Development of Visual Diagnostic Expertise in Pathology

Rebecca S. Crowley, MD[1,2], Gregory J. Naus MD[3], and Charles P. Friedman PhD[1]

[1]Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh PA
[2]Center for Pathology Informatics, University of Pittsburgh, Pittsburgh PA
[3] Magee-Women's Hospital, Pittsburgh PA

ABSTRACT

In this study, we use methods from information-processing to explore the visual diagnostic processes of novice, intermediate, and expert pathologists. Subjects were asked to examine and interpret a set of slides while we collected think-aloud verbal protocols and captured on digital video the actual visual data they examined from the microscope. We performed an in-depth combined video and protocol-based analysis of processes and errors occurring as the task was performed. Additionally, we collected measures of accuracy, certainty, and difficulty for all cases. Our preliminary analysis identified significant differences between groups in all three major aspects of this task: searching skills, perceptual skills and cognitive/reasoning skills. We describe the implications of our preliminary cognitive task analysis on the design of a developing intelligent educational system in Pathology.

INTRODUCTION

How do pathologists make diagnoses and how might this complex skill be taught most efficiently? Experienced pathologists often speak of particularly talented diagnosticians as "having a good eye," reflecting a commonly held belief among pathologists that diagnosis is best characterized as a kind of complex implicit pattern matching. Although pattern matching is certainly part of expert pathology diagnosis, we are interested in exploring additional features of expertise in the domain, including diagnostic reasoning and search strategies.

In this study we use methods from information-processing to compare the visual diagnostic processes of novice, intermediate, and expert pathologists. Our primary motivation in performing this study is to use our findings to inform the development of an intelligent computer-based educational system in Pathology. Pathology residencies typically last 5 years, during which approximately half of the training is devoted to acquisition of skills in diagnostic surgical pathology. Many trainees opt for subsequent sub-specialty fellowships. Long residencies and sub-specialty fellowships are needed because there are a vast number of patterns that must be recognized and because many patterns are infrequent, requiring an extended training interval to accrue sufficient exposure. In general, computer-based education in this domain could augment training by providing exposure to a large number of rare patterns, in a short focused interval. But unlike standard computer-based instructional systems, intelligent tutoring systems (ITS) have the distinct advantage that they can provide individualized coaching and feedback in a simulated, but realistic task environment. Substantial previous work on ITS has shown that development of successful systems is closely coupled with empirical research which is needed to (1) define the task to be tutored, (2) characterize the steps to expertise that will be *scaffolded* by the tutoring system, (3) determine the "rules of thinking" which form the basis of expertise in that domain, and (4) identify the important errors and misconceptions made by students. Following the example of others working in similar domains [1,2,3], our approach is to combine our own analyses of expertise in this domain with previous empirical and theoretical work from the fields of cognitive psychology, education and perception, in order to direct our design.

Our work draws from related research on expertise in internal medicine and radiology. Studies of medical cognition by Patel and colleagues, have shown that clinical problem solving differs among four identifiable levels of expertise: novices, intermediates, sub-experts, and experts[4,5,6]. Experts, with their extensive domain knowledge, filter irrelevant information and arrive at a diagnosis early in the process. Sub-experts, who hold general domain knowledge but more limited specialized knowledge arrive at a correct diagnosis later in the process. Intermediates, in the process of building a knowledge base, are characterized by attempts to apply developing knowledge, often with difficulty, to real world problems. Novices, who have little to no domain knowledge, represent problems in their most simplistic form. Using a variety of methods, other researchers have studied diagnostic expertise among radiologists examining chest X-Rays[7], and mammograms[1,2,8]. Describing the development of expertise in reading chest x-rays, Lesgold and colleagues [9] have shown that experts evoke a pertinent schema quickly, but tune these schemata flexibly, and thus can alter their representation when conflicting data are encountered.

Although the pathology diagnosis is often among the most important factors in determining patient prognosis and treatment options, there have been no prior studies focused on identifying the components of skill in this domain. As a medical diagnostic task, microscopic pathology is unusual because it requires the diagnostician to visually search an image, which cannot be seen in its totality at one time. With the aid of a microscope, the pathologist moves around a slide from area to area using objectives of different magnification to examine the tissue. This aspect of the task significantly slows the process of visual

1067—5027/01/$5.00

classification. Unlike many other visual classification tasks, the image is encoded and understood in many small pieces through a process of serial search. As she physically searches a slide, the pathologist focuses on areas of interest and ignores irrelevant detail, identifies important features and assigns significance to them, considers and tests hypotheses, recalls and rules out other alternatives, and eventually converges on the diagnosis that best fits the visual pattern encountered. In this study we begin to define how this complex set of processes combine during the development of expertise.

METHODS

This analysis includes fifteen subjects (5 novices, 5 intermediates, and 5 experts) selected from a total of thirty subjects who were recruited for participation. Each subject examined 4 cases of breast pathology.

Case materials: Cases were selected from the files of a single University Hospital. In addition to the report of the original pathologist, a second pathologist reviewed all cases, and made an independent diagnosis. Only cases in which the two diagnoses agreed were considered for inclusion. 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. Each subject saw four of eight possible breast pathology cases, two considered by our expert collaborator to be "easy" and two considered to be "moderate-difficult" for a junior resident.

Subjects: Novices were 3rd year medical students who had recently completed the required second-year course in Pathology including a one-month course in Reproductive Pathology. Intermediates were 2nd and 3rd year residents in Pathology, who had completed at least one year of surgical pathology and the equivalent of one rotation in breast and gynecologic pathology. Experts were practicing board-certified pathologists, with special expertise in Reproductive Pathology, and an average of 24.2 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.

Data collection: Subjects were instructed to give think-aloud protocols[10], and demonstrated this skill on a practice case before proceeding to the diagnostic test set. Think-aloud methods are a standard technique of cognitive science, and have been used to study tasks in a wide range of domains. Subjects are asked to verbalize all of their thoughts without filtering them as they perform a task. With minimal coaching, most subjects are able to provide a running stream of verbalizations revealing the cognitive processes associated with task performance.

We asked subjects to first examine each slide without benefit of clinical history, talking out loud until they reached a diagnostic conclusion. They were then given a brief clinical history indicating the patient's age and gender, anatomic site, procedure, and relevant clinical history, and permitted to return to the slide and revise their diagnosis before issuing a final diagnosis. Before proceeding to the next case, subjects were asked to rate certainty of their diagnosis and difficulty of the case on a 10 cm visual analog scale. Responses were measured to the nearest cm. Video feed of the entire session was captured from the microscope, synchronized with audio from think-aloud protocols and stored as digital video files on CD-ROM. An additional audiotape recording was made for transcription purposes. Think-aloud protocols were transcribed verbatim, and segmented into individual protocol statements.

Protocol Coding: Sixty individual cases were coded along two axes: First, each protocol statement was coded for operators (process) and for knowledge states (content). An initial coding scheme was adapted from Hassebrock and Prietula [11], but extensively modified during the iterative coding scheme development process. Twenty-four of 120 individual cases, from 16 different subjects across all levels of expertise, were used to develop the final coding scheme. The final scheme contained 56 operators representing aspects of (1) data examination, (2) data explanation, (3) data interpretation and hypothesis testing, and (4) control processes, such as meta-reasoning. Second, in correlation with the video record, cases were coded for errors in the search, perceptual, and reasoning aspects of the task (Table 3). Error codes included errors made at the case level (e.g. never finding the diagnostic area) and at the level of individual protocol statements (e.g. assigning an incorrect significance to a particular finding).

Data Analysis: Diagnostic accuracy was determined for final diagnosis before and after the clinical history was reviewed. In both cases, the diagnosis was coded as correct or incorrect for the specific diagnosis and the general diagnostic category by assessing agreement with a pre-determined list of correct specific and category diagnoses for each case. Transcripts were coded and analyzed with the Protocol Analyst's Workbench, a Macintosh software package for protocol coding, model and process tracing. One way Analyses of Variance (ANOVA) was performed on protocol counts, times to particular protocol events, analog scale ratings and measures of accuracy, with subject as the unit of analysis. Statistical significance was set at 0.05.

RESULTS

Results of our preliminary analyses point to four developmental sequences with important implications for the design of an educational system [12]. We first review some general descriptive measures off this task (Table 1) and then present evidence from our process and error analyses supporting our emerging model of skill acquisition (Tables 2 and 3).

Protocol process	Novice (n=5)		Intermediate (n=5)		Expert (n=5)		ANOVA	
	Mean	SD	Mean	SD	Mean	SD	F-value	P-value
Time to diagnosis (minutes)	4.0	0.9	5.9	3.1	3.7	1.0	1.86	<.05
Number of protocol statements	60	21	84	35	53	8	2.03	<.05
Certainty ratings; 10 = most certain	4.0	1.7	7.0	.9	9.3	0.5	26.32	<.001
Difficulty ratings; 10 = most difficult	5.5	1.0	4.3	2.2	2.5	1.6	4.38	<.05
Accuracy before clinical history								
% correct specific diagnoses	5	12	50	18	83	25	17.77	<.001
% correct diagnostic category	25	25	60	14	100	14	15.88	<.001
% Diagnoses altered after history	50	40	15	22	0	0	4.78	<.05
Accuracy after clinical history							13.29	<.001
% correct specific diagnoses	20	11	55	11	83	25		
% correct diagnostic category	35	29	65	14	100	14	9.58	<.01

Table 1. Task descriptive statistics

General descriptive measures (Table 1). As expected, accuracy was significantly different among groups. Only one correct diagnosis was made in the 20 cases seen by 5 medical students, and our video analysis showed that in this case the lesion was never actually observed under the microscope: the student had erroneously identified normal glands as infiltrating ductal carcinoma. Experts were highly accurate, identifying the correct diagnosis 83% of the time, and placing it into the correct broader category 100% of the time. Intermediate performance was approximately midway between novice and expert. On average, novices changed their diagnoses most frequently after reading the clinical history, and experts changed their diagnoses least frequently. Clinical history was critical for novices because they were often unable to identify the anatomic location of the lesion from the slide alone, and knowledge of the anatomic location typically constrains further decisions in this domain. Certainty ratings and difficulty ratings closely paralleled level of expertise: novices reported the highest difficulty and lowest certainty levels and experts reported the lowest difficulty and highest certainty levels. Time to diagnosis, and number of protocol statements did not significantly differ among groups.

Process and error analyses (Tables 2 and 3). We have identified four developmental sequences of interest, and outline below support for these sequences from both the process and error coding analyses.

1. A transition to accurate, and then automatic physical searching. Novices demonstrate frequent search errors, regularly missing the diagnostic area entirely (Table 3). In contrast, intermediates rarely erred in the physical search of the slide. They were able to find the diagnostic area, even if they were not able to interpret it accurately. This data suggests that (1) searching the slide is an important component of expertise that is poorly developed initially and (2) this skill develops relatively early during the training period. We also observed significant differences among groups, in verbalizations about the operational aspects of using a microscope, such as changing to a different power, or changing one's attention to another aspect or area of the slide (Table 2). Early in the development of expertise, the use of the microscope requires explicit attention and effort. Like the novices, intermediates remain quite aware of this aspect of the task. With experience, the explicit actions required to search a slide become automatic – the microscope becomes an extension of the expert. Thus our model suggests a transition from error-prone to accurate to efficient and automatic searching.

2. Evolution of visual efficiency. Novices and intermediates see an abundance of visual information. Our protocol analysis shows significant differences in the frequency of all Identify operators, by level of expertise (Table 2). On average, novices and intermediates explicitly identified more individual visual findings than experts. In our coding scheme, Identify operators are used to encode statements in which the subject verbalizes identification of any visual feature. Identify statements are sub-classified into codes for particular groups of features such as normal structures, histopathologic cues and descriptive cues. Novices mainly identified visual information descriptively (e.g. "big blue blobs") and noted normal structures, as would be expected with their limited knowledge. In contrast, intermediates identified more histopathologic cues (e.g. "central necrosis" and "sharply-punched-out-spaces"). We suspect that experts explicitly identified fewer discrete visual elements because they (1) have developed a highly efficient search strategy, (2) restrict their diagnostic options quickly, and (3) process visual information as a whole. Our error analysis shows significant differences among groups for errors related to the perceptual component of this task (Table 3). Although they appear to make fewer errors than their novice counterparts, intermediates at this stage of training continue to make errors in feature detection and identification.

Protocol process (# per case unless otherwise stated)	Novice(n=5)		Intermediate (n=5)		Expert (n=5)		ANOVA	
	Mean	SD	Mean	SD	Mean	SD	F-value	P-value
Verbalize operational aspects	3.05	1.52	4.30	1.67	1.50	1.41	4.16	<.05
Identify (aggregate)	20.15	3.28	24.45	8.11	12.43	6.84	4.51	<.05
Identify descriptive cue	2.65	1.90	0.35	0.38	0.00	0.00	8.21	<.01
Identify histopathologic cue	6.55	4.56	12.75	5.19	6.57	4.56	3.74	=.05
Knowledge-driven search for finding	0.65	0.55	3.00	1.07	2.57	1.76	5.15	<.05
Hypotheses considered (# unique per case)	1.32	.41	2.68	1.06	2.56	0.96	3.180	=.05
Protocol statement where final diagnosis first suggested (Statement number)	39.20	15.66	24.08	7.04	16.32	3.33	6.640	<.05

Table 2. Differences in protocol processes

3. A shift from reliance on explicit feature identification to rapid implicit pattern matching. We suggest that with developing expertise, the early phase of the diagnostic process (leading to hypothesis formation) is characterized by a shift from (1) reliance on identification of features to support explicit hypothesis formation, to (2) rapid implicit pattern matching. We identified significant differences in the protocol point at which the final diagnosis was first suggested (Table 2), supporting our contention of increased speed of hypothesis formation. The reliance on identification of features to trigger hypothesis formation is perhaps best demonstrated by the following example - one of many among the intermediate protocols. This excerpt is taken from an intermediate subject arriving at the correct diagnosis of lobular carcinoma. The visual pattern of lobular carcinoma is distinctive, and was almost instantaneously recognized by all experts and some intermediates. In contrast, this particular resident spent several minutes engaged in an unconstrained search of different areas of the tumor, before finding a particularly salient area:

> *I think now I begin to see the cellular elements in terms of how they are arranged. It's not forming a defined gland. It's more like...AHA!...I'm beginning to think I'm seeing indian filing. It's arranged linearly...That makes me think of Lobular Carcinoma.*

4. Development of goal - structured search and discrimination with expanding domain knowledge. The acquisition of domain knowledge is obviously a critical component of developing expertise. Relevant domain knowledge includes, among other things, knowledge of the diagnostic categories and knowledge about the kinds of visual features or criteria that argue for or against a particular diagnostic category. Not surprisingly, we detected significant differences among groups for number of hypotheses considered per case. (Table 2). On average, novices considered fewer hypotheses of any type than experts or intermediates. As domain knowledge expands, it changes the quality of the search and identification process. The developing expert can now explicitly set the goal to search for features that support or refute a diagnostic hypothesis. Our protocol analysis shows differences among groups for this knowledge-driven search for diagnostic features (Table 2). Preliminary analyses suggest that intermediates are using this operator more than novices, but about as frequently as experts. We suspect that intermediates and experts may be satisfying very different goals as they engage in knowledge-driven search. Informally, we observed that in expert protocols this operator was mainly used to either (1) rule out other less likely possibilities after rapid consideration and acceptance of another hypothesis or (2) identify a key feature that discriminates between two possible diagnoses with similar visual patterns. In contrast, knowledge-driven search among intermediates was more characteristically associated with an effort to support the leading diagnosis. In future analyses we will more formally test the hypothesis that use of these operators satisfies different goals depending on level of expertise.

Error category	Examples of specific errors included	Novice (n=5)		Intermediate (n=5)		Expert (n=5)		ANOVA	
		Mean # per case	SD	Mean # per case	SD	Mean # per case	SD	F-value	P-value
Search	Lesion entirely missed during search Magnification use error	0.30	0.27	.05	.11	0	0	4.43	<.05
Perceptual	Lesion traversed but not noticed Pathologic finding misidentified Normal finding misidentified	2.30	0.54	0.60	.38	0.05	0.12	45.86	<.0001
Reasoning	Assign wrong significance to finding Use wrong discriminator Insufficient evidence to accept/reject hypothesis	2.30	3.14	0.45	.33	0.15	0.12	2.17	.16

Table 3. Differences in protocol errors

DISCUSSION

In our preliminary analysis, we report on the identification of four developmental sequences that begin to describe the development of three basic groups of skills – physical search, visual feature recognition, and diagnostic reasoning – as they apply to this domain. Although we have attempted to explicitly separate them for the purposes of cognitive modeling, we recognize that each of these skills builds upon the others to enable expert performance. As search skills improve, it becomes possible to find the diagnostically relevant areas, and thus to practice the skills that are needed to reinforce accurate feature detection. As domain knowledge increases and accuracy in feature detection improves, one can begin to use the findings as evidence to support hypothesis formation, and ultimately more complex processes such as discrimination between diagnostic alternatives.

This work has the potential to contribute to research aimed at understanding the basic cognitive processes underlying diagnostic expertise, and to impact the development of instructional technologies in this domain. Our findings echo the work of investigators exploring diagnostic reasoning in other domains. For example, the increased identification we observed among novices and intermediates, when compared with experts, is reminiscent of previous work showing that intermediates utilize more information explicitly [4]. However, our findings suggest that intermediates are not unlike novices in this regard, except that they typically identify features in the language of the expert as opposed to more descriptive form.

In addition to contributing to our basic understanding of reasoning in this domain, our emerging cognitive model of expertise is of value in the design and development of computer-based instructional systems in Pathology. In parallel with our information-processing studies, we are developing a prototype Lisp-based model-tracing tutor in diagnostic pathology [12]. The cognitive model underlying our tutor is based on a set of generic production-rules for observing, identifying, and interpreting histopathologic findings, a frame-based knowledge representation encoding the diagnostic criteria and a frame-based problem representation of the slide. The production-rules were derived from our protocol analysis, and model the cognitive processes and errors that we observed in novice, expert and intermediate subjects. We are currently working on a virtual-microscope interface to the existing model that will simulate the authentic diagnostic context, enabling the tutor to scaffold novice performance in searching and interpreting a virtual slide.

REFERENCES

1. Azevedo R, Lajoie SP, Desaulniers M, Fleiszner DM, and Bret PM. RadTutor: The theoretical and empirical basis for the design of a mammography interpretation tutor. In: du Boulay B and Mizoguchi R, editors. Frontiers in Artificial Intelligence and Application, Amsterdam: IOP Press; 1997. p. 386-393.
2. Azevedo R, and Lajoie SP. The cognitive basis for the Design of a Mammography Interpretation Tutor. International Journal of Artificial Intelligence in Education 1998 9:32-44.
3. Lillehaug S-I and LaJoie SP. AI in medical education, another grand challenge for medical informatics. Artificial Intelligence in Medicine 1998 12: 197-225.
4. Arocha JF, Patel VL and Patel YC. Hypothesis generation and the coordination of theory and evidence in novice diagnostic reasoning. Medical Decision Making 1993 13: 198-211.
5. Patel VL and Groen GJ. Knowledge-based solution strategies in medical reasoning. Cognitive Science 1986 10:91-116.
6. Patel VL, Groen GJ, Arocha JF. Medical expertise as a function of task difficulty. Memory and Cognition 1990 18:394-406.
7. Kundel HL, Nodine CF, and Carmody D. Visual scanning, pattern recognition and decision-making in pulmonary nodule-detection. Investigative Radiology 1978 13:175-181.
8. Nodine, C.F., Kundel H.L., Lauver S.C., & Toto, L.C. Nature of Expertise in Searching Mammograms for Breast Masses. Academic Radiology 1996 3: 1000-1006.
9. Lesgold, A.M., Rubinson, H., Feltovich, P., Glaser, R., Klopfer, D., & Wang, Y. Expertise in a Complex Skill: Diagnosing x-ray Pictures. In Chi MTH., Glaser R, and Farr MJ, editors. *The nature of expertise.* Hillsdale, NJ: Lawrence Earlbaum Associates; 1988. p. 311-342
10. Ericsson K, and Simon H. Verbal reports as data. Psychological Review 1980 87: 215-250.
11. Hassebrock, F. and Prietula, M.J. A protocol-based coding scheme for the analysis of medical reasoning. International Journal of Man-Machine Studies 1992 37, 613-652.
12. Crowley RS and Monaco V. Development of a model-tracing intelligent tutor in diagnostic pathology. Proceedings of the AMIA Fall Symposium, 2001.

ACKNOWLEDGEMENTS

Rebecca Crowley is supported by National Library of Medicine Medical Informatics Training Grant number 5-T15-LM07059. Support for this project was provided by a grant from the University of Pittsburgh School of Medicine – Pathology Postdoctoral Research Training Program (PPRTP).