approaches may prove valuable for accelerating and expanding diagnostic competency in pathology postgraduate training.

Relation to Prior Work

Failure to detect the critical region was considerably lower in this study than rates previously reported in radiology and pathology (30–35%).^{23,55} The hematoxylin and eosin (H&E) staining used on the present biopsy images tend to make some histopathological features visually salient, attracting attention.^{34,56,57} Also, unlike prior studies, we did not restrict interpretation times or the ability to zoom and pan the image. The ability to zoom and pan is critical to the clinical relevance of our study, but we acknowledge that this design could have increased the likelihood that critical regions were eventually detected.

Failure to recognize the relevance of critical features was higher in this study than in prior work,²² but was not a critical determinant of correctly describing features as successful feature descriptions still occurred 64% of the time whether or not there was an overlap of pROI and cROI. Our design examined the accuracy of feature description wherever the participants' ROI was drawn, allowing us to examine the possibility that an accurate diagnosis could occur without focusing attention on our pre-determined expert consensus defined ROIs (i.e., leaving open the possibility that valuable diagnostic information could be found elsewhere on a slide).

Limitations

While the present study allowed relatively naturalistic image zooming and panning by participants, the interpretations were still done within an experimental context. Viewing, annotation, and diagnostic behavior may differ from routine clinical practice, with errors arising at different phases of what is very likely an iterative interpretive process. Furthermore, while we used a small, standardized set of 14 cases to maximize efficiency and minimally interfere with pathologists' busy schedules, a larger and more diverse set of cases is preferable for promoting generalizability. Replication in true clinical settings with more diverse cases is important.

Our histology form required participants to select a single most advanced diagnostic category, which facilitated statistical analyses but also masks some of the nuance and variety of disease discovered during review. In contrast, our open-ended text responses allowed for more nuanced descriptions but also introduced subjectivity and variability, particularly when interpreted by others (including our raters). While combining these two techniques was an important feature of our design, it also highlighted the inherent challenge faced by pathologists (especially post-graduate trainees) when describing and effectively communicating perceived histopathological features. In clinical contexts, pathologists must balance the objectivity and prescriptive value of a categorical diagnosis while also ensuring they are communicating the nuances of disease processes. While innovative, we appreciate the inherent subjectivity of scoring written annotations. The two-phase scoring procedure with independent expert raters increased the reliability of this process, as evidenced by the 94% inter-rater agreement. It is possible, however, that our experts developed specific terminology and associations for describing histopathological features that may not overlap entirely with intended meaning of participants' descriptions, particularly descriptions written by those with less experience. While the raters were intentionally flexible in their interpretation of annotations and discussed discrepancies at length, when non-universally accepted descriptors were used, judging their intended meaning became more difficult.

Finally, while an established gold standard diagnosis of a case is necessary for research purposes, a perfect gold standard for pathology diagnosis is unlikely to exist.¹⁴ The consensus reference diagnoses and consensus cROIs used in the present study were established through years of expert pathologist meetings, including three experts independently interpreting all cases followed by consensus meetings to come to agreement on a reference diagnosis for each case (using a modified Delphi approach).^{16,17} To our knowledge, this process was much more comprehensive than what is typically done to reach

consensus (e.g., via second opinions) in clinical practice. Our novel conceptual framework for evaluating diagnostic errors in pathology found confirmatory evidence that detecting and recognizing the relevance of critical regions are important first steps in the interpretive process. In our study, however, the only variable associated with a successful diagnosis was the ability to accurately describe the histologic features pathologists deemed most important on a slide. Investing educational efforts into developing organ specific vocabulary and fluency in feature recognition and description may prove an ideal intervention point for continuing pathologist education and training, with potential implications for improving diagnostic accuracy.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Ventana Medical Systems, Inc., a member of the Roche Group, for use of iScan Coreo Au™ whole slide imaging system, and HD View SL for the source code used to build our digital viewer. For a full description of HD View SL please see <http://hdviews.codeplex.com/>. We also wish to thank the staff, faculty, and trainees at the various pathology training programs across the U.S. for their participation and assistance in this study.

Funding

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under award number R01 CA225585, R01 CA172343, R01 CA140560. The content is solely the responsibility of the authors and does not necessarily represent the views of the National Cancer Institute or the National Institutes of Health. Researchers are independent from funders, and the funding agency played no role in study design, conduct, analysis or interpretation.

Data Availability Statement

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. Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine, The National Academies of Sciences, Engineering, and Medicine. *Improving Diagnosis in Health Care*. (Balogh EP, Miller BT, Ball JR, eds.). National Academies Press (US); 2015. Accessed July 28, 2022. <http://www.ncbi.nlm.nih.gov/books/NBK338596/>
2. Brunyé T, Drew T, Weaver DL, Elmore JG. A review of eye tracking for understanding and improving diagnostic interpretation. *Cogn Res Princ Implic*. 2019;4:1–16. [PubMed: 30693393]
3. Kundel HL, Nodine CF. Studies of eye movements and visual search in radiology. In: Senders JW, Fisher DF, Monty RA, eds. *Eye Movements and the Higher Psychological Processes*. Lawrence Erlbaum Associates; 1978:317–327.
4. Elstein AS, Schwarz A. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *BMJ*. 2002;324(7339):729–732. doi:10.1136/bmj.324.7339.729 [PubMed: 11909793]

5. Krupinski EA, Tillack AA, Richter L, et al. Eye-movement study and human performance using telepathology virtual slides. Implications for medical education and differences with experience. *Hum Pathol.* 2006;37(12):1543–1556. [PubMed: 17129792]
6. Al-Moteri MO, Symmons M, Plummer V, Cooper S. Eye tracking to investigate cue processing in medical decision-making: A scoping review. *Comput Hum Behav.* 2017;66:52–66. doi:10.1016/j.chb.2016.09.022
7. Logothetis NK, Sheinberg DL. Visual object recognition. *Annu Rev Neurosci.* 1996;19:577–621. doi:10.1146/annurev.ne.19.030196.003045 [PubMed: 8833455]
8. Gauthier I, Tarr MJ. Visual Object Recognition: Do We (Finally) Know More Now Than We Did? *Annu Rev Vis Sci.* 2016;2(1):377–396. doi:10.1146/annurev-vision-111815-114621 [PubMed: 28532357]
9. Sox HC, Blatt MA, Higgins MC, Marton KI. *Medical Decision Making.* Butterworths; 1988. doi:10.1002/9781118341544
10. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis. *Science.* 1959;130:9–21. doi:10.1126/science.130.3366.9 [PubMed: 13668531]
11. Kassirer JP, Kopelman RI, Wong JB. *Learning Clinical Reasoning.* Williams & Wilkins; 1991:332.
12. Hébert TM, Cole A, Panarelli N, et al. Training the Next Generation of Pathologists: A Novel Residency Program Curriculum at Montefiore Medical Center/Albert Einstein College of Medicine. *Acad Pathol.* 2019;6:2374289519848099. doi:10.1177/2374289519848099
13. Black-Schaffer WS, Morrow JS, Prystowsky MB, Steinberg JJ. Training Pathology Residents to Practice 21st Century Medicine: A Proposal. *Acad Pathol.* 2016;3:2374289516665393. doi:10.1177/2374289516665393
14. Elmore JG, Longton GM, Carney PA, et al. Diagnostic Concordance Among Pathologists Interpreting Breast Biopsy Specimens. *JAMA.* 2015;313(11):1122. doi:10.1001/jama.2015.1405 [PubMed: 25781441]
15. Elmore JG, Nelson HD, Pepe MS, et al. Variability in Pathologists' Interpretations of Individual Breast Biopsy Slides: A Population Perspective. *Ann Intern Med.* 2016;164(10):649–655. doi:10.7326/M15-0964 [PubMed: 26999810]
16. Oster N, Carney PA, Allison KH, et al. Development of a diagnostic test set to assess agreement in breast pathology: Practical application of the Guidelines for Reporting Reliability and Agreement Studies (GRRAS). *BMS Womens Health.* 2013;13(1):3.
17. Allison KH, Reisch LM, Carney PA, et al. Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel. *Histopathology.* 2014;65(2):240–251. doi:10.1111/his.12387 [PubMed: 24511905]
18. Ventana Medical Systems I. iScan Coreo Au Product Page.
19. Onega T, Weaver DL, Frederick PD, et al. The diagnostic challenge of low-grade ductal carcinoma in situ. *Eur J Cancer.* 2017;80:39–47. doi:10.1016/j.ejca.2017.04.013 [PubMed: 28535496]
20. Hooge ITC, Holleman GA, Haukes NC, Hessels RS. Gaze tracking accuracy in humans: One eye is sometimes better than two. *Behav Res Methods.* 2019;51:2712–2721. doi:10.3758/s13428-018-1135-3 [PubMed: 30350022]
21. Bogart S. SankeyMATIC: A Sankey diagram builder for everyone. Published online 2022. <https://github.com/nowthis/sankeymatic>
22. Nagarkar DB, Mercan E, Weaver DL, et al. Region of interest identification and diagnostic agreement in breast pathology. *Mod Pathol.* 2016;29(9):1004–1004. doi:10.1038/modpathol.2016.85 [PubMed: 27198567]
23. Kundel HL, Nodine CF. Studies of eye movements and visual search in radiology. In: Senders JW, Fisher DF, Monty RA, eds. *Eye Movements and the Higher Psychological Processes.* Lawrence Erlbaum Associates; 1978:317–327.
24. Mello-Thoms C, Hardesty L, Sumkin J, et al. Effects of lesion conspicuity on visual search in mammogram reading. *Acad Radiol.* 2005;12(7):830–840. doi:10.1016/j.acra.2005.03.068 [PubMed: 16039537]
25. Kundel HL, Nodine CF, Carmody DP. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Invest Radiol.* 1978;13(3):175–181. doi:10.1097/00004424-197805000-00001 [PubMed: 711391]

26. Carmody DP, Nodine CF, Kundel HL. An analysis of perceptual and cognitive factors in radiographic interpretation. *Perception*. 1980;9:339–344. doi:10.1068/p090339 [PubMed: 7454514]
27. Manning D, Ethell S, Donovan T, Crawford T. How do radiologists do it? The influence of experience and training on searching for chest nodules. *Radiography*. 2006;12(2):134–142. doi:10.1016/j.radi.2005.02.003
28. Brunyé TT, Mercan E, Weaver DLDL, Elmore JGJG. Accuracy is in the eyes of the pathologist: The visual interpretive process and diagnostic accuracy with digital whole slide images. *J Biomed Inform*. 2017;66:171–179. doi:10.1016/j.jbi.2017.01.004 [PubMed: 28087402]
29. Heekeren HR, Marrett S, Ungerleider LG. The neural systems that mediate human perceptual decision making. *Nat Rev Neurosci*. 2008;9(6):467–479. doi:10.1038/nrn2374 [PubMed: 18464792]
30. Brunyé TT, Gardony AL. Eye tracking measures of uncertainty during perceptual decision making. *Int J Psychophysiol*. 2017;120:60–68. doi:10.1016/j.ijpsycho.2017.07.008 [PubMed: 28732659]
31. Callan DJ. Eye Movement Relationships to Excessive Performance Error in Aviation. *Proc Hum Factors Ergon Soc Annu Meet*. 1998;42(15):1132–1136. doi:10.1177/154193129804201516
32. Anderson BA, Laurent PA, Yantis S. Learned Value Magnifies Salience-Based Attentional Capture. *PLOS ONE*. 2011;6(11):e27926. doi:10.1371/journal.pone.0027926 [PubMed: 22132170]
33. Gaspar JM, McDonald JJ. Suppression of Salient Objects Prevents Distraction in Visual Search. *J Neurosci*. 2014;34(16):5658–5666. doi:10.1523/JNEUROSCI.4161-13.2014 [PubMed: 24741056]
34. Brunyé TT, Carney PA, Allison KH, Shapiro LG, Weaver DL, Elmore JG. Eye movements as an index of pathologist visual expertise: A pilot study. *PLoS ONE*. 2014;9(8). doi:10.1371/journal.pone.0103447
35. Sauter M, Liesefeld HR, Zehetleitner M, Müller HJ. Region-based shielding of visual search from salient distractors: Target detection is impaired with same- but not different-dimension distractors. *Atten Percept Psychophys*. 2018;80(3):622–642. doi:10.3758/s13414-017-1477-4 [PubMed: 29299850]
36. Gegenfurtner A, Lehtinen E, Jarodzka H, Säljö R. Effects of eye movement modeling examples on adaptive expertise in medical image diagnosis. *Comput Educ*. 2017;113:212–225. doi:10.1016/j.compedu.2017.06.001
37. Gegenfurtner A, Lehtinen E, Säljö R. Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educ Psychol Rev*. 2011;23(4):523–552. doi:10.1007/s10648-011-9174-7
38. Kramer MR, Porfido CL, Mitroff SR. Evaluation of strategies to train visual search performance in professional populations. *Curr Opin Psychol*. 2019;29:113–118. doi:10.1016/j.copsy.2019.01.001 [PubMed: 30731261]
39. Abernethy B. Expertise, Visual Search, and Information Pick-up in Squash. *Perception*. 1990;19(1):63–77. doi:10.1068/p190063 [PubMed: 2336337]
40. Goulet C, Bard C, Fleury M. Expertise Differences in Preparing to Return a Tennis Serve: A Visual Information Processing Approach. *J Sport Exerc Psychol*. 1989;11(4):382–398. doi:10.1123/jsep.11.4.382
41. Gauthier I, Williams P, Tarr MJ, Tanaka J. Training ‘greeble’ experts: a framework for studying expert object recognition processes. *Vision Res*. 1998;38(15):2401–2428. doi:10.1016/S0042-6989(97)00442-2 [PubMed: 9798007]
42. Kellman PJ, Massey CM, Son JY. Perceptual Learning Modules in Mathematics: Enhancing Students’ Pattern Recognition, Structure Extraction, and Fluency. *Top Cogn Sci*. 2010;2(2):285–305. doi:10.1111/j.1756-8765.2009.01053.x [PubMed: 25163790]
43. Schwartz DL, Bransford JD. A Time for Telling. *Cogn Instr*. 1998;16(4):475–522.
44. Gick ML, Holyoak KJ. Schema induction and analogical transfer. *Cognit Psychol*. 1983;15(1):1–38. doi:10.1016/0010-0285(83)90002-6
45. Krasne S, Hillman JD, Kellman PJ, Drake TA. Applying perceptual and adaptive learning techniques for teaching introductory histopathology. *J Pathol Inform*. 2013;4(1):34. doi:10.4103/2153-3539.123991 [PubMed: 24524000]

46. Faber P. A Cognitive Linguistics View of Terminology and Specialized Language. De Gruyter Mouton; 2012. doi:10.1515/9783110277203
47. Genesee F. Integrating Language and Content: Lessons from Immersion. Center for Applied Linguistics, National Center for Research on Cultural Diversity and Second Language Learning; 1994.
48. Vitak SA, Ingram JE, Duchowski AT, Ellis S, Gramopadhye AK. Gaze-augmented think-aloud as an aid to learning. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '12. Association for Computing Machinery; 2012:2991–3000. doi:10.1145/2207676.2208710
49. Pinnock R, Young L, Spence F, Henning M. Can think aloud be used to teach and assess clinical reasoning in graduate medical education? *J Grad Med Educ.* 2015;7:334–337. doi:10.4300/JGME-D-14-00601.1 [PubMed: 26457135]
50. Jagannath AD, Dreicer JJ, Penner JC, Dhaliwal G. The cognitive apprenticeship: advancing reasoning education by thinking aloud. *Diagnosis.* Published online December 1, 2022. doi:10.1515/dx-2022-0043
51. Markman AB, Ross BH. Category use and category learning. *Psychol Bull.* 2003;129:592–613. doi:10.1037/0033-2909.129.4.592 [PubMed: 12848222]
52. Pashler H, Mozer MC. When does fading enhance perceptual category learning? *J Exp Psychol Learn Mem Cogn.* 2013;39:1162–1173. doi:10.1037/a0031679 [PubMed: 23421513]
53. Meagher BJ, McDaniel MA, Nosofsky RM. Effects of feature highlighting and causal explanations on category learning in a natural-science domain. *J Exp Psychol Appl.* 2022;28:283–313. doi:10.1037/xap0000369 [PubMed: 34110857]
54. Eglinton LG, Kang SHK. Interleaved Presentation Benefits Science Category Learning. *J Appl Res Mem Cogn.* 2017;6(4):475–485. doi:10.1016/j.jarmac.2017.07.005
55. Brunyé TT, Drew T, Kerr KF, Shucard H, Weaver DL, Elmore JG. Eye tracking reveals expertise-related differences in the time-course of medical image inspection and diagnosis. *J Med Imaging.* 2020;7(5):051203. doi:10.1117/1.JMI.7.5.051203
56. Parkhurst D, Law K, Niebur E. Modeling the role of salience in the allocation of overt visual attention. *Vision Res.* 2002;42(1):107–123. doi:10.1016/S0042-6989(01)00250-4 [PubMed: 11804636]
57. Tatler BW, Hayhoe MM, Land MF, Ballard DH. Eye guidance in natural vision: reinterpreting salience. *J Vis.* 2010;11(5):5–5. doi:10.1167/11.5.5

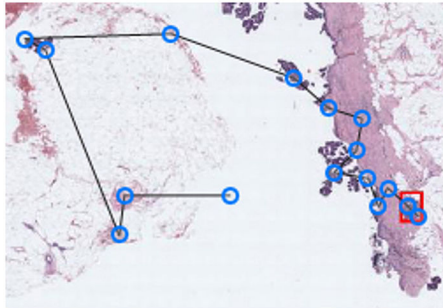
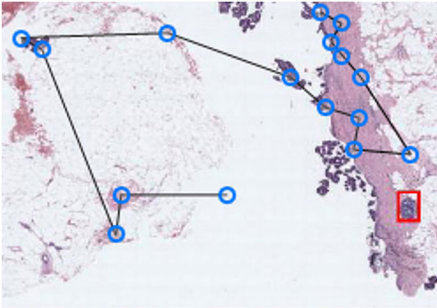
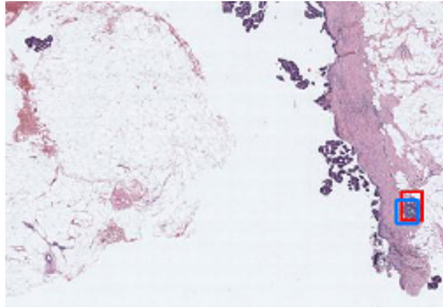
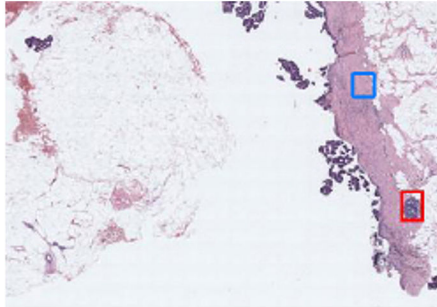
Interpretive Phase	Example of Success	Example of Error
Detecting Critical Region	 <p><i>Eyes fixate the cROI.</i></p>	 <p><i>Eyes do not fixate the cROI.</i></p>
Recognizing Relevance	 <p><i>Participant-drawn pROI overlaps the consensus reference cROI.</i></p>	 <p><i>Participant-drawn pROI does not overlap the consensus reference cROI.</i></p>
Describing Features	<p>Papillary lesion and adjacent duct with atypically proliferating cells.</p> <p><i>Participant's annotation accurately describes pROI features.</i></p>	<p>Atypical cells invading through a fibrotic desmoplastic stroma with no ductal architecture.</p> <p><i>Participant's annotation inaccurately describes pROI features.</i></p>
Decision Making	<ul style="list-style-type: none"> <input type="checkbox"/> Non-Proliferative changes <input type="checkbox"/> Proliferative lesion without atypia <input checked="" type="checkbox"/> Atypical lesion <input type="checkbox"/> Carcinoma in situ <input type="checkbox"/> Invasive carcinoma (ductal, lobular or other special type) <p><i>Diagnosis matches consensus.</i></p>	<ul style="list-style-type: none"> <input type="checkbox"/> Non-Proliferative changes <input type="checkbox"/> Proliferative lesion without atypia <input type="checkbox"/> Atypical lesion <input type="checkbox"/> Carcinoma in situ <input checked="" type="checkbox"/> Invasive carcinoma (ductal, lobular or other special type) <p><i>Diagnosis mismatches consensus.</i></p>

Figure 1. The four interpretive phases in our framework, along with examples of success and error in each phase. Example case depicts consensus atypia. Top panel images depict eye fixation sequences (blue rings, black lines) and cROI (red rectangle). Second panel image depicts overlapping cROI (red) and pROI (blue). Third panel depicts participant-provided feature annotation. Bottom panel depicts participant-decided categorical diagnosis.



Figure 2. A pathology trainee (face blurred for privacy) interpreting a WSI during the study; the eye tracker is attached to the bottom of the computer monitor, and the trainee is navigating the zoomed case using the computer mouse.

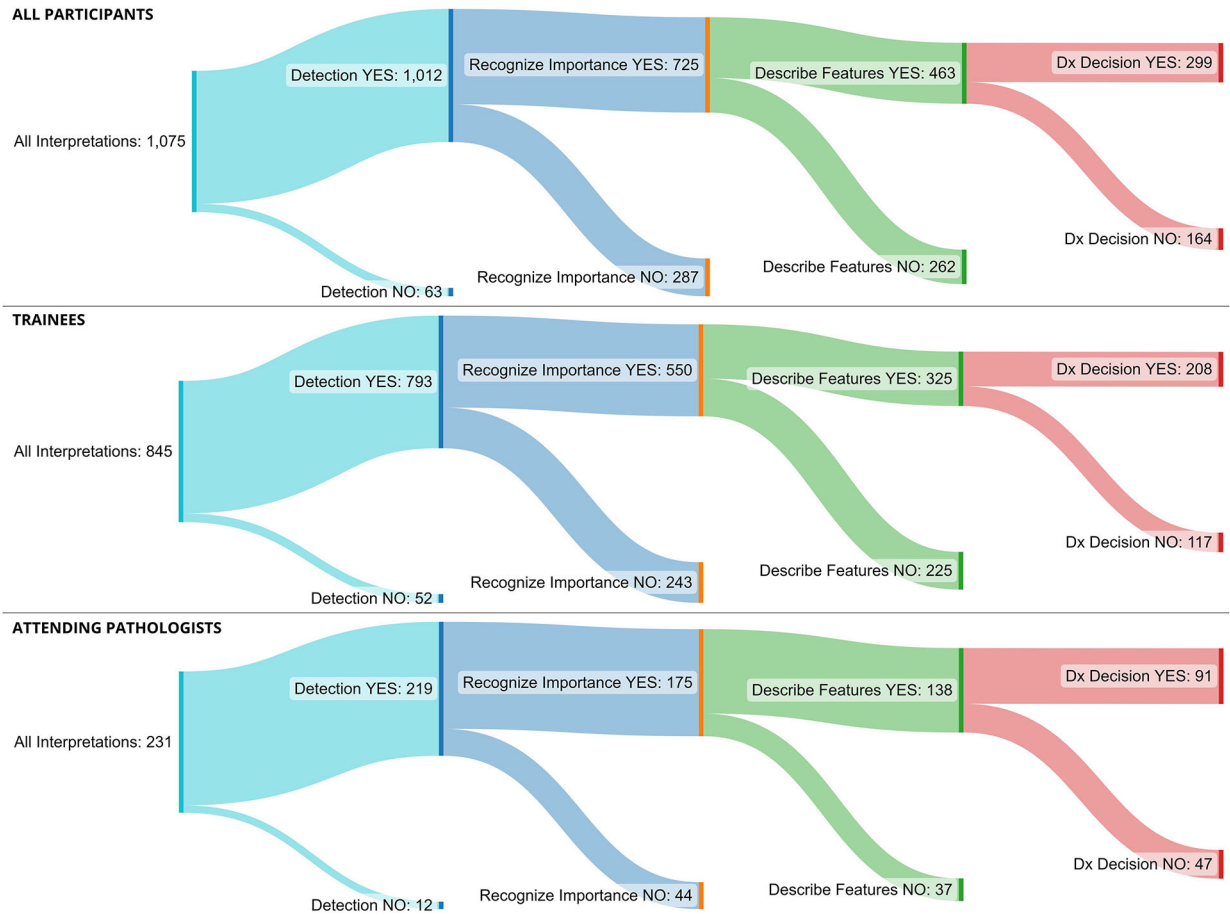


Figure 3. Sankey diagrams showing the flow of the interpretive process through the four interpretive phases: detecting the lesion, recognizing importance, describing features, and diagnostic decision. The width of a flow is proportional to its quantity (i.e., the number of interpretations), branches represent change in flow (e.g., detecting or not detecting a cROI), and colors represent phases (e.g., detection, recognition). Upper image displays all participants, middle displays pathology trainees, and lower displays attending pathologists.

Table 1.

Method used to assess accuracy of participant's description of histopathologic features

Rater 1 response	Rater 2 response					
	Stated feature not found	Accurately represents features	Stated features underrepresent actual features	Stated features overrepresent actual features	Stated features are contradictory	Ambiguous, difficult to judge
Stated feature not found	25 ^{a,b}	2 ^c	1 ^a	1 ^a	0 ^a	0 ^c
Accurately represents features	0 ^c	685 ^{b,d}	13 ^c	23 ^c	0 ^c	1 ^c
Stated features underrepresent actual features	0 ^a	8 ^c	172 ^{a,b}	0 ^a	0 ^a	0 ^c
Stated features overrepresent actual features	0 ^a	2 ^c	0 ^a	188 ^{a,b}	0 ^a	1 ^c
Stated features are contradictory	0 ^a	1 ^c	0 ^a	1 ^a	2 ^{a,b}	0 ^c
Ambiguous, difficult to judge	1 ^c	12 ^c	3 ^c	0 ^c	0 ^c	69 ^{b,c}

Agreement/disagreement between expert rater 1 and rater 2 after round 2 of evaluating study participants' feature annotations of their individually drawn pROI. The frequency of agreement between the 2 raters is along the (upper left to lower right) diagonal. The 2 raters' assessments agreed for 1141 of 1211 annotations (94.2%).

^a Inaccurate

^b Indicates agreement

^c Excluded from analyses due to disagreement between the 2 raters or ambiguous descriptions from participants

^d Accurate

Table 2.

Number (and proportion) of successful interpretations within each of the 4 phases of the interpretive process, as a function of consensus diagnosis and participant experience level

Interpretive phase	Expert consensus diagnostic category of the case	Trainees	Attending pathologists	All participants
Detecting critical region	Benign without atypia	NA (94 [80%])	NA (28 [88%])	NA (122 [81%])
	Atypia	232 (98%)	62 (97%)	294 (98%)
	Low-grade DCIS	229 (94%)	63 (97%)	292 (94%)
	High-grade DCIS	110 (93%)	30 (88%)	140 (92%)
	Invasive carcinoma	128 (100%)	36 (100%)	164 (100%)
Recognizing relevance	Benign without atypia	NA (36 [31%])	NA (12 [38%])	NA (48 [32%])
	Atypia	164 (69%)	53 (83%)	217 (72%)
	Low-grade DCIS	178 (73%)	57 (88%)	235 (76%)
	High-grade DCIS	71 (60%)	25 (74%)	96 (63%)
	Invasive carcinoma	103 (80%)	28 (78%)	131 (80%)
Describing features	Benign without atypia	NA (90 [76%])	NA (27 [84%])	NA (117 [78%])
	Atypia	121 (51%)	48 (75%)	169 (56%)
	Low-grade DCIS	143 (59%)	49 (75%)	192 (62%)
	High-grade DCIS	30 (25%)	24 (71%)	54 (36%)
	Invasive carcinoma	118 (92%)	35 (97%)	153 (93%)
Diagnostic decisions	Benign without atypia	NA (104 [88%])	NA (29 [91%])	NA (133 [89%])
	Atypia	52 (22%)	32 (50%)	84 (28%)
	Low-grade DCIS	57 (22%)	18 (28%)	75 (24%)
	High-grade DCIS	7 (6%)	11 (32%)	18 (12%)
	Invasive carcinoma	122 (95%)	36 (100%)	158 (96%)

DCIS, ductal carcinoma in situ; NA, not applicable.

Table 3.

Number (and proportion) of interpretations that proceeded through every possible combination of detection (yes, no), recognizing importance (yes, no), recognizing features (yes, no), and achieving an accurate diagnosis (yes, no)

Row no.	Detecting region	Recognizing relevance	Describing features	Diagnosing accurately	Frequency (proportion)
1	YES	YES	YES	YES	299 (0.28)
2	YES	NO	YES		102 (0.09)
3	YES	YES	NO		28 (0.03)
4	NO	NO	YES		17 (0.02)
5	YES	NO	NO		15 (0.01)
6	NO	NO	NO		7 (0.01)
7	NO	YES	NO		0 (0)
8	NO	YES	YES		0 (0)
9	YES	YES	NO	NO	234 (0.22)
10	YES	YES	YES		164 (0.15)
11	YES	NO	NO		93 (0.09)
12	YES	NO	YES		77 (0.07)
13	NO	NO	YES		26 (0.02)
14	NO	NO	NO		11 (0.01)
15	NO	YES	NO		2 (0)
16	NO	YES	YES		0 (0)

Combinations ending with an accurate diagnosis are presented first (rows 1–8), ordered from most to least frequent. These are followed by combinations ending with an inaccurate diagnosis (rows 9–16), ordered from most to least frequent.