Perceptual Analysis of the Reading of Dermatopathology Virtual Slides by Pathology Residents

Claudia Mello-Thoms, MS, PhD; Carlos A. B. Mello, PhD; Olga Medvedeva, MS; Melissa Castine, BS; Elizabeth Legowski, BS; Gregory Gardner, MS; Eugene Tseytlin, MS; Rebecca Crowley, MD, MSIS

• ***Context.***—**The process by which pathologists arrive at a given diagnosis—a combination of their slide exploration strategy, perceptual information gathering, and cognitive decision making—has not been thoroughly explored, and many questions remain unanswered.**

Objective.—**To determine how pathology residents learn to diagnose inflammatory skin dermatoses, we contrasted the slide exploration strategy, perceptual capture of relevant histopathologic findings, and cognitive integration of identified features between 2 groups of residents, those who had and those who had not undergone their dermatopathology rotation.**

Design.—**Residents read a case set of 20 virtual slides (10 depicting nodular and diffuse dermatitis and 10 depicting subepidermal vesicular dermatitis), using an in-house–developed interface. We recorded residents' reports of diagnostic findings, conjectured diagnostic hypotheses, and final (or differential) diagnosis for each case, and time stamped each interaction with the interface. We created search maps of residents' slide exploration strategy.**

Results.—**No statistically significant differences were observed between the resident groups in the number of correctly or incorrectly reported diagnostic findings, but residents with dermatopathology training generated significantly more correct hypotheses (mean improvement of 88.5%) and correct diagnoses (70% of all correct diagnoses).**

Conclusions.—**Two types of slide exploration strategy were identified for both groups: (1) a focused and efficient search, observed when the final diagnosis was correct; and (2) a more dispersed, time-consuming strategy, observed when the final diagnosis was incorrect. This difference was statistically significant, and it suggests that initial interpretation of a slide may bias further slide exploration.**

(*Arch Pathol Lab Med.* 2012;136:551–562; doi: 10.5858/arpa.2010-0697-OA)

Pathology is the gold standard in medical diagnosis, and the pathologist's decision has the power to uniquely determine the outcome of a given patient's treatment. However, interpretation of either glass or virtual slides requires enormous amounts of perceptual and cognitive skills, which are gathered during the course of years of apprenticeship and clinical practice. Not surprisingly, residency in pathology takes an average of 5 years to complete, with additional training, in the form of fellowship, required for subspecialty certification. Perhaps because of the long training or the better opportunities in the job market, a significant fraction of pathologists receive board certification as general practitioners, and are thus entitled to interpret a large number of different tissue types.

As a consequence, significant disagreement has been observed when subspecialty-trained pathologists are compared with general practitioners. For example, studies comparing the diagnoses rendered by non-subspeciality-trained pathologists with those rendered by dermatopathologists in skin biopsies have shown a disagreement rate ranging from at least 2.7%[1] to 7.4%[2] to 14.3%.[3] On average, 21% of the biopsies disagreed upon could have had some clinical impact,[4] with a false-negative rate for cancer of 2%.[1]

Behind such disagreement and error statistics is the fact that there is variability in the reading of pathology slides, both among different pathologists and for the same pathologist at different points in time. This variability usually depends on the specific type of pathology examined, the grading system applied, and the previous knowledge and experience of the pathologist. Often variability is assessed using the Cohen κ statistic, which corrects for chance agreement. Hence, a κ of 1 indicates perfect agreement, whereas a value of 0 indicates pure chance agreement. In dermatopathology, agreement has been shown to range from 0.34 (poor) to 0.5 (moderate)[5,6] when different pathologists diagnose the same slides.

However, most studies of disagreement and error in the reading of pathology slides arrive at their conclusions based solely on the pathologists' final diagnosis of the case, without consideration for the process by which the decision was made. This process—an integration of the slide exploration strategy, perceptual information gathering,

Accepted for publication July 26, 2011.

From the Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania (Drs Mello-Thoms and Crowley, Mss Medvedeva, Castine, and Legowski, and Messrs Gardner and Tseytlin); and the Centro de Informatica, Universidade Federal de Pernambuco, Recife, Brazil (Dr Mello).

The authors have no relevant financial interest in the products or companies described in this article.

Presented in part at Pathology Informatics 2010; September 20, 2010; Boston, Massachusetts.

Reprints: Claudia Mello-Thoms, PhD, MS, Department of Biomedical Informatics, UPMC Shadyside Hospital, Cancer Pavilion, 5150 Centre Ave, Ste 301, Room 304, Pittsburgh, PA 15232 (e-mail: mellothomsc@upmc.edu).

and cognitive decision making—has been thoroughly studied in other medical domains (such as radiology[7–19]), but it has only recently begun to receive attention in pathology.[20–24] For example, one of the better understood parts of this process is the multiscale approach used by pathologists in their cognitive decision making.[25] According to this model, pathologists identify suspicious regions and form initial hypotheses about the case at low magnification, but they need the higher resolving power of high magnification to make a final determination on the nature of the disease process present. Nonetheless, a full understanding of the basic principles behind the entire process, not only its cognitive component, is crucial for understanding acquisition and characterization of expertise, and for devising ways to reduce the gap between the subspecialty-trained expert and the general pathologist.

In the cognitive sciences there are many different theories that aim to explain how people acquire expertise in a given domain. We will focus on 2 of these theories, which are arguably the most commonly used. The first is based upon the Dreyfus model,[26] and it is the basis underneath the ''forward reasoning'' strategy, which for decades was prevalent in most medical schools in the United States. According to this theory, careful consideration and identification of all histopathologic features present in an image should precede the formation of diagnostic hypotheses and hence the generation of a diagnosis on the case. This supposedly minimizes the risk that a novice may lock in on an incorrect diagnosis before having considered all possible differential diagnoses that are supported by the histopathologic findings in the case. One possible problem with this strategy is that identification of too many irrelevant features may lead to the assembly of several competing diagnoses, which may worsen the performance of residents.[27] In the Dreyfus model,[26] as expertise develops, the pathologist needs to rely less and less on the identification of histopathologic features, as the slide starts to be recognized holistically (ie, in a global fashion). Thus, in this model, the acquisition of expertise is marked from a move from pure cognitive processing by novices to pure perceptual processing by experts.[28] In summary, if we apply this model to pathology residents before and after they receive subspecialty training, the model predicts that before the residents receive subspecialty training (for example, in dermatopathology), they should be able to detect and identify a large number of histopathologic features and generate several competing diagnoses, but ultimately fail in the selection of the final diagnosis. On the other hand, after dermatopathology training, as their expertise level in the domain is increased, they should still be able to detect and identify a significant number of histopathologic findings, but now should generate fewer competing diagnoses and struggle less to arrive at the correct diagnosis.

In contrast with this model, Lesgold et al[29] proposed that novices primarily carry out almost pure perceptual processing, and as they learn and acquire expertise they advance to a mix of perceptual and cognitive processing. Lesgold et al[29] asserted that in the acquisition of expertise in a visually based domain, perceptual learning of abnormal feature characteristics precedes cognitive-inferential decision-making processes, which are related to disease diagnosis. Hence, according to this model, novices should be able to identify histopathologic findings by their visual characteristics, even if they do not know what those findings are or whether they are relevant diagnostically. Experts, on the other hand, possess a large internal database of findings and diagnoses, use at most 4 cues to diagnose a slide,[30] and make decisions using both perceptual and cognitive processing by integrating these cues almost instantly into the overall context of the image. In summary, this model predicts that before their dermatopathology rotation residents should be able to detect many histopathologic findings in the slide by using perceptual processing alone, but they should mostly fail to properly identify such findings and to generate correct diagnoses for the cases. After their dermatopathology rotation, residents should be able to detect and identify more findings. Furthermore, as cognitive processing is expanding, they should be better able to integrate perceptual and cognitive elements to generate a few correct differential diagnoses.

This study was carried out in order to determine which of these 2 models best explains the acquisition of expertise in dermatopathology during pathology residency. We concentrated on the detection of inflammatory skin lesions for 2 primary reasons: because (1) this lesion type ranks among the 3 highest (along with melanocytic and squamoproliferative lesions) in yielding disagreement and error in the reading of skin biopsies by general pathologists[4]; but (2) for the other 2 lesion types, specific clinical guidelines are in place to refer the cases to specialists (ie, dermatopathologists), but no such guidelines exist for inflammatory dermatoses. This leads the majority of these biopsies to be interpreted by general pathologists,[4] but given the histologically similar appearance of many different inflammatory dermatoses, these diagnoses can present quite a perceptual and cognitive challenge.

Research questions:

Research question 1. How does dermatopathology training influence the residents' ability to detect and identify diagnostically significant findings and to generate appropriate diagnoses on slides depicting inflammatory skin biopsies?

Research question 2. Does the additional experience acquired during the dermatopathology rotation influence the way in which the residents visually explore, pan, and zoom the digital slide?

MATERIALS AND METHODS

This study was approved by our institutional review board (IRB No. PRO09030260).

Subjects

Eleven pathology residents were recruited from 2 large academic training institutions in our area through mass e-mail sent to the residency program advertising the opportunity to participate in the experiment. The inclusion criterion required only that participants not be in their first year of residency. Upon receiving e-mail, residents self-selected to respond to us, and they were included in this experiment on a first come, first served basis. Our sample contained 1 postgraduate year 2, 7 postgraduate year 3, and 3 postgraduate year 4 residents. Of these, 7 residents had gone through a dermatopathology rotation, and 4 had not. All were paid for their participation.

Study Design

The study consisted of working through 2 sets of 10 cases (one set containing subepidermal vesicular [SV] dermatoses, the other nodular and diffuse [ND] dermatoses), identifying diagnostic

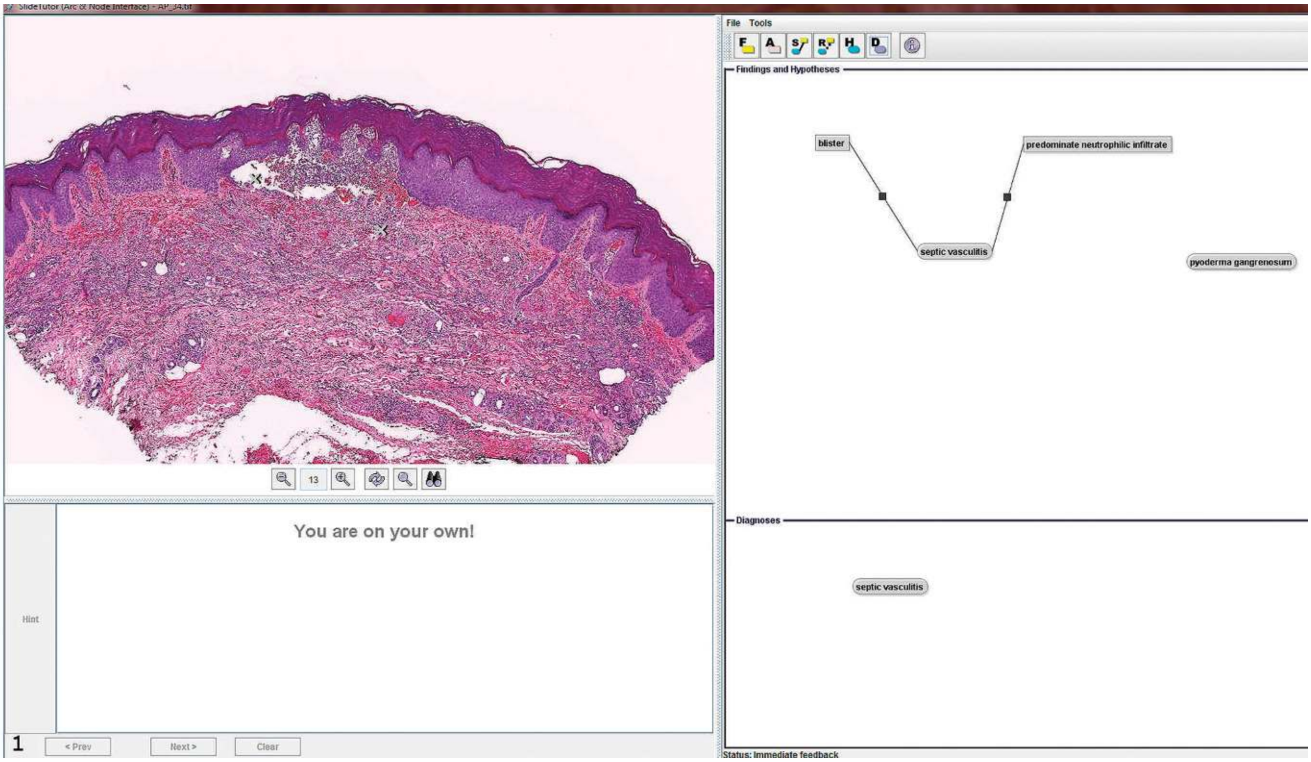


Figure 1. *Screen shot of an ongoing case reading, depicting findings and their associated hypotheses as identified by a given resident.*

findings, and providing a final diagnosis for each case. Subjects were instructed to think aloud while working through the cases in order to provide additional information about feature identification, but think-aloud data were not analyzed in this study. Time on task was allowed to vary so that all participants completed all of the cases.

The experiment started with a 30-minute training session in which the residents became acquainted with the system to be used for data collection. This was accomplished by having the residents watch a 20-minute video demonstration of the system, followed by live practice under a research assistant's guidance as they explored the functionality of the interface. After that, reading of the first case set started. To avoid reading order effects, half of the subjects started with the ND cases, and the other half started with the SV cases. Subjects could take a break of approximately 15 minutes between the 2 case sets. On average, reading of both sets lasted approximately 4 hours.

The residents' data were collected using a ''light'' version of SlideTutor, which is an intelligent tutoring system developed to aid in the teaching of diagnostic reading of pathology slides by Crowley and Medvedeva.[31] In SlideTutor, subjects are provided with a digital slide to explore, by zooming (up to ×20) and panning, and can perform a number of action types, such as identifying findings they see on the slide, reporting initial hypotheses (which are potential diagnoses the user is considering), drawing support for and refuting links between findings and hypotheses (to aid in the reasoning process), and recording the final diagnosis, which can be a single disease or a set of differential diagnoses. Figure 1 illustrates a case in progress.

The light version used here differed from the standard version of SlideTutor in that users were not provided with any feedback while working on the cases; therefore, they were never given any indication as to the correctness of their decisions regarding the features they were identifying, the hypotheses they were generating, or the diagnosis(es) they were reporting for each case.

Unbeknownst to the residents, SlideTutor's internal reasoning interface was still active, and it classified the residents' actions as being correct (for example, a reported feature was indeed present at the indicated location), incorrect (eg, a given diagnosis to a case did not match the truth table), or missing (when a feature listed in the truth table as being diagnostically important was not reported by the residents). This reasoning was not displayed to the residents, but was recorded and used for data analysis.

Case Selection

This study entailed the use of 20 cases, of which 10 had SV dermatitides and 10 had ND dermatitides. Table 1 lists the differential diagnoses and the associated diagnostic findings for each of the SV and ND cases. Although many other features were present in each case, only the ones listed in the table were considered to be pertinent to the correct diagnosis. The findings and diagnoses in Table 1 were established by a knowledge engineer (M.C.) who had been trained by a dermatopathologist to identify and diagnose skin diseases. As such, they were used as the gold standard for this study, and Table 1 (in addition to information regarding the best magnification range to identify each feature) is also referred to as the ''truth table.'' For the magnification ranges, any digital zoom levels corresponding to lower magnifications than ×4 were deemed to be low magnification; digital zoom levels corresponding to magnifications between ×4 and ×10 were deemed to be medium magnification; and finally, digital zoom levels corresponding to magnifications greater than ×10 (up to ×20) were deemed to be high magnification.

Data Analysis

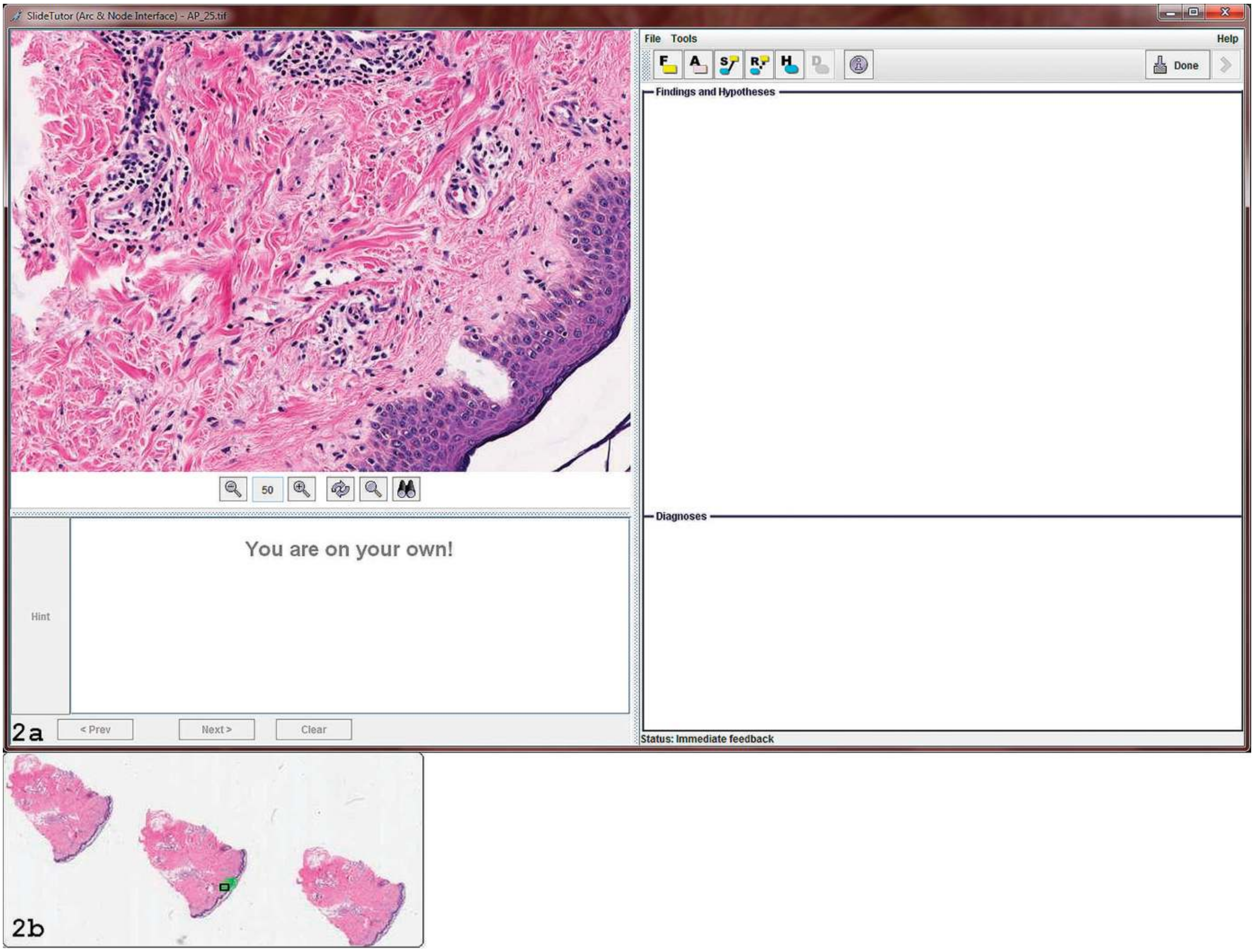
In order to determine which model best explains acquisition of expertise in dermatopathology, we assessed the residents' ability to identify the diagnostic findings present on the cases, and the relationship between perceptual parameters (such as dwell time, time to hit, reporting time, etc) and feature identification, as explained next. In these analyses, all tests were used with a significance level of $P < .05$.

Research Question 1. Perceptual Analysis of Feature Identification

Perceptual parameters indicative of visual search behavior and typically associated with abnormality detection and identification

Table 1. Diagnosis and Associated Findings, per Case, for the Subepidermal Vesicular and the Nodular and Diffuse Sets

Case Type	Case No.	Case Findings	Case Diagnosis/Differential Diagnosis
Nodular and diffuse	01	Diffuse histiocytic inflammatory infiltrate Histiocytic giant cells Lipophages	Xanthogranuloma Generalized eruptive histiocytoma Malacoplakia Silica granuloma Paraffinoma
Nodular and diffuse	02	Sarcoidal granuloma Naked tubercle Nodular histiocytic inflammatory infiltrate No foreign body	Sarcoidosis Crohn disease Melkersson-Rosenthal syndrome
Nodular and diffuse	03	Diffuse lymphocytic inflammatory infiltrate Spongiosis Parakeratosis Top-heavy inflammatory infiltrate Isolated small lymphocytes	Pseudolymphoma
Nodular and diffuse	04	Nodular histiocytic inflammatory infiltrate Palisaded granuloma Mucin	Granuloma annulare
Nodular and diffuse	05	Nodular histiocytic inflammatory infiltrate Tuberculoid granuloma Marked necrosis in the dermis	Miliary tuberculosis Scrofuloderma tuberculosis
Nodular and diffuse	06	Nodular histiocytic inflammatory infiltrate Palisaded granuloma Degenerated collagen	Necrobiosis lipoidica
Nodular and diffuse	07	Sarcoidal granuloma Nonpolarizable foreign body Nodular histiocytic inflammatory infiltrate	Tattoo
Nodular and diffuse	08	Diffuse histiocytic inflammatory infiltrate Prominent neutrophils Bacterial infectious cause	Botryomycosis Blastomycosis-like pyoderma Rhinoscleroma
Nodular and diffuse	09	Nodular lymphocytic infiltrate Bottom-heavy inflammatory infiltrate	Lymphoma
Nodular and diffuse	10	Diffuse histiocytic inflammatory infiltrate Lipophages No histiocytic giant cells	Xanthelasma Xanthoma Xanthoma disseminatum Verruciform xanthoma
Subepidermal vesicular	01	Deep eosinophilic dermal inflammatory infiltrate Deep neutrophilic dermal inflammatory infiltrate Subepidermal blister	Arthropod bite
Subepidermal vesicular	02	Minimal lymphocytic dermal inflammatory infiltrate Subepidermal blister Subepidermal fibrosis	Blister above scar
Subepidermal vesicular	03	Moderate neutrophilic dermal inflammatory infiltrate Subepidermal blister Nuclear dust No mucin in reticular dermis	Dermatitis herpetiformis Dermatitis herpetiformis-like drug eruption Linear immunoglobulin A dermatosis Epidermolysis bullosa acquired
Subepidermal vesicular	04	Moderate neutrophilic dermal inflammatory infiltrate Thrombi in superficial dermal vessels Subepidermal blister	Septic vasculitis
Subepidermal vesicular	05	No lymphocytic dermal inflammatory infiltrate Subepidermal blister	Epidermolysis bullosa dermolytic Epidermolysis bullosa junctional Bart syndrome Epidermolysis bullosa acquired
Subepidermal vesicular	06	Minimal lymphocytic dermal inflammatory infiltrate Isolated dermal eosinophils Subepidermal blister	Bullous pemphigoid Herpes gestationis
Subepidermal vesicular	07	Moderate lymphocytic dermal inflammatory infiltrate Subepidermal blister Subepidermal fibrosis Thick collagen bundles in reticular dermis Sclerosis of papillary dermis	Lichen sclerosus et atrophicus
Subepidermal vesicular	08	Moderate neutrophilic dermal inflammatory infiltrate Nuclear dust Mucin in reticular dermis Subepidermal blister	Systemic lupus erythematosus
Subepidermal vesicular	09	No or minimal lymphocytic dermal inflammatory infiltrate Subepidermal blister Dermal papillae preserved Perivenular rims of homogeneous material Extensive solar elastosis	Porphyria cutanea tarda Erythropoietic protoporphyria Protoporphyria Variegate porphyria
Subepidermal vesicular	10	Moderate lymphocytic dermal inflammatory infiltrate Subepidermal blister Individual necrotic keratinocytes Ballooning	Erythema multiforme Mucha-Habermann disease



You are on your own!

Status: Immediate feedback

Figure 2. *Example of slide exploration. a, Area at high magnification in the image viewer. b, Area projected back in the search map.*

were obtained from the data set. Even though eye-position tracking was not used in this experiment, we used the actual panning and zooming movements of the slide to gauge where the subject's visual attention was located. In this context we defined *time to hit* as how long it took the residents, from image onset, to first position the image viewer (shown in Figure 1) so that a given diagnostic finding for a given case first became visible within the best magnification range for identification of that feature.

In addition, *dwell time* was calculated as the accrued time in which a given feature was visible in the image viewer, at any magnification level within the best magnification range, so as to allow for perception and identification of that feature. In this case, feature visibility in the image viewer was determined if at least 10% of the area of the feature was depicted in the image viewer. Thus, for each feature listed as being diagnostically significant on each case, dwell time always started as 0.000 seconds, and the counter was increased by the corresponding amount of time during which the viewer depicted the feature. When the image viewer moved away from the feature, the counter was stopped; if the image viewer came back to the feature's location, the counter restarted accruing time, using as baseline wherever it was last.

Finally, *reporting time* was calculated as how long it took the residents to report a given diagnostic finding on a given case; similarly, *diagnosis time* was calculated as how long it took the residents to provide a diagnosis (or a differential diagnosis) on the case. Both times were calculated from image onset.

Research Question 1. Analysis of the Diagnoses Made

We determined the diagnoses that were more often correctly reported by the residents, as well as those that were less often correctly reported. We also determined the effects of correct, incorrect, and missed feature identification on the correctness of the final diagnosis reported.

Research Question 2. Analysis of Slide Exploration Strategy

Search Maps.—In order to better understand the link between expertise, the slide exploration strategy used by the residents, and the residents' decision-making process, which ultimately led to the diagnosis(es) assigned to the cases, we created dynamic representations of the slide exploration strategy employed, namely, animated movies of the residents' zooming, panning, and focusing on certain areas on a given slide. This was done by recording and time stamping (using SlideTutor's internal interfaces) all of the residents' actions while reading the cases. This record allowed for a playback of the residents' slide exploration strategy. However, for interreader comparison and statistical analyses, we needed a static representation of slide exploration, and hence we created the search maps. In these maps, residents' interactions with the image viewer were first separated according to the magnification range being used at the time; so, for each case and each resident, a 3-part search map was formed, where 1 part contained information about low-magnification exploration, 1 about medium-magnification exploration, and 1 about high-magnification exploration. For place-keeping

Table 2. Counts of Corresponding Image Sampling at a Given Location by 2 Distinct Readers[a]

	Reader 2		
Reader 1	Yes	No	Total
Yes	C1	C2	total_yes1
No	C3	C4	total_no1
Total	total_yes2	total_no2	Total

[a] These data were used to calculate observer agreement.

purposes, all of the residents' viewer movements within each of these ranges were projected back onto the initial slide, which was at the lowest possible magnification level (×1), and this allowed for comparisons among different residents. Thus, if a given area was visually explored by the residents, its projection on the low-magnification slide was painted with a shade of green. In this way, the search maps allowed for a conversion between a dynamic exploration strategy and a static representation of where in the slide the residents' visual attention was focused. However, the search maps did not preserve temporal sequencing of actions, and they looked different from what the resident was seeing in the image viewer. For example, if a resident spent some time analyzing a given feature using any magnification in the high-magnification range, in the interface that particular area covered the entire image viewer, whereas in the high-magnification portion of the search map for that case that exploration was reflected as green painting in just a small area (corresponding to the projection of the originally explored location), as shown in Figure 2.

In order to compare the slide exploration strategy used by each resident and test for statistically significant differences, we quantified visual sampling as the ratio between the area of the slide visually explored by the resident (painted in green in the search maps) and the overall area of each slide, by magnification level. This normalized measure allowed us to determine whether residents with different experience levels had visually covered similar amounts of tissue in the slides.

Observer Agreement—Cohen κ.—The slide coverage measure derived in "Search Maps" did not provide any information regarding the similarities in the slide exploration strategies used by the residents. In order to assess that, we compared the search maps between the residents, by magnification level and by case, on a pixel-by-pixel basis, and we created 4 counters, all of which always started at zero at the beginning of each new comparison: (C1) number of pixels sampled by both residents; (C2) number of pixels sampled by resident i but not by resident j; (C3) number of pixels sampled by resident j but not by resident i; and (C4) number of pixels not sampled by either of the 2 residents. These values then allowed us to generate the 2 × 2 table depicted in Table 2.

Hence, observed sampling agreement can be calculated as

$$p_o = (C1 + C4)/Total$$

whereas chance sampling agreement is computed as

$$p_e = ((total\_yes1 * total\_yes2) + (total\_no1 * total\_no2))/Total^2$$

Under these conditions, observer agreement is measured by κ, which is given by

$$\kappa = (p_o - p_e)/1 - p_e$$

In general, $\kappa = 0$ means no agreement; $0.0 < \kappa \leq 0.20$, slight agreement; $0.21 \leq \kappa \leq 0.40$, fair agreement; $0.41 \leq \kappa \leq 0.60$, moderate agreement; $0.61 \leq \kappa \leq 0.80$, substantial agreement; and $0.81 \leq \kappa \leq 1.0$, almost perfect agreement.

RESULTS

Research Question 1. Perceptual Analysis of Feature Identification

For a feature-based analysis, Table 3 displays the median time to hit and the median dwell time for the ND and the SV cases according to whether the feature was correctly reported, incorrectly reported, or not reported (i.e., missed). Data are shown separately for the residents who had undergone their dermatopathology rotation and those who had not. Times are given in seconds.

As Table 3 shows, either the residents knew the features and identified them correctly or they missed (i.e., did not report) the features. There are very few instances of features that were incorrectly reported. For example, for the entire set of readings of the ND cases, only one resident made a single incorrect feature identification. For these cases, most errors were in the missed identification of diagnostic features (252 instances of such errors). In the ND cases, the residents correctly identified 77 instances of diagnostic features being present. Certain features, such as predominate lymphocytic infiltrate, were always correctly identified, whereas other features, such as parakeratosis and spongiosis, were never reported by the residents. As a result of this discrepancy, paired means comparison analysis showed a statistically significant difference in the number of diagnostic features correctly reported and those not reported ($t = -15.909$, $P < .001$).

For the SV cases, the residents incorrectly identified 40 diagnostic features, correctly identified 156, and missed 175. Contrary to the ND cases, in the SV cases there were no diagnostic features that were always correctly identified; even blisters were incorrectly identified in 13

Table 3. Median Time to Hit and Dwell Time for the Nodular and Diffuse (ND) and the Subepidermal Vesicular (SV) Cases[a]

Completed dermatopathology rotation	ND			SV		
	Correctly Identified	Incorrectly Identified	Not Reported	Correctly Identified	Incorrectly Identified	Not Reported
Time to hit, s (No. features)						
Untrained	40.984 (23)	. . .	35.844 (97)	19.625 (47)	29.109 (16)	22.422 (71)
Trained	28.734 (54)	29.485 (01)	32.062 (155)	15.312 (109)	22.015 (24)	18.203 (104)
Dwell time, s (No. features)						
Untrained	206.686 (23)	. . .	417.838 (97)	106.422 (47)	95.203 (16)	214.838 (71)
Trained	145.746 (54)	365.669 (01)	501.697 (155)	86.484 (109)	81.461 (24)	269.386 (104)

[a] Separated according to whether the residents correctly or incorrectly identified the feature, or whether the feature was not reported. Data are shown separately for residents who had (Trained) and those who had not (Untrained) undergone their dermatopathology rotation.

Table 4. Median Time Taken to Generate Correct and Incorrect Hypotheses in Nodular and Diffuse (ND) and Subepidermal Vesicular (SV) Cases by Residents With (Trained) and Without (Untrained) Dermatopathology Training, and Number of Hypotheses

	Time to Hypothesis, s (No. Hypotheses)			
	Correct		Incorrect	
Case Type	Untrained	Trained	Untrained	Trained
SV	402.0 (09)	327.0 (31)	407.0 (64)	361.0 (154)
ND	382.0 (13)	473.0 (41)	594.0 (67)	505.5 (154)

instances. Paired means comparison showed statistically significant differences between the number of incorrectly and correctly reported diagnostic features ($t = -10.545$, $P < .001$) and between the number of incorrectly identified and missed diagnostic features ($t = -12.273$, $P < .001$), but not between the number of correctly identified and missed diagnostic features ($t = -1.727$, $P = .56$).

Using the nonparametric Mann-Whitney *U* test, we compared the number of correctly reported findings between the 2 resident groups. We also ran this comparison for the number of incorrectly reported findings and for the number of not reported findings. We found no statistically significant differences between the 2 groups. We also used the Mann-Whitney *U* test to determine whether significant differences existed between the 2 resident groups in both dwell time and in time to hit the locations of correctly reported, incorrectly reported, and not reported findings. We found no statistically significant differences.

We also wanted to determine whether statistically significant differences existed between the 2 groups of residents when we compared how long they took to report the findings. For this we used analysis of variance, and used a Scheffé post hoc test to determine whether any observed differences were statistically significant. We found significant differences for both ND and SV cases, with residents who had not undergone their dermatopathology rotation taking significantly longer to report findings (for ND cases, median reporting times for residents who had undergone training was 264 seconds, whereas for those who had not it was 381 seconds, $F_1 = 39.265$, $P < .001$; for SV cases, median reporting times for residents who had undergone training in dermatopathology was 203 seconds, whereas for those who had not it was 239 seconds, $F_1 = 15.543$, $P < .001$).

Research Question 1. Analysis of the Diagnoses Made

Table 4 shows the total number of hypotheses generated by the residents in the ND and SV cases according to whether the residents had or had not undergone their dermatopathology rotation. Averaging the results shown by the number of residents in each category and by the number of cases, we can see that for the ND cases, residents without dermatopathology training generate on average, per case, 0.325 correct hypotheses and 1.675 incorrect hypotheses, whereas residents with subspecialty training generate 0.586 correct hypotheses and 2.20 incorrect hypotheses. For the SV cases, residents without dermatopathology training generate, on average, per case, 0.225 correct hypotheses and 1.60 incorrect hypotheses, whereas the residents with dermatopathology training generate on average 0.443 correct hypotheses and 2.20 incorrect hypotheses. In order to determine whether these differences were statistically significant, we used the Mann-Whitney *U* test, and we found that there were significant differences in the number of correct hypotheses generated ($z = -2.525$, $P = .01$) and in the number of incorrect hypotheses generated ($z = -2.082$, $P = .04$) between the 2 groups of residents.

We also sought to determine whether the 2 resident groups differed in how long they took to report their hypotheses. For this we used the Mann-Whitney *U* test, and for the SV cases we found no significant differences, either when the generated hypotheses were correct ($z = -0.567$, $P = .57$) or incorrect ($z = -0.897$, $P = .37$). For the ND cases, there were no differences for the correct hypotheses ($z = -0.212$, $P = .83$), but significant differences for the incorrect hypotheses ($z = -3.307$, $P < .001$).

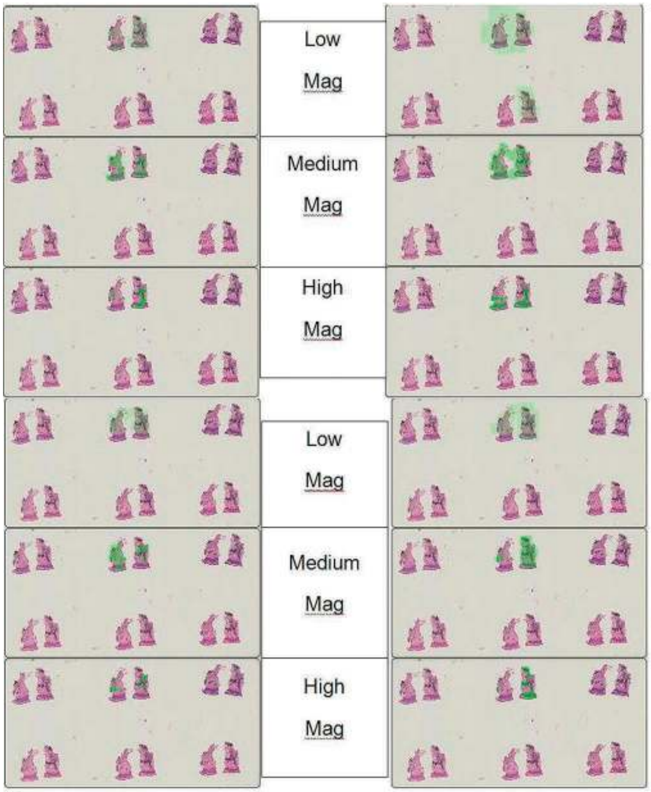
In each domain (ND and SV), 2 cases were not correctly diagnosed by any of the residents, regardless of whether or not they had had their dermatopathology rotation. These cases were numbers 06 (necrobiosis lipoidica) and 08 (botryomycosis) for the ND dermatitides and numbers 05 (epidermolysis bullosa dermolytic) and 08 (systemic lupus erythematosus) for the SV dermatitides, as listed in Table 1. The cases that most of the residents diagnosed correctly were numbers 01 (xanthogranuloma), 02 (sarcoidosis), 07 (tattoo) and 10 (xanthelasma) for the ND domain, and for the SV domain, they were cases 02 (blister above scar) and 03 (dermatitis herpetiformis). For the ND cases, 73% of the correct diagnoses were made by residents who had undergone their dermatopathology rotation, whereas for the SV cases that fraction was 76%.

In order to determine whether the number of features correctly, incorrectly, or not reported influenced the final diagnosis, we used paired means comparison to determine whether statistically significant differences existed between the cases for which the final diagnoses had been correct and those for which it had been incorrect. Surprisingly, we found no statistically significant differences either for the SV cases (incorrectly identified features, $t_{10} = -1.226$, $P = .25$; correctly identified features, $t_{10} = -1.158$, $P = .27$; and not reported features, $t_{10} = 1.870$, $P = .09$), or for the ND cases (correctly identified features, $t_{10} = 0.827$, $P = .43$; not reported features, $t_{10} = 0.975$, $P = .35$).

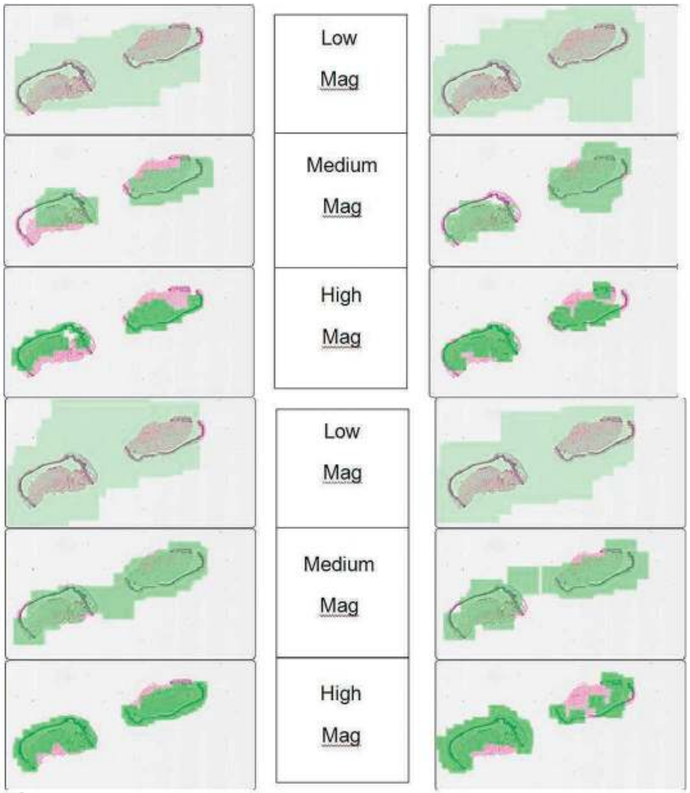
Because that result was surprising, we decided to investigate it further by determining whether dermatopathology training contributed in any way to the lack of statistical significance (for example, by having the results of one group of residents overwhelm the overall results

Table 5. Median Times Taken to Report the Final Diagnosis on the Cases, Separated According to Whether the Diagnosis Was Correct or Incorrect and to Whether the Resident Had Undergone Dermatopathology Training (Trained) or Not (Untrained)

	Time to Diagnosis, s			
	Correct		Incorrect	
Case Type	Trained	Untrained	Trained	Untrained
ND	672.0	451.5	654.0	696.0
SV	492.0	509.0	554.0	471.0



3



4

Figure 3. *Slide exploration of 4 residents in a nodular and diffuse case for which all provided correct diagnosis: upper panel represents exploration at low magnification; middle panel, medium magnification; lower panel, high magnification. Green denotes areas visually inspected.*

Figure 4. *Slide exploration strategy of 4 residents in a subepidermal vesicular case where none provided correct diagnosis: upper panel represents exploration at low magnification; middle panel, medium magnification; lower panel, high magnification. Green denotes areas visually inspected.*

and thus disguise any existing differences). We used analysis of variance, where the independent variable was binary, namely, whether the final diagnosis was correct or incorrect, and the dependent variables were the number of features in each category (incorrectly, correctly, or not reported) and dermatopathology training (with or without for each resident). Again, no statistically significant effects were found, either for the SV or for the ND cases.

Finally, we looked at how long it took the residents with and without dermatopathology training to report their final diagnosis on the case (diagnosis time). This is shown in Table 5, and separated according to whether the diagnosis was correct or incorrect. We used analysis of variance (with Scheffé post hoc test) to determine whether there were statistically significant differences in the diagnosis time between the 2 groups of residents. Again, we found statistically significant differences for both ND and SV cases, with residents who had not undergone their rotation taking significantly longer to provide their diagnosis on the cases (for ND cases, median diagnosis time for residents who had undergone dermatopathology rotation was 668 seconds, whereas for those who had not it was 780 seconds, $F_1 = 28.070$, $P < .001$; for the SV cases, median diagnosis time for residents who had completed their dermatopathology rotation was 553 seconds, whereas for those who had not it was 547 seconds, $F_1 = 11.447$, $P < .001$).

Research Question 2. Analysis of Slide Exploration Strategy

Search Maps.—For each ND and each SV case explored by the residents, we created search maps that depicted the slide exploration strategy used, by magnification level range. Figure 3 shows an example of the search maps for 4 different residents as they examined a ND case in which all 4 residents agreed on the correct diagnosis (tattoo). In contrast, Figure 4 shows an example of the search maps for 4 residents as they examined an SV case in which none of the observers arrived at the correct diagnosis (epidermolysis bullosa dermolytic).

Initial visual inspection of the patterns shown in Figures 3 and 4 suggests that when residents have an initial idea about what the actual diagnosis on a case may be, their slide exploration strategy is fairly similar in that they are able to discard most of the distracting information available in the image and concentrate only in the areas that will help them to arrive at the diagnosis. In these instances their search strategy is very focused, and it starts with limited exploration at low and medium power, with very selective use of high power. On the other hand, as shown in Figure 4, when residents do not have any idea what a diagnosis may be, they freely explore the slide at all power levels, as if seeking a clue that may solve the mystery. In these cases search is laborious and unfocused, and as they cannot interpret the perceptual information being captured (but not processed) from the image at all power levels, the diagnostic space cannot be constrained.

In order to test whether the search strategy is indeed different between cases for which the diagnosis is correct and those for which it is incorrect, we started by calculating the relative area covered by the residents' slide exploration strategy, by magnification level. We used

Table 6. Cohen Coefficient κ by Magnification Range (Low, Medium, and High) and by Case Type (Nodular and Diffuse [ND] and Subepidermal Vesicular [SV])

	Magnification		
Case Type	Low	Medium	High
SV	0.452	0.389	0.302
ND	0.500	0.358	0.254

the Wilcoxon signed rank test to determine whether there were statistically significant differences in the relative coverage between cases that were correctly diagnosed and those that were incorrectly diagnosed. This yielded significant differences at all magnification levels for the ND cases (low magnification, $z = -2.534$, $P = .01$; medium magnification, $z = -2.667$, $P = .01$; high magnification, $z = -2.934$, $P = 0.003$) but only at the high magnification levels for the SV cases (low magnification, $z = -1.540$, $P = .12$; medium magnification, $z = -0.770$, $P = .44$; high magnification, $z = -2.380$, $P = .02$).

Finally, we have sought to determine whether dermatopathology training influences image coverage in 2 conditions: (1) when the final diagnosis is correct; and (2) when the final diagnosis is wrong. We used analysis of variance (Scheffé post hoc test) in this determination, and we found no statistically significant differences in image coverage between the residents who had undergone their dermatopathology rotation and those who had not, for either the ND or the SV cases.

Observer Agreement—Cohen κ.—Table 6 lists the values of kappa for the ND and SV cases, by magnification range. As the table shows, agreement is moderate for both ND and SV cases at low magnification range, and it is slowly reduced as the magnification range increases.

Kappa, as we defined it, is a measure of the agreement in the slide exploration strategy used by the residents. If we restrict the pairings to those formed by either 2 residents who had undergone their dermatopathology rotation or 2 residents who had not undergone the rotation (and hence leave out all of the mixed pairs, in which one resident had and the other had not undergone the rotation), we can use the Mann-Whitney *U* test to determine whether statistically significant differences exist between the 2 types of pairs of residents. In carrying out such analysis we found significant differences only for the ND cases, when examined at high magnification ($z = -2.45$, $P = .01$). In this case, we found more similarities when we compared the exploration strategies employed by the residents who had undergone their dermatopathology rotation than when we compared the exploration strategy of residents who had not undergone such rotation yet.

COMMENT

Pathology may be the gold standard of medical diagnoses, but the process by which pathologists arrive at the decisions they make is still very much unknown. This process, a combination of the pathologists' slide exploration strategy, their perceptual gathering of information, and the cognitive integration of this information into decisions, has had some of its components explored, particularly in what relates to the cognitive aspects of decision making. For example, there is the multiscale approach model, which advocates that pathologists identify suspicious regions at low magnification and then use the data at higher magnifications to confirm or refute the diagnostic hypotheses originally formed.[25]

Aiming to better understand the integration between the perceptual and the cognitive components of the process, Crowley et al[20] carried out a study in the development of visual diagnostic expertise in the reading of breast histopathology slides by analyzing search patterns and verbal protocol (think-aloud) data from 3 observer groups: board-certified pathologists (''experts''), pathology residents (''intermediates'') and third-year medical students (''novices''). These authors found that the residents could accurately detect a significant fraction of the diagnostic findings in the case set, but they could not properly integrate these findings into a clinically coherent diagnosis, which suggested that these observers possessed good perceptual learning abilities for detection of disease characteristics but either had poor formation of initial hypothesis on the case or had faulty access to cognitive schema, which would have resulted in difficulties when integrating the findings with proper disease classification.

This interpretation supports a model of medical image interpretation proposed by Kundel and Nodine[32] according to which detection of relevant diagnostic information and correct rendition of diagnosis depends on (not necessarily in this order) (1) the observer's visual search strategy; (2) the observer's ability to disambiguate relevant perceptual information from background noise; (3) the observer's cognitive interpretation of perceived findings; and (4) the observer's experience integrating the interpreted findings into a decision about the case. In this model, experts perceive (2) and interpret (3) findings, reach a decision about the case (4), and then search to see whether anything else is present (1), in a *detect-then-search* approach, whereas intermediates and novices follow the traditional *search-then-detect* route.[28] Experts' ability to reorder the traditional steps stems from their increased ability to holistically integrate perceived features, which are captured by the central and the peripheral vision at image onset.[33] In this context holistic integration is a direct function of experience reviewing thousands of images, and perceptually learning to identify, very quickly, informative features. As shown by Crowley et al,[20] intermediates and novices lack such a large internal dictionary of what differentiates signal from noise in a medical image; hence, they must visually scan it in order to find the locations of interest. Furthermore, in seeking to acquire expertise in the reading of breast histopathology slides, intermediates fail on (3), whereas novices fail on (2).

Nonetheless, the role of visual search in diagnostic interpretation of histopathologic slides is still unclear. In order to further identify the influence of visual scanning of slides on decision making strategy, Krupinski et al[23] used eye-position tracking, wherein expert pathologists, pathology residents, and medical students examined a set of fixed, low-magnification digital slides of breast core biopsies. The data found by Krupinski et al suggested 2 general types of scanning patterns: (1) a strategy in which many different areas of the slide were fixated on for a short period of time (adopted primarily by residents and medical students), and (2) a strategy in which fewer areas of the slide attracted visual attention, but for longer periods of time (observed among the experts). This dichotomous search strategy is in agreement with Kundel and Nodine's model,[32] in that it suggests that because experts have greater perceptual integration of features, they need to examine fewer areas in the slide, but each examined area takes longer to process because of cognitive decision making at the location. On the other hand, trainees have to search the slides in order to find locations with potentially relevant information, which yields many fixation points, but because no decision-making process is necessarily carried out in each of these areas, dwell time is short at each location.

In the study of Krupinski et al,[23] the observers viewed the digital slides at fixed magnification, but this is not how pathologists read images in their clinical practice, and it may have biased the observers' scanning strategies. In order to avoid this possible pitfall, Treanor et al[24] conducted an eye-position tracking study in histopathology in which trainees and experts were allowed to zoom in and to pan on virtual slides, thus mimicking the clinical practice. These authors found, similarly to Crowley et al,[20] that trainees could correctly identify the areas where the abnormal tissue was located, but they failed to cognitively integrate these findings into an appropriate diagnosis. Furthermore, even though trainees and experts spent similar amounts of time examining the slides at low magnification (<×5), trainees spent significantly longer at high magnification (>×10), perhaps suggesting a greater difficulty to integrate the information viewed at high power with the overall percept of the image.

Residents in our study used an interface similar to the method used by Treanor et al[24] that allowed for zooming and panning of the virtual slide. However, unlike that study, our study did not use eye-position tracking, and instead used the residents' slide exploration strategy as a proxy for determining where their visual attention was directed. Our primary goal was to acquire more information about the development of expertise, and with it to better understand the process by which pathologists make decisions.

To this end we tried to answer 2 research questions. In research question 1, our focus was to determine when in the training of a pathologist perceptual learning and cognitive integration of findings with diagnosis(es) develops. Clearly this is a very broad and ambitious question, and hence we have focused our analyses in a limited area, a subdomain of dermatopathology, namely, inflammatory skin disease.

Our results suggested that there were no statistically significant differences in the number of findings that were correctly, incorrectly, or not reported by the residents who either had or had not undergone their dermatopathology rotation. This was surprising, and it may suggest that perceptual learning for some abnormal feature characteristics in inflammatory skin disease is acquired perhaps as early as during the initial rotations in pathology residence. This possibility is supported by the lack of significant differences between the 2 groups of residents in time to hit, a measure traditionally employed in perceptual research to determine how quickly abnormal feature configuration, which differs from anticipated schemata of the image, attracts visual attention. In this context time to hit is a measure of global (ie, textural) image characteristics, as it is the disturbance itself, and not the nature of the disturbance (namely, the specific type of finding), that causes visual attention to shift to a specific area of the slide, which is then examined at greater length using high-resolution vision. At each of these locations attentional engagement is measured by dwell time, and, although we found statistically significant differences for some findings when we compared dwell time at the locations where these findings had been correctly reported or not reported, these differences were not significant between the 2 resident groups.

One may argue that a possible reason why we did not find any significant differences in finding identifications between the 2 groups of residents is that recognition of isolated diagnostic criteria may not necessarily precede diagnostic assessment. However, we do not believe this to be the case, as such a collapsed decision-making process has been shown only for experts who operate in a holistic (or integrative) manner,[33] and not for novices. Nonetheless, in this study we had to follow this serial model because both theories of learning being investigated propose some type of identification of criteria before diagnostic decision making.

In addition, feature identification did not seem to have any statistically significant effect on the correctness of the final diagnosis, for either group of residents, for both ND and SV cases. This finding is compatible with the model of acquisition of expertise proposed by Lesgold et al,[29] in which proficiency in reading medical images starts with the acquisition of subsymbolic discrimination abilities at the beginning of residency, and then progresses to processing the image using only perceptual cues (that is, identifying findings without understanding their diagnostic relevance). In this stage diagnoses are formed by generating all hypotheses that are compatible with the perceived findings. Unfortunately, this often leads to discrimination insufficiency,[34] which in the current study was represented by the large proportion of incorrect diagnoses reported. Furthermore, according to the model of Lesgold et al,[29] as a resident becomes more practiced at reading slides, cognitive processing begins to develop, and contextual information from the image is used to arrive at a diagnosis. In our data this is reflected by the fact that more than 70% of the correct diagnoses were made by the residents who had undergone their dermatopathology rotation.

In our second research question, we asked, ''How does this budding expertise manifest itself in the exploration of the virtual slide?'' In the late 1960s, Neisser[35] reasoned that image perception is a 2-stage process, in which (1) a *preattentive stage* analyzes the entire slide in parallel using the central and peripheral vision, and a global view of what is being displayed is acquired; this is followed by (2) a *focal stage* in which items or groups of items are examined under greater scrutiny by the foveal vision. In this process the preattentive stage may bias the selection of the areas that will be subjected to further examination.[36]

As there are no direct ways to measure preattentive processing, time to first fixate diagnostically significant findings (ie, time to hit) is commonly used as a proxy to determine the degree of information from the image that the observer has acquired from the global view. In this study, as previously mentioned, no statistically significant differences were noted for time to hit between the 2 groups of residents, which suggests that their initial global views of the slides probably contained similar amounts of information. On the other hand, attentional engagement in the focal stage, that is, local processing of diagnostically relevant features, can be measured in many ways, such as by contrasting the differences in image coverage between the 2 groups of residents. In this comparison the underlying hypothesis is that better training will lead to more focused search, with less exploration of the overall slide, which is a behavior often seen in experts, as shown by Krupinski et al.[23]

Our results for this comparison initially seemed contradictory. First, we did not find significant differences between the 2 resident groups for slide exploration at low, medium, or high magnification levels. However, we did find significant differences in image coverage according to whether the final diagnosis on the case was correct or incorrect. When the final diagnosis was correct, the residents' search strategy was focused and efficient, which suggests that they properly assessed the information on the case in order to arrive at the appropriate diagnosis. On the other hand, when their final diagnosis was incorrect, the observed search strategies were spread out and time consuming, suggesting that the residents did not know how to interpret the information present in the case. This supports Antes and Penland's[36] hypothesis that the preattentive stage may impose an initial bias in the reading of the cases; hence, when the residents start the reading with some idea about what the final diagnosis may be, this bias guides them to efficiently explore the slide (as shown in Figure 3), whereas when the preattentive stage does not offer any possible clues about the nature of disease in the case (or it offers too many contradicting possibilities), a fishing expedition ensues, in which residents must seek information over all the tissue portions of the slide, using a costly strategy.

Although our data did not completely support either the Dreyfus model[26] of acquisition of expertise or the model proposed by Lesgold et al,[29] for the most part our results seemed biased towards the inference by Lesgold et al that perceptual learning precedes cognitive-inferential decision making when one is receiving training in a visually based domain. This assertion is supported by 4 results: (1) the lack of statistically significant differences in the number of correct, incorrect, and not reported findings between the residents with and without dermatopathology training; (2) the significantly longer dwell times and reporting times observed for diagnostically relevant findings for residents who had undergone their dermatopathology rotation, which suggests the beginnings of a process of cognitive discrimination. Also, (3) these residents generated more correct and incorrect hypotheses, on average, per case, than their peers. Interestingly, the mean number of correct hypotheses generated with training increased by 80% for ND cases and by 97% for SV cases, which far surpassed the increase in the mean number of incorrect hypotheses generated (31% for ND cases and 38% for SV cases). This increase suggests greater cognitive integration between perceived diagnostic findings and disease schemata. Finally, (4) residents with training reported their final diagnosis in less time than their peers, and were more accurate.

Interestingly, neither model truly explains why significant differences were observed in image coverage according to whether the final diagnosis was correct or not, but not between the resident groups. According to Neisser's theory,[35] the efficient and focused exploration strategy observed when the final case diagnosis was correct could be the result of preattentional bias. This bias, primarily caused by subsymbolic perceptual features, would support the model of Lesgold et al[29] had it been stronger after dermatopathology training. Given that it was similar for both groups of residents, it seems to suggest that subsymbolic (ie, perceptual) processing did not change much during the dermatopathology rotation; rather, what differentiated the residents with training from those without was better cognitive processing of perceived findings. If this was the case, when does subsymbolic processing develop? During the medical school years, as medical students learn by book-based examples? During the first year of pathology residency, as they learn by practicing? Although our data do not allow us to answer these questions, future research will be carried out to look at this specific issue.

Our study has several limitations. Among these was the small sample size used, which impacted the statistical power of our analyses. Although we are cognizant of that, we felt that in this preliminary, exploratory study of 2 theories of learning in a visual domain, a small caseload was needed because we attempted to capture every step in the residents' decision making process. This led to unusually long case-reading times (as compared to reading times in the clinic), because in our experiment the residents had to mark diagnostic findings, report diagnostic hypotheses, etc. In this scenario, a large caseload, although highly beneficial statistically, would have significantly reduced enrollment in the experiment, because of the residents' time constraints. Hence, there is the possibility that a Type II error occurred, as this type of error is related to the statistical power of a given test. We intend to address this issue with larger studies in the future, in which the number of required actions from the observers will be significantly reduced, thus allowing us to use a larger case set.

Another limitation is that we only had 4 residents who had not undergone dermatopathology rotation, which may reduce the generalizability of our results to this population. Moreover, all residents from our institution who had undergone their rotation were trained by the same dermatopathologist, so they could have had a tendency to report findings similarly. Because of the small data set, we could not further divide the analysis into residents from institution 1 versus residents from institution 2. Other resident-related factors may have impacted this study as well. For example, because this was a self-referred, paid experiment, the commitment of the residents to the task varied. For the most part, they seemed interested and engaged, but we did not formally evaluate whether that was true. In addition, residents' prior rotations may have influenced their identification of diagnostic findings and slide exploration strategy, but we did not keep track of their previous rotations, nor would we have been able to analyze their effect given our small observer sample. Finally, it is possible that our data reflect to some extent the ''checklist effect,'' because our user interface—the light version of SlideTutor—did not allow residents to write in the name of the findings that they wished to report, but rather forced them to choose from a preselected list of findings. We do not believe this to be a major limitation, because the list of preselected findings contains all possible findings that existed on all slides of a given type (ND or SV). Furthermore, we reasoned that it would reduce observer variability if we did not allow the residents to come up with the names of the findings themselves, but instead had them choose from a preselected list.

This work was supported by the following grants: NIH/NLM R01 LM007891-05 and AHRQ K01 HS018365-01. We also would like to thank the anonymous reviewers for their suggestions in improving the clarity and the quality of the manuscript.

References

1. Renshaw AA, Pinnar NE, Jiroutek MR, Young ML. Quantifying the value of in-house consultation in surgical pathology. *Am J Clin Pathol*. 2002;117(5):751–754.

2. Renshaw AA, Gould EW. Comparison of disagreement and amendment rates by tissue type and diagnosis. *Am J Clin Pathol*. 2006;126(5):736–739.

3. Shoo BA, Sagebiel RW, Kashani-Sabet M. Discordance in the histopathologic diagnosis of melanoma at a melanoma referral center. *J Am Acad Dermatol*. 2010;62(5):751–756.

4. Trotter MJ, Bruecks AK. Interpretation of skin biopsies by general pathologists: diagnostic discrepancy rate measured by blinded review. *Arch Pathol Lab Med*. 2003;127(11):1489–1492.

5. Piepkorn MW, Barnhill RL, Cannon-Albright LA, et al. A multiobserver, population-based analysis of histologic dysplasia in melanocytic nevi. *J Am Acad Dermatol*. 1994;30(5):707–714.

6. Farmer ER, Gonin R, Hanna MP. Discordance in the histopathologic diagnosis of melanoma and melanocytic nevi between expert pathologists. *Hum Pathol*. 1996;27(6):528–531.

7. Kundel HL, Nodine CF, Carmody DP. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Invest Radiol*. 1978;13(3):175–181.

8. Carmody DP, Nodine CF, Kundel HL. An analysis of perceptual and cognitive factors in radiographic interpretation. *Perception*. 1980;9(3):339–344.

9. Nodine CF, Kundel HL. Using eye movements to study visual search and to improve tumor detection. *Radiographics*. 1987;7(6):1241–1250.

10. Kundel HL, Nodine CF, Krupinski EA. Visual dwell indicates locations of false-positive and false-negative decisions. *Invest Radiol*. 1989;24:472–478.

11. Renfrew DL, Franken Jr EA, Berbaum KS, Weigelt FH, Abu-Yousef MM. Error in radiology: classification and lessons in 182 cases presented at a problem case conference. *Radiology*. 1992;183(1):145–150.

12. Berbaum KS, Franken EA Jr, Anderson KL, et al. The influence of clinical history on visual search with single and multiple abnormalities. *Invest Radiol*. 1993;28(3):191–201.

13. Nodine CF, Kundel HL, Mello-Thoms C, et al. How experience and training influence mammography expertise. *Acad Radiol*. 1999;6(10):575–585.

14. Nodine CF, Mello-Thoms C, Kundel HL, Weinstein SP. Time course of perception and decision making during mammographic interpretation. *AJR Am J Roentgenol*. 2002;179(4):917–923.

15. Manning DJ, Ethell SC, Donovan T. Detection or decision errors?: missed lung cancer from the posteroanterior chest radiograph. *Br J Radiol*. 2004;77(915):231–235.

16. 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.

17. Berbaum K, Franken EA, Caldwell RT, Schartz KM. Can a checklist reduce SOS errors in chest radiography? *Acad Radiol*. 2006;13(3):296–304.

18. Mello-Thoms C. How does the perception of a lesion influence visual search strategy in mammogram reading? *Acad Radiol*. 2006;13(3):275–288.

19. Kundel HL, Nodine CF, Krupinski EA, Mello-Thoms C. Using gaze-tracking and mixture distribution analysis to support a holistic model for the detection of cancers on mammograms. *Acad Radiol*. 2008;15(7):881–886.

20. Crowley RS, Naus GJ, Stewart J III, Friedman CP. Development of visual diagnostic expertise in pathology: an information-processing study. *J Am Med Inform Assoc*. 2003;10(1):39–51.

21. Tiersma ESM, Peters AAW, Mooij HA, Fleuren GJ. Visualising scanning patterns of pathologists in the grading of cervical intraepithelial neoplasia. *J Clin Pathol*. 2003;56(9):677–680.

22. Schrader T, Niepage S, Leuthold T, et al. The diagnostic path, a useful visualization tool in virtual microscopy. *Diagn Pathol*. 2006;1:40.

23. 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 expertise. *Hum Pathol*. 2006;37(12):1543–1556.

24. Treanor D, Lim CH, Magee D, Bulpitt A, Quirke P. Tracking with virtual slides: a tool to study diagnostic error in histopathology. *Histopathology*. 2009;55(1):37–45.

25. Doyle S, Rodriguez C, Madabhushi A, Tomaszeweski J, Feldman M. Detecting prostate adenocarcinoma from digitized histology using a multi-scale hierarchical classification approach. *Conf Proc IEEE Eng Med Biol Soc*. 2006;1:4759–4762.

26. Dreyfus HL, Dreyfus SE. *Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. New York, NY: Free Press; 1986.

27. Kulatunga-Moruzi C, Brooks LR, Norman GR. Using comprehensive feature lists to bias medical diagnosis. *J Exp Psychol Learn Mem Cogn*. 2004;30(3):563–572.

28. Nodine C, Mello-Thoms C. The role of expertise in radiologic image interpretation. In: Samei E, Krupinski E, eds. *Medical Image Perception and Techniques*. New York, NY: Cambridge University Press; 2010:139–156.

29. Lesgold AM, Rubinson H, Feltovich P, et al. Expertise in a complex skill: diagnosing x-ray pictures. In: Chi MTH, Glaser R, Farr MJ, eds. *The Nature of Expertise*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988:311–342.

30. Harries C, Evans JSBT, Dennis I. Measuring doctors' self-insight into their treatment decisions. *Appl Cogn Psychol*. 2000;14(5):455–477.

31. Crowley RS, Medvedeva O. An intelligent tutoring system for visual classification problem solving. *Artif Intell Med*. 2006;36(1):85–117.

32. Kundel HL, Nodine CF. A visual concept shapes image perception. *Radiology*. 1983;146(2):363–368.

33. Kundel HL, Nodine CF, Conant EF, Weinstein SP. Holistic component of image perception in mammogram interpretation: gaze-tracking study. *Radiology*. 2007;242(2):396–402.

34. Raufaste E, Eyrolle H, Marine C. Pertinence generation in radiological diagnosis: spreading activation and the nature of expertise. *Cogn Sci*. 1998;22(4):517–546.

35. Neisser U. *Cognitive Psychology*. Englewood Cliffs, NJ: Prentice-Hall Inc; 1967.

36. Antes JR, Penland JG. Picture context effects on eye movement patterns. In: Fisher DF, Monty RA, Senders JW, eds. *Eye Movements: Cognition and Visual Perception*. Hillsdale, NJ: Lawrence Erlbaum Publishers; 1981:157–170.