Research Paper ■

Development of Visual Diagnostic Expertise in Pathology: An Information-processing Study

REBECCA S. CROWLEY, MD, MS, GREGORY J. NAUS, MD,
JIMMIE STEWART III, MD, CHARLES P. FRIEDMAN, PHD

Abstract **Objective**: To identify key features contributing to trainees' development of expertise in microscopic pathology diagnosis, a complex visual task, and to provide new insights to help create computer-based training systems in pathology.

Design: Standard methods of information-processing and cognitive science were used to study diagnostic processes (search, perception, reasoning) of 28 novices, intermediates, and experts. Participants examined cases in breast pathology; each case had a previously established gold standard diagnosis. Videotapes correlated the actual visual data examined by participants with their verbal "think-aloud" protocols.

Measurements: Investigators measured accuracy, difficulty, certainty, protocol process frequencies, error frequencies, and times to key diagnostic events for each case and subject. Analyses of variance, chi-square tests and post-hoc comparisons were performed with subject as the unit of analysis.

Results: Level of expertise corresponded with differences in search, perception, and reasoning components of the tasks. Several discrete steps occur on the path to competence, including development of adequate search strategies, rapid and accurate recognition of anatomic location, acquisition of visual data interpretation skills, and transitory reliance on explicit feature identification.

Conclusion: Results provide the basis for an empirical cognitive model of competence for the complex tasks of microscopic pathology diagnosis. Results will inform the development of computer-based pedagogy tools in this domain

■ **J Am Med Inform Assoc.** 2003;10:39–51. DOI 10.1197/jamia.M1123.

Affiliations of the authors: Center for Pathology Informatics, Center for Biomedical Informatics, Department of Pathology, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania (RSC); Department of Pathology, University of Pittsburgh School of Medicine, Pittsburgh Pennsylvania (GJN); Center for Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh Pennsylvania (CPF); Department of Pathology and Laboratory Medicine, University of Kentucky Medical Center, Lexington, Kentucky (JS).

The authors thank Michael Sendek for assistance with digital video collection, Guido Gatti for help with statistical analyses, and the late Herbert Simon, PhD, for use of the Protocol Analyst's Workbench (PAW). This work was supported by the National Library of Medicine Medical Informatics Training Grant Number 5-T15- LM07059 and a grant from the University of Pittsburgh School of Medicine–Pathology Postdoctoral Research Training Program (PPRTP). Preliminary findings from this study were reported in the Proceedings of the AMIA Fall Symposium, 2001.

Correspondence and reprints: Rebecca S. Crowley, MD, MS, Center for Pathology Informatics, UPMC Shadyside Cancer Pavilion–307, 5230 Centre Avenue, Pittsburgh, PA 15232; e-mail: <crowleyrs@msx.upmc.edu>.

Received for publication: 3/14/02; accepted for publication: 8/28/02.

Microscopic pathology, a subdiscipline of pathology, focuses on diagnosis of disease by histologic examination. Patients' tissues and cells obtained during biopsies, aspirates, and operations are permanently affixed to glass slides, stained, and examined. The assignment of a pathologic diagnosis is critical for any patient with cancer and for many other diseases. Pathologic diagnostic classification is reported to the referring physician and determines therapy and prognosis. Although models of the diagnostic process in microscopic pathology have been advanced,[1,2] few, if any, have achieved empirical validation. No previous studies have identified key features required to establish human expertise in this domain.

Pathology residencies typically require 5 years of training, approximately half of which is devoted to acquiring skills in diagnostic surgical pathology. To develop expertise, trainees require long residencies (and often additional subspecialty fellowships) to expose them to a sufficiently large number of cases, including a wide variety of rare and unusual patterns.

This study used information-processing and cognitive science methodologies to compare the visual diagnostic processes of novice, intermediate, and expert pathologists. Two goals were sought: (1) to understand basic cognitive processes underlying visual diagnostic expertise and (2) to gain insights useful for developing an intelligent Pathology tutoring system.[3]

Computer-based educational applications in microscopic pathology diagnosis can supplement traditional training by exposing trainees to large numbers of rare patterns in a short time. Intelligent tutoring systems (ITS) have an advantage over standard computer-assisted instruction (CAI) because they provide a simulated, realistic task environment in which individualized coaching and feedback can occur. Previous ITS work indicates that successful systems closely couple content with empirical research to (1) define the tutoring task, (2) characterize the scaffold of steps to expertise for tutoring, (3) determine the cognitive rules forming the basis of expertise in that domain, and (4) identify important errors and misconceptions made by students. Following previous work in similar domains,[4–6] we designed our study to analyze the development of expertise.

Background

Substantial previous work has identified some of the cognitive mechanisms underlying development of diagnostic expertise. Early work attempted to define medical diagnostic expertise in terms of general heuristics useful across many domains. Widely generalizable heuristics proved difficult to find, other than frequent reliance on the hypothetico-deductive method.

Elstein concluded that problem-solving was characterized by content specificity.[7] Subsequently, a variety of theories have attempted to characterize differences between novice and expert clinicians. Schmidt et al. described clinical expertise in terms of the development of exemplars of disease, termed "illness scripts."[8] Patel and colleagues observed differences in the "direction" of reasoning.[9,10] Both novices and experts who made erroneous diagnoses employed "top-down" or "backward" approaches, reasoning from hypotheses to evidence, whereas experts who issued accurate diagnoses typically reasoned in a bottom-up or "forward" manner from evidence to hypotheses. Other investigators have focused on the evoking of an initial problem representation (in the form of a differential diagnosis and associated tasks) as a method for creating structure in an ill-structured domain.[11]

The authors found no prior observational studies of the development of expertise in microscopic diagnosis. However, previous studies in radiology and dermatology—domains with prominent visual components—have explored diagnostic expertise, employing a variety of methods, such as "think-aloud" protocols, eye-tracking, and theoretical frameworks including information-processing, classical decision-making, and signal detection theory.[12–16] Studying the interpretation of complex chest x-rays, Lesgold et al. used information-processing methods to describe differences among novice, intermediate, and expert radiologists.[12] Experts reported more findings, verbalized more causes and effects, and showed more and longer reasoning chains than novices. Experts typically built mental representations of patient anatomy, evoked a pertinent schema quickly, and exhibited flexibility in tuning of these schemata. The authors raise the interesting suggestion that the balance of recognition and inference in diagnosis seems to vary with experience. They further suggest that purely perceptual learning occurs earlier in the course of learning than learning of the cognitive processes associated with inference. Consequently, the "emerging cognitive capability will have to contend with a stronger perceptual ability already in place" (p 337).

To inform the development of RadTutor (an intelligent tutor for mammography interpretation[4,5]), Azevedo and Lajoie used similar methods to study expertise in mammogram interpretation. Studying ten radiology residents and ten attending radiologists, they found no significant differences in accuracy, total time on task, number of radiologic findings, number of observations, number of diagnoses, frequency of operators, or frequency of errors between residents and attending radiologists.

In a series of studies describing expertise in dermatology, Norman et al. used stimuli consisting of color kodachrome slides of skin lesions.[17–19] They developed two different models of expertise for visual diagnostic tasks: (1) "Independent Cues" interpretation, in which "Learners gain expertise mainly by acquiring knowledge about the specific features that are best able to differentiate among diseases" (p 1064),[17] and (2) "Instance-based Categorization,"[18] in which expertise derives from rapid pattern recognition mechanisms that help experts match the case at hand to previously encountered examples. The authors measured response time and accuracy for typical and atypical cases. The "Independent Cues" hypothesis predicts that (1) errors would be more likely in atypical cases and (2) this principle should interact with expertise (i.e., experts should be relatively more accurate on typical slides). Their findings confirmed the first prediction but not the second, leading them to conclude that the results are at odds with a rule-based model of expertise and support the competing case-based hypothesis. The authors argue that "errors are not predictable on the basis of stable characteristics of the features of the lesion" (p. 1067).[17]

A significant body of work relates to perception and search in the field of radiology. Beginning with the pioneering work of Kundel in the 1960s,[13] visual search and feature recognition have been studied using eye-tracking experiments. Subjects' prior knowledge was determined, and experimenters varied relevant task-related aspects. Analysis was done using a framework of signal detection theory. Kundel et al. showed that development of expertise in reading mammograms is associated with faster search times, greater efficiency, and improved discriminatory abilities.[14] In later work they suggested that basic perceptual-cognitive units "feed the interpretive decision-making process."[15] In contrast to Patel et al., they hypothesized that a "top-down" or "backwards" reasoning strategy predominates as expertise evolves.[16] Furthermore, they theorized that expertise in radiology search and perception depends on a four-stage serial process composed of (1) a global impression, (2) a discovery search stage, (3) a reflective search stage, and (4) post-search recall.[15]

Microscopic pathology is unusual because diagnosticians must search images too large to be seen entirely at one time. Microscopic examination involves moving around slides from area to area, examining different regions under varying magnification. This activity significantly slows the process of visual classification. The image is encoded and understood in many small pieces through a process of serial search. The authors' research methods exploited unique aspects of diagnostic microscopy to produce a rich corpus of videotape data for analysis of cognitive processes and errors.

Development of expertise can be delineated along dimensions including cognitive, sociocultural, and organizational axes. The authors chose information-processing as a theoretical framework, and therefore focused on finely granular cognitive analysis of tasks. Nevertheless, more general theoretical frameworks regarding acquisition of skills are directly relevant to this work. Dreyfus and Dreyfus[20] describe development of skilled human performance in five sequential stages: novice, advanced beginner, competent, proficient, and expert. Their stages reflect transitions in three general aspects of performance:

1. Switching from reliance on abstractions and rules to use of past experience (instances).

2. Evolution in situational perception. Experts perceive situations less as a set of equally weighted parts, and more as a whole in which parts vary in importance and relevance.

3. Changed perspective from detached observer to involved performer.

Use of this model has produced significant insights into the development of clinical competence in nursing[21] and raises intriguing questions regarding educational implications.

Research Questions Regarding Microscopic Diagnostic Pathology

1. How does diagnostic accuracy differ among three levels of expertise for a standardized case set? Does the assessment of case difficulty or diagnostic certainty parallel accuracy across all levels?

2. What are the cognitive processes underlying task performance, and how do these processes differ as expertise develops?

Table 1 ■

Case Materials

Case	Gold Standard Diagnosis	Description
1	Infiltrating ductal carcinoma	Focal lesion of poorly differentiated cancer adjacent to biopsy site and scar.
2	Ductal carcinoma in situ (DCIS)	Widespread solid and cribriform in-situ carcinoma present throughout majority of sample.
3	Infiltrating lobular carcinoma	Widespread classical type infiltrating lobular carcinoma. Scant adjacent normal tissue.
4	Lobular carcinoma in situ (LCIS)	Small focus of LCIS with retrograde extension in otherwise normal breast.
5	Fibroadenomas, sclerosing adenosis and intraductal papilloma	Multiple focal lesions, including sclerosing adenosis—a benign lesion that shares some visual features with cancer.
6	Paget's disease	Nipple with focal area of intra-epidermal Paget's disease. No underlying carcinoma.
7	Adenomyoepithelioma	Small circumscribed lesion with uniform features.
8	Atypical papilloma	Large lesion with numerous atypical features

3. What errors occur in the performance of this task, and what is the distribution of errors across levels of expertise?

4. What is the relative proportion of time spent in the different phases of this diagnostic task as a function of level of expertise?

Methods

Research Design

We performed a process-tracing, expert-novice study with level of expertise as the independent variable. Dependent variables included diagnostic accuracy; certainty and difficulty ratings; protocol process and error measures; and latencies. The study employed a case set of eight breast pathology slides (one per case) with pre-established gold standard diagnoses. Each of the 28 participants was randomly assigned to see four of eight possible cases, yielding a total of 112 subject-cases. The cases were distributed so that the number of cases for each case-level combination was roughly equivalent. Four of 112 protocols were not collected because of technical failure, resulting in a corpus of 108 protocols for study.

Case Materials

Cases were selected from the files of a single university hospital. Each case consisted of a single glass microscopic slide of tissue stained with hematoxylin and eosin and a 1–2 sentence case history. A second pathologist reviewed all cases and made an independent diagnosis in addition to that of the original pathologist. The study accepted only cases in which the two diagnoses agreed. The case set was designed to span multiple continua including diagnostic difficulty, size of lesion relative to size of tissue, typicality, and incidence of disease. "Controversial" or "borderline" cases were excluded as follows: (1) cases representing disorders for which no generally accepted consensus exists in the field regarding the criteria for diagnosis or no agreement regarding existence of the disorder as a separately classifiable entity and, (2) "borderline lesions" in which the histologic features of the specific case lie at the boundary between two possible diagnoses. Diagnoses included in the case set are shown in Table 1, along with a brief description of each case.

Participants

The study included 10 novices (third-year medical students who had recently completed the required second-year course material in pathology), 10 intermediates (second- and third-year residents in pathology, who had completed at least one year of surgical pathology), and 8 experts (board-certified pathologists, many with special expertise in breast pathology, with an average of 26.5 years of training and practice experience). All subjects were volunteers, recruited by a combination of e-mail, regular mail, and poster solicitations. Medical students and residents received a small honorarium for their participation. All medical students were from a single medical school (University of Pittsburgh). Pathology residents were from multiple residency training programs across the country; however, the majority of participants were from the University of Pittsburgh. Expert pathologists were from multiple hospitals within the University of Pittsburgh Medical Center system. The study design and use of human subjects were approved by the University Institutional Review Board.

Task

Participants first examined each slide without benefit of clinical history, talking out loud until they reached a diagnostic conclusion. Then they were given the brief clinical history and permitted to return to the slide to revise their opinion before issuing a final diagnosis. This procedure is similar to one previously used in cognitive studies of radiologists.[12] At the conclusion of each case, subjects rated the certainty of their diagnosis, and difficulty of the case, on a visual analog scale.

Data Collection

We collected think-aloud protocols as participants examined the case materials. Think-aloud protocols[22] are a standard technique of cognitive science, in which participants are asked to verbalize all of their thoughts without filtering them. With minimal coaching, most participants can reveal cognitive processes associated with task performance. Concurrently, and synchronized with the subjects' verbalizations, we videotaped the entire microscopic pathology session using a camera mounted on the subject's microscope. The video captured the visual data available to the participant and the magnification. A permanent audiovisual record of the diagnostic process was created and stored as digital video files on CD-ROM. Think-aloud protocols were also transcribed verbatim and segmented into individual protocol statements.

Coding Schemes

Investigators developed two independent coding schemes: one for the coding of cognitive processes and one for the coding of errors. Both coding schemes were derived from an initial analysis of 24/108 individual cases from 16 different subjects across all cases and levels of expertise. The theoretical background for coding cognitive processes has been previously described.[23] Briefly, operators are viewed as "information processes" that produce new states of knowledge by acting on existing states of knowledge. Individual protocol statements (segments) are encoded as (1) an operator or action that defines the process and (2) a list of arguments composed of descriptors and their values that encode the content or knowledge. The complete list of descriptors is defined with the action, but the selection of descriptors and the values that they take vary with each segment coding.

For process coding, an initial coding scheme was adapted from Hassebrock and Prietula[24] but modified during the iterative coding scheme development process. The development of the coding scheme was incremental and iterative: each of the 24 protocols was exhaustively coded, while building the set of operators and a description of the criteria required for coding that operator. The template of process codes was constructed in the Protocol Analyst's Workbench (PAW)—a Macintosh software package for protocol analysis.[25] PAW assists human coders by facilitating development of a consistent coding vocabulary and by providing domain-neutral methods for data entry and common protocol analysis tasks. Eleven precursor versions of the coding scheme preceded the final coding scheme. The final scheme consisted of 48 operators covering five general categories: (1) data examination, (2) data exploration and explanation, (3) data interpretation, (4) control processes, and (5) operational processes. Data examination included actions involving the selection and examination of visual and historical data; for example, *Identify-normal-structure, Identify-histopathologic-cue*, and *Compare-findings-from-multiple-locations*. Data exploration and explanation included higher-order abstractions, such as *Evaluate-certainty-finding, Evaluate-salience*, and *Associate-location*. Data interpretation included hypothesis generation and testing (e.g., *Statement-of-hypothesis* and *Confirm-hypothesis-with-present-finding*) as well as more basic actions that supported interpretation (*Recall-evidence-hypothesis-relationship*) and the evaluation of the output of data interpretation actions (*Evaluate-certainty-hypothesis*). Control processes included operators for more global evaluation of the progress or quality of one's own reasoning (meta-reasoning) and for diagnostic planning subsequent to the current task. Operational processes included verbalizations of motor and attentional processes related to the use of the microscope. Table 2 depicts a small fragment of protocol and the codes for actions and descriptors that were assigned to each segment. The complete set of operator codes and coding criteria are available on request from the first author.

Error coding involved a similar incremental, iterative approach through combined video and protocol analysis. All events that could be considered "diagnostic errors" were categorized. Error codes covered the case level (e.g., never finding the diagnostic area)

Table 2 ■

Fragment of Protocol Showing Process and Content Coding

			Content codes (Arguments)	
Line	Protocol segment	Process codes (Operators)	Descriptor	Value
137	"I have to figure out areas of obvious necrosis"	Set-goal-identify	Finding	areas of obvious necrosis
138	"No hemorrhage."	Note-absent-finding	Absent-Finding	hemorrhage
139	"I still think it is malignant."	Statement-of-hypothesis	Hypothesis # Hypothesis Category Hypothesis	4 benign/malignant malignant
140	"I still think it is breast."	Identify-anatomic-location	Location	breast
141	"I think..."	Not-coded	Reason	incomplete thought
142	"Could it be a lymph node that is replaced by something?"	Statement-of-hypothesis	Hypothesis # Hypothesis Category Hypothesis	5 general replaced lymph node
143	"Except for the fat, I don't see anything else."	Confirm-with-present-finding	Hypothesis # Hypothesis category Hypothesis Present-finding	5 General Replaced lymph node Fat

and the level of individual protocol statements (e.g., assigning an incorrect significance to a particular finding). Table 3 shows the complete set of error codes and example errors for each category.

Protocol Coding

The first author coded each segment of the 108 protocols and videotapes using the process and error coding schemes and the Protocol Analyst's Workbench (PAW). Text files were exported to Excel, SPSS, and Statview for subsequent analysis.

Measurements

Coding Diagnostic Accuracy. Diagnostic accuracy was coded (as correct or incorrect) before and after the clinical history was reviewed for both the specific diagnosis and the general diagnostic category. Each case had a gold-standard specific diagnosis and a general diagnostic category. For specific diagnoses, synonymous names were accepted; for general diagnostic classifications, true "parent" categories as well as "close misses" (alternative specific diagnoses deemed suboptimal) were accepted. For example, for a case in which the gold standard diagnosis was "infiltrating ductal carcinoma," a diagnosis of "invasive ductal carcinoma" was considered a correct specific diagnosis, whereas a diagnosis of "cancer" was considered correct for category but not for specific diagnosis.

Measures of Certainty and Difficulty. Ratings on visual analog scales were converted to an integer value between 0 and 10 for each case (using the nearest tenth of the scale as the value).

Process and Error Measures. Frequencies of all processes and all errors were determined for each case. Additional measures included operator sequences (the number of times a particular operator preceded or followed another operator), the number of unique diagnostic hypotheses, and whether the participant ever considered the gold-standard diagnosis during the course of the examination.

We performed both intra-rater and inter-rater reliability studies using a randomly selected set of 12 protocols (11%). For intra-rater reliability, the first author (RSC) recoded the same protocols after a 3-month wash-out period. For inter-rater reliabilities, we taught the coding scheme to a board-eligible pathologist (JS) who had no previous knowledge of this study and no prior experience with protocol analysis. The reliability coder was trained over three 2-hour sessions, during which he coded 15 protocols of increasing size and difficulty, discussing all discrepant codes with the primary coder (RSC). After training was complete, he coded 12 randomly selected protocols entirely on his own without any interaction, assistance, or feedback. We determined agreement by calculating the percentage of *individual operator codes* in which both coders assigned the same code. The intra-rater reliability was 89%, and the inter-rater reliability was 79% for the individual codes (not for the larger aggregated indices). Thus,

Table 3 ■

Errors

Errors	Description	Novice	Intermediate	Expert	Statistics			
Case Level	Errors coded as present or absent in each case	Number of cases/ Total (%)	Number of cases/ Total (%)	Number of cases/ Total (%)	Chi-Square	P Value	Pairwise Comparison	P Value
1	Lesion never brought under objective	8/30 (26.7 %)	1/28 (3.6 %)	0/23 (0 %)	13.25	0.0021	N, I	0.0059
							N, E	0.0006
							E, I	0.3189
2	Lesion traversed without recognition	7/30 (23.3 %)	0/28 (0 %)	0/23 (0 %)	8.11	0.0209	N, I	0.0056
							N, E	0.0058
							E, I	—
3	Error in identifying anatomic location	14/40 (35.0 %)	2/38 (5.3 %)	0/32 (0 %)	22.76	<0.0001	N, I	0.0008
							N, E	<0.0001
							E, I	0.1612

Segment Level	Errors counted for each case	Mean/ case	SD	Mean/ case	SD	Mean/ case	SD	F Value	P Value	Pairwise Comparison (Tukey HSD)	P Value
4	Incorrectly names normal structure	0.35	0.74	0.11	0.31	0	0	5.28	0.012	N, I	0.068
										N, E	0.013
										E, I	0.656
5	Incorrectly names histopathologic cue	0.93	1.4	0.76	1.08	.003	.18	7.05	0.004	N, I	0.669
										N, E	0.024
										E, I	0.004
6	Error in assigning significance, declarative knowledge failures	0.48	0.82	0.32	0.57	.003	.18	3.17	0.059	N, I	0.552
										N, E	0.048
										E, I	0.298

the reliabilities are lower bounds for the indices presented in this work.

Measuring Latencies. Using combined videotape-protocol analysis, we timed the following intervals for each case: (1) total time on task, (2) time to final identification of the anatomic location, (3) time to the first statement of a hypothesis, and (4) time to first statement of hypothesis ultimately accepted. For a subset of cases that contained a focal lesion, we also identified (5) the time of lesion identification. The criteria for identification of this event was that (1) the lesion was in the field of view and simultaneously (2) the participant made any statement identifying the area as different, such as "Aha," "That's something important," or "Oh, that's cancer there."

Analysis

Statistical analysis was performed using SPSS and SAS software. After determining that cases did not differ statistically with respect to dependent variables, we aggregated continuous dependent measures across cases to generate a mean for each participant and for each measure. One-way analyses of variance (ANOVA) included level of expertise as the factor and subject as the unit of analysis for protocol counts, times to particular protocol events, analog scale ratings, and continuous error measures. Tukey-HSD post-hoc tests determined whether differences were significant for pairwise comparisons. Chi-square tests were performed on accuracy measures and categorical error measures. For all tests, statistical significance was set at 0.05.

Results

Diagnostic Accuracy

Without additional clinical history, experts' mean diagnostic performance was highest at 78%; intermediates performed at 40% accuracy; and, novices at only 2.5% mean diagnostic accuracy (Table 4). Other measures of accuracy closely paralleled these findings. The addition of clinical history did not significantly improve specific or categoric accuracy among intermediates or experts. Categoric accuracy among novices did significantly increase with the addition of history, when compared with experts ($\chi^2 = 12.5$, n = 112, df = 2, p = 0.0004) and novices ($\chi^2 = 4.46$, n = 112, df = 2, p = 0.03). Post-hoc analyses showed significant differences between all pairs for all four measures of accuracy. Only novices had frequent difficulty in correctly identifying the anatomic location in the case.

Ratings of Certainty and Difficulty

Experts expressed the highest degree of certainty in their diagnoses and offered the lowest ratings of case

Table 4 ■

Accuracy

Measure	Novice	Intermediate	Expert	Chi-Square	P Value
Accuracy before clinical history					
Correct specific diagnoses/ Total (%)	1 / 40 (2.5 %)	16 / 40 (40%)	25 / 32 (78.1%)	26.32	<0.0001
Correct categoric diagnoses/ Total (%)	9 / 40 (22.5%)	21 / 40 (52.5%)	31 / 32 (96.9%)	23.32	<0.0001
Number diagnoses changed by history / Total (%)	20 / 40 (50%)	9 / 40 (22.5 %)	2 / 32 (6.3 %)	13.33	0.0013
Accuracy after clinical history					
Correct specific diagnoses/ Total (%)	4 / 40 (10 %)	18 / 40 (45 %)	26 / 32 (81.3 %)	40.39	<0.0001
Correct categoric diagnoses/ Total (%)	14 / 40 (35 %)	23 / 40 (57.5 %)	31 / 32 (96.9 %)	17.93	0.0001

difficulty (Table 5). Novices were least certain and thought cases were the hardest. Intermediates rated their certainty as lower in cases in which their diagnoses were inaccurate compared with cases where their diagnoses were accurate (t = –2.45, p = 0.019). Difficulty rating scores did not differ significantly between cases with accurate and inaccurate diagnoses (t = 0.45, p = 0.65).

Process Measures

Analysis of process code (operator) counts showed significant differences among groups. To summarize the results, we divide the findings by parent process: (1) data examination, (2) data exploration and explanation, (3) data interpretation, (4) control processes, and (5) operational processes. Wherever possible, we give examples of differences for individual operators. A complete analysis of all individual operator counts is beyond the scope of this publication and will be the subject of a subsequent communication.

1. **Data examination**. We divided data examination operators into three groups: (1) primarily involving visual identification, (2) primarily involving comparisons between or among visual features or areas of the slide, and (3) involving data examination related to the additional clinical history itself. Examples of data examination operators include *Identify-normal-structure*, *Identify-histopathologic-cue*, and *Compare-findings-from-multiple-locations*.

- *Visual identification*. Statements related to visual identification, taken together, constituted about 35% of all protocol statements (Table 6). Experts verbalized visual identifications less frequently than intermediates, and there was a trend toward fewer identification verbalizations when compared with the novice group. Furthermore, significantly different kinds of visual features were identified as a function of participants' levels of expertise. Whereas novices verbalized identification of normal structures most often, intermediates verbalized histopathologic cues most often.
- *Visual comparison*. Visual comparison statements constituted approximately 5% of protocol statements, and were most frequent in intermediate protocols (see Table 6).
- *Examination of history*. Statements related to the clinical history accounted for about 4% of all protocol statements and did not vary with subjects' expertise (see Table 6).

2. **Data exploration and explanation**. Approximately 7% of all protocol statements related to data exploration and explanation. Example operators include *Evaluate-certainty-finding*, *Evaluate-salience*, and *Associate-location*. These statements were most frequent among intermediates (see Table 6). Numerous differences were observed for individual operators. Novices and intermediates discussed the certainty of a finding more often than experts.

Table 5 ■

Difficulty and Certainty Ratings

	Novice		Intermediate		Expert			
Measure	Mean	SD	Mean	SD	Mean	SD	F Value	P Value
Difficulty Score	5.8	1.1	4.7	2.1	1.9	1.8	50.08	<0.001
Certainty Score	4.2	1.3	6.7	1.2	9.5	0.5	11.95	<0.001

Table 6 ■

Protocol Measures

	Novice		Intermediate		Expert		ANOVA			
Measure	Mean	SD	Mean	SD	Mean	SD	F Value	P Value	Pairwise Comparison (Tukey HSD)	P Value
Protocol lines	63.6	16.5	78.0	27.9	44.6	14.2	5.7	0.009	N, I	0.285
									N, E	0.154
									E, I	0.006
Data examination										
Identification	14.1	5.4	20.5	8.2	8.0	2.8	9.4	0.001	N, I	0.065
									N, E	0.110
									E, I	0.001
Comparison	2.2	1.2	3.3	1.5	0.93	0.66	8.7	0.001	N, I	0.132
									N, E	0.077
									E, I	0.001
History	1.4	0.5	1.7	0.7	1.2	0.8	1.7	0.210	N, I	0.511
									N, E	0.741
									E, I	0.191
Data exploration and explanation	2.6	1.6	4.5	2.4	2.0	1.6	4.4	0.023	N, I	0.081
									N, E	0.794
									E, I	0.027
Data interpretation	4.9	1.6	11.4	4.1	8.4	3.6	9.8	0.001	N, I	0.000
									N, E	0.084
									E, I	0.153
Control Processes	0.4	0.5	0.9	1.3	0.4	0.4	0.8	0.454	N, I	0.500
									N, E	1.000
									E, I	0.557
Operational processes	2.6	2.1	4.4	1.8	1.5	1.6	5.4	0.011	N, I	0.119
									N, E	0.408
									E, I	0.009
Goal-setting	0.4	0.4	1.7	1.3	1.8	2.0	3.3	0.052	N, I	0.094
									N, E	0.083
									E, I	0.981
Unique hypotheses	2.2	0.6	4.0	0.9	3.3	1.2	11.3	<0.001	N, I	<0.001
									N, E	0.031
									E, I	0.210

3. **Data interpretation**. Data interpretation accounted for approximately 20% of all statements. Example operators include *Statement-of-hypothesis*, *confirm-hypothesis-with-present-finding*, *Evaluate-certainty-hypothesis*, and *Recall-evidence-hypothesis-relationship*. Statements involving data interpretation were most frequent among intermediates (see Table 6).

4. **Control processes**. Statements related to control processes accounted for less than 2% of total statements and did not vary among groups (Table 6).

5. **Operational processes**. Operational statements related to the use of the microscope, such as a change of magnification, position, or attention to a particular area, and accounted for approximately 6.5% of the total. Intermediates had the highest frequency of these statements (see Table 6).

We analyzed *goal-setting statements* across all general categories; they accounted for fewer than 5% of statements. Goal setting statements explicitly (1) limit the diagnoses under consideration, (2) rule out a hypothesis, or (3) identify a particular finding that has not yet been found (see Table 6). Novices expressed fewer such statements, but this difference did not reach statistical significance.

Intermediates and experts considered significantly more unique *hypotheses* than novices (see Table 6). In cases involving an incorrect diagnosis, novices considered the correct diagnosis in less than 15% of the

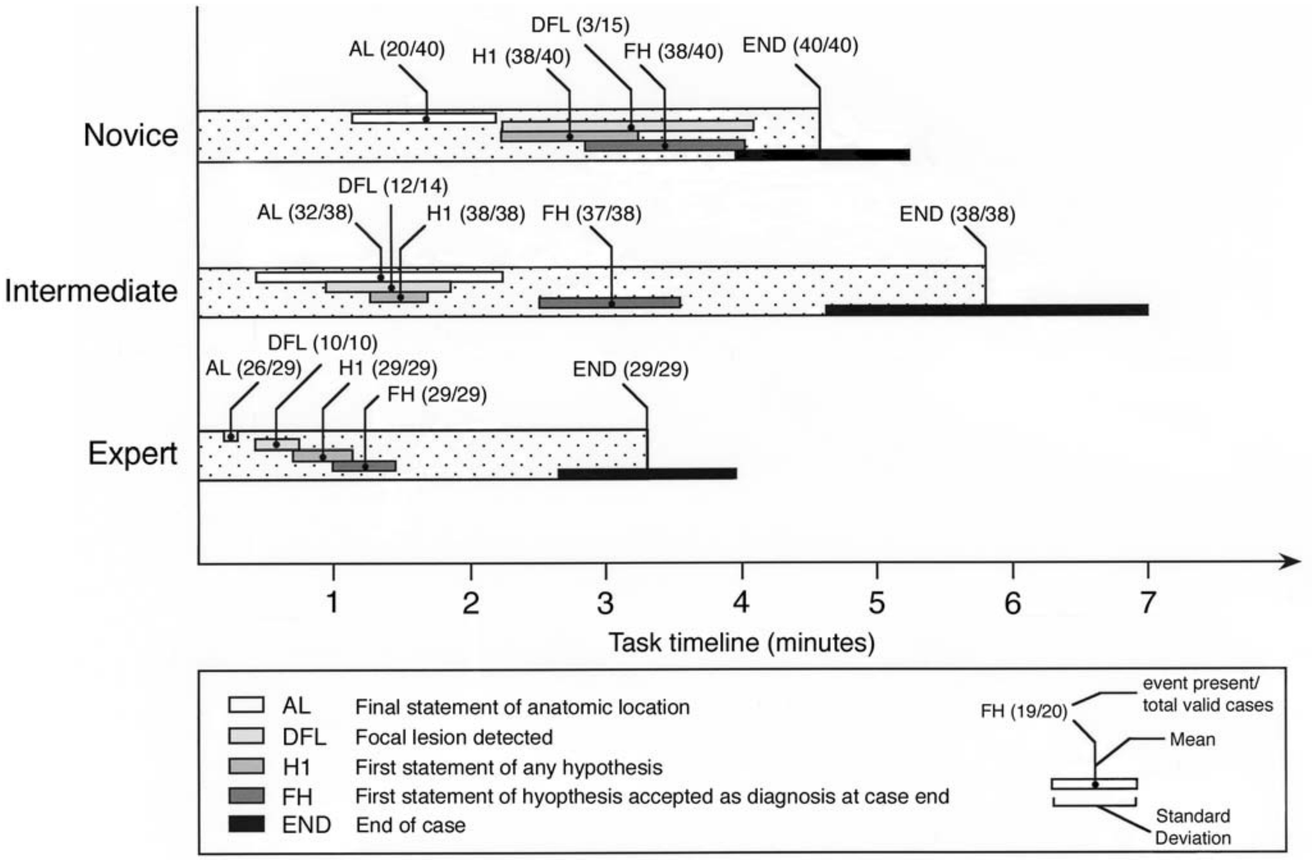

Figure 1. Task timeline depicting latencies and event counts.

cases, whereas experts and intermediates making errors considered the correct diagnosis approximately 45% of the time.

Latencies

Aggregated event latencies, depicted as timelines, appear in Figure 1. Ratios indicate the number of events detected per valid case. Events related to hypothesis formation occur early in experts, but later among intermediates and novices. Experts rapidly and uniformly identified the anatomic location and made early explicit statements about it. Novices often failed to identify any anatomic location. Experts detected focal lesions more rapidly than novices (Tukey HSD, p = 0.002), but experts and intermediates did not significantly differ in the time to lesion detection (Tukey HSD, p = 0.164).

Statement of a first hypothesis often marks a change in the protocol focus from exploring the slide to testing of possible diagnoses. Compared with novices, experts (Tukey HSD, p = <0.001) and intermediates (Tukey HSD, P=.001) made this change more rapidly. Experts verbalized the diagnosis that they will ultimately accept earlier compared with both intermediates (Tukey HSD, p = 0.002), and novices (Tukey HSD, p = <0.001). Close proximity between first hypothesis (H1) and final hypothesis (FH) among experts indicates that the first diagnosis was often accepted as final. Intermediates interjected more intermediate hypotheses between first and final hypotheses, taking longer to come to closure. Nevertheless, experts and intermediates spent more time between first statement of final hypothesis (FH) and the end of the case (END) than novices—typically, in testing their hypotheses, searching for confirming evidence, and ruling out alternatives.

Error Measures

We quantified six specific types of errors for each case (see Table 3): three errors at the case level (categorical) and three at the protocol segment level (continuous). Case-level errors included (1) failure to view the diagnostic area (search error), (2) traversal of diagnostic area without apparent recognition (gross perceptual error), and (3) failure to identify the anatomic location (an aggregate of (a) no attempt to identify the anatomic location, (b) location attempted but not completed, and (c) anatomic location misidentified). Segment-level errors included (4) participant incorrectly identified a normal structure, (5) participant incorrectly identified an abnormal histopathologic feature, and (6) participant incorrectly interpreted data (an aggregate of (a) wrong significance assigned to a feature and (b) incorrect recall of evidence-hypothesis relationships).

In all cases, error types 1–3 were significantly more frequent among novices (see Table 3)—and rare to nonexistent among intermediates and experts. Error types 4–6 decreased in frequency as level of expertise increased (see Table 3). Novices had greater difficulty in identifying normal structures; higher error rates in identifying histopathologic cues; and higher error rates for data interpretation. We examined participants' accuracy in the subset of cases in which focal lesions were correctly detected. On these cases, intermediates' diagnostic accuracy was 36% and experts' accuracy was 90% ($\chi^2 = 14.27$, $n = 54$, df = 1, p = <0.001). Although intermediates found the diagnostic area as quickly and as often as experts, they did not interpret these lesions as accurately once they found them.

Discussion

Significance

This study offers a first view of the features important to skilled performance in diagnostic microscopic pathology. Significant differences along the continuum of expertise occurred on all aspects of task performance—search, detection, feature identification, and data interpretation. Early in the development of expertise, the use of the microscope requires conscious attention and effort. Intermediates appear to apply explicit strategies in searching the slide, such as examining the entire slide at low power first, and selecting particular areas to revisit at higher power. Such explicit and conscious strategies were verbalized. Experts verbalized less often the operational aspects of changing power, position, and attention because they were more "automatic." Like Dreyfus' experts,[20] experienced pathologists behave as involved participants, using the microscope as direct extensions of their perceptual processes.

That intermediate participants accurately find lesions but do not classify them accurately suggests that diagnostic reasoning is not fully coupled to the search and detection process. The physical searching, perception, and attention required to find a lesion involve different skills than those needed to classify it. The important perceptual skill during the first process is recognition of "something that does not belong" and merits further investigation. Lesgold contended that perceptual learning occurs earlier than cognitive processes associated with inference.[12] Intermediates are in a unique position of recognizing important visual information, without having fully developed the ability to process it to diagnostic closure.

That intermediates are better at visual feature identification than novices suggests that increasing exposure to stimuli helps participants to recognize their symbolic meanings. Other work in nonvisual domains also found that intermediates verbalize more findings compared to both novices and experts.[26] Novices may not have the vocabulary or perceptual abilities to reduce complex visual cues in this fashion. Intermediates tend to identify and interpret individual features compared with experts, who arrive at the diagnosis sooner using a higher level, implicit "pattern-matching" approach. These findings are consistent with previous studies that identified compilation of elaborated to abridged networks as part of gaining expertise.[8,27] In addition, the Dreyfus model of expertise[20] recognizes evolution from perception of a set of "parts" to recognition of the "whole." "Pattern-matching" in visual diagnosis may reflect the composition or compilation[23,28] of processes that convert longer sequences of feature-identification and evidence-hypothesis matching into shorter sequences of nonverbalizable, higher-level "pattern-matching." Norman et al. asserted that accurate visual diagnosis of skin lesions is associated with increased speed, suggesting that rapid instance-based classification is central to skilled performance.[17] The authors suggest that pattern-detection is accurate, but strategies in which feature identification are prominent (Independent Cues Hypothesis) reflect failure of the first-line pattern recognition process.

An alternative, developmental perspective is that explicit feature identification necessarily precedes development of accurate, rapid, and implicit "pattern-matching" abilities. Pathologists' speed and accuracy gradually increase as features are learned and integrated into an increasing knowledge base. Feature identification and evidence-hypothesis matching represent a critical intermediate step during skill acquisition. Such a developmental sequence has important educational implications for this domain.

The question of reasoning "forward" versus "backward" has been debated with respect to diagnostic reasoning over time. Patel and colleagues concluded that novices' predominantly backward reasoning transitions to predominantly forward reasoning in experts.[9,10] Other data suggest that mixed strategies are common.[4,5] Analysis of verbal protocols at the level of individual protocol statements in this domain could address a number of interesting questions. How uniformly are "backward" and "forward" strategies applied? When do diagnosticians switch between them? What factors alter the balance between the two strategies? How does the structure and volume of knowledge affect the moment-to-moment struggle between competing approaches? Studying diagnostic microscopic pathology provides a critical advantage because the process of serial search (observing nonadjacent features one by one at different magnifications) slows the diagnostic process in a manner that facilitates observation. Additional studies in this field could produce a more finely granular model of skilled performance using think-aloud methods.

The interaction of forward and backward reasoning can be framed as part of a future global inquiry into diagnostic reasoning. How do various levels of search, perception, and reasoning interact? How does abstract declarative knowledge shape perceptual and search processes? How do search and perception influence development of strategies that also bear on the problem? How does an unconstrained search for cues become a goal-directed effort to find distinguishing features, and why? How do these processes change as knowledge structures are expanded and perceptual abilities are refined? What role do these strategies play in the generation of errors? And how can we exploit understanding of these processes to develop better methods for training diagnosticians?

Limitations of the Present Study

Several limitations potentially apply to this study. The time-intensive nature of information-processing methods limits the number of participants and tasks that can be included, reducing ability to generalize to other populations or subdomains. Second, participants were aware that they were being studied. The "Hawthorne effect" tendency for performance to improve when subjects know that they are being studied may have altered performance on our diagnostic tasks.[29] If all levels of subjects experience the Hawthorne effect equivalently, relative comparisons may still remain valid. Finally, the think-aloud methods impose their own set of potential limitations. It has been shown that subjects instructed to minimize effects of "mediated processes" such as filtering and self-explanation during think-aloud protocols do not behave significantly differently than when they do not verbalize during the same tasks.[22] However, verbalizations of processes are associated with increases in task time. Although relative comparisons hopefully remain valid, absolute latencies may not be accurate.

Implications for Future Microscopic Pathology Education Systems

One important motivation in performing this study was to use our findings to develop effective educational systems in microscopic diagnostic pathology. Study results indicate that physical search and detection are extremely important and poorly developed among novices. Traditional computer-assisted instruction using static images omits training in tasks related to search and detection. Study results suggest that explicit feature identification may be a critical intermediate strategy prior to reaching expert-level rapid "compiled" visual categorization. Systems should incorporate direct feedback on visual feature identification, separate from subsequent inferences.

Systems should use information about the student's ability to identify important visual features in selecting cases for presentation. Finally, different behaviors observed during hypothesis-testing point out that novices' lack of structured knowledge limits their ability to utilize the perceptual relationships they learn. A key advance in educational systems might couple training in visual aspects with presentation of appropriate concepts and relationships from pathology knowledge bases.

Conclusion

The authors' study of expertise in microscopic pathology offers early understanding of a highly complex visual diagnostic task. Results suggest that different cognitive skills contribute to expertise, providing a number of sources for diagnostic errors. Enumeration of differences in these skills and related errors along the continuum of expertise elucidates how these skills develop and provides an empirical foundation for computer-based pedagogy in microscopic diagnostic pathology.

References ■

1. van Ginneken AM and van der Lei J. Understanding differential disagreement in Pathology. Proc Annu Symp Comput Appl Med Care 1991:99–103.
2. Bartels PH. The diagnostic pattern in Histopathology. AJCP 1989; Suppl 1:S7–S13.
3. Crowley RS and Monaco V. Development of a model-tracing intelligent tutor in diagnostic pathology. Proc AMIA Symp 2001:805.
4. Azevedo R, Lajoie SP, Desaulniers M, Fleiszner DM, Bret PM. RadTutor. The theoretical and empirical basis for the design of a mammography interpretation tutor. In: du Boulay B and Mizoguchi R (eds): Frontiers in Artificial Intelligence and Application, Amsterdam, IOP Press; 1997;386–393.
5. Azevedo R, Lajoie SP. The cognitive basis for the design of a mammography interpretation tutor. Int J Artif Intell Educ 1998;9:32–44.
6. Lillehaug S-I, LaJoie SP. AI in medical education, another grand challenge for medical informatics. Artif Intell Med 1998;12:197–225.
7. Elstein AS, Shulman LS, Sprafka SA. Medical Problem Solving: An Analysis of Clinical Reasoning. Cambridge, MA, Harvard University Press, 1978.
8. Schmidt HG, Norman GR, Boshuizen HPA. A cognitive perspective on medical expertise: theory and implications. Acad Med. 1992;65:611–621.
9. Patel VL, Groen GJ. Knowledge-based solutions in medical reasoning. Cogn Sci 1986;10:91–116.
10. Patel et al. Primer on Cognition for Medical Informatics. J Am Med Inform Assoc 2001; 8:324–343.
11. Pople HE. Heuristic methods for imposing structure on ill-structured problems: the structuring of medical diagnostics. In Szolovits P (ed): Artificial Intelligence in Medicine. Boulder, CO, Westview Press, 1982, pp 119–190 (AAAS Symposium Series, no. 51).
12. Lesgold AM, Rubinson H, Feltovich P, Glaser R, Klopfer D, Wang Y. 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.
13. Kundel HL, Wright DJ. The influence of prior knowledge on visual search strategies during the viewing of chest radiographs. Radiology 1969;93:315–320.
14. Nodine CF, Kundel HL, Lauver SC, et al. Nature of expertise in searching mammograms for breast masses. Acad Radiol 1996;3:1000–1006.
15. Nodine CF, Kundel HL. The cognitive side of visual search in radiology. In O'Regan JK, Levy-Schoen A (eds): Eye Movements: From Psychology to Cognition. North Holland, Elsevier Science, 1987:572–582.
16. Kundel HL, Nodine CF. A visual concept shapes image perception. Radiology 1983;146:363–368.
17. Norman GR, Rosenthal DR, Brooks LR, Allen SW, Muzzin LJ. The development of expertise in dermatology. Arch Dermatol. 1989;125:1063–1068.
18. Norman GR, Brooks LR, Allen SW, Rosenthal D. Sources of observer variation in dermatologic diagnosis. Acad Med 1990; 65(suppl):S19–S20.
19. Regehr G, Cline J, Norman GF, Brooks L. Effect of processing strategy on diagnostic skill in dermatology. Acad Med 1994; 69(suppl):S34–S36.
20. Dreyfus HL, Dreyfus SE. Mind over machine: The Power of Human Intuition and Expertise in the Era of the Computer. New York, The Free Press, 1986.
21. Benner P. From Novice to Expert: Excellence and Power in Clinical Nursing Practice. Menlo Park, CA, Addison Wesley; 1984.
22. Ericsson KA, Simon HA. Protocol Analysis. Cambridge, MA, MIT Press, 1993.
23. Newell A, Simon HA. Human Problem Solving. Englewood Cliffs, NJ, Prentice Hall, 1972.
24. Hassebrock F, Prietula MJ. A protocol-based coding scheme for the analysis of medical reasoning. Int J Man-Machine Stud 1992;37:613–652.
25. Fisher CA. Protocol Analyst's Workbench: Design and Evaluation of Computer-Aided Protocol Analysis [dissertation]. Pittsburgh, PA, Carnegie Mellon University, 1991.
26. Arocha JF, Patel VL, Patel YC. Hypothesis generation and the coordination of theory and evidence in novice diagnostic reasoning. Med Decision Mak 1993;13:198–211.
27. Bordage G. Elaborated knowledge: A key to successful diagnostic thinking. Acad Med 1994; 68:883–885.
28. Anderson JA. Rules of the Mind. Hillsdale, NJ, Lawrence Erlbaum Associates, 1993.
29. Roethligsburger FJ Dickson WJ. Managemement and the Worker. Cambridge, MA, Harvard University Press, 1939.