DOCUMENT RESUME

ED 415 753 HE 030 897

AUTHOR Papa, Frank; Shores, Jay
TITLE Expert Systems Based Clinical Assessment and Tutorial Project.
INSTITUTION Texas Coll. of Osteopathic Medicine, Fort Worth.
SPONS AGENCY Fund for the Improvement of Postsecondary Education (ED), Washington, DC.
PUB DATE 1992-08-30
NOTE 42p.
CONTRACT P116B90120-90
PUB TYPE Reports - Descriptive (141)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Artificial Intelligence; *Clinical Diagnosis; Diseases; Evaluation Methods; *Expert Systems; Higher Education; Instructional Effectiveness; Knowledge Level; *Medical Education; Problem Solving; Student Evaluation; Symptoms (Individual Disorders); Test Reliability; Test Validity; Tutorial Programs
IDENTIFIERS *Expert Novice Problem Solving; Texas College of Osteopathic Medicine

ABSTRACT

This project at the Texas College of Osteopathic Medicine (Fort Worth) evaluated the use of an artificial-intelligence-derived measure, "Knowledge-Based Inference Tool" (KBIT), as the basis for assessing medical students' diagnostic capabilities and designing instruction to improve diagnostic skills. The instrument was designed to address the problem that, in medicine, diagnostic expertise is problem-specific and appears to be more a factor of the student's knowledge base than cognitive skills. This study determined that the KBIT produced reliable and valid (based on comparisons of diagnostic accuracy of experts with those of novices) for four different problem areas: Weakness, Red Eye, Papulosquamous Disorders, and Elevated Creatinine. Additionally the study showed that two expert/KBIT-derived instructional approaches significantly improved the diagnostic accuracy of treatment student groups when compared to a control group and to students conventionally trained. After the executive summary and a project overview, this report describes the project's background and origins, its components and activities, and results. Attached is a related article titled "An Expert Program Shell Designed for Extracting Disease Prototypes' and Their Use as Models for Exploring the 'Strong Problem-Solving Methods' Employed in Clinical Reasoning" (F.J. Papa; S. Meyer). (DB)

Expert Systems Based Clinical Assessment and Tutorial Project

FIPSE FINAL REPORT/COVER SHEET

Grantee Organization: Texas College of Osteopathic Medicine
Department of Medical Education
3500 Camp Bowie
Fort Worth, TX 76107

Grant Number: P116B90120-90

Project Dates: Starting Date: 9-1-89
Ending Date: 8-30-92
Number of Months: 36

Project Director: Frank Papa &
Jay Shores
Department of Medical Education
Texas College of Osteopathic Medicine
Department of Medical Education
3500 Camp Bowie
Fort Worth, TX 76107

Grant Award:

Year 1	$87,726
Year 2	$87,726
Year 3	$87,463
Total	$262,915

SUMMARY

In medicine, diagnostic expertise is problem-specific. Furthermore, diagnostic expertise appears to be 'knowledge base' and not 'cognitive skills' dependent. Unfortunately, conceptual and logistical problems associated with current medical assessment methodologies make it difficult to obtain reliable and valid, problem-specific/knowledge base dependent measures of diagnostic capabilities.

In the early 1980's, one author (FJP) demonstrated that an artificial intelligence-derived tool could be used to acquire a problem-specific knowledge base from medically trained individuals. This tool made it possible to draw reliable, valid, and logistically feasible inferences about diagnostic capabilities in a given problem area. With funding from FIPSE, we set out to determine: 1) if the AI tool (called KBIT - Knowledge Based Inference Tool) could provided reliable and valid measures of diagnostic capabilities across a number of problem areas (i.e., is KBIT generalizable?), and 2) if KBIT derived instruction could result in improved diagnostic capabilities.

We have recently demonstrated KBIT's generalizability by producing reliable and valid (diagnostic accuracy of experts > novices) measures of diagnostic performance in each of four distinctly different problem areas (Weakness, Red Eye, Papulosquamous Disorders and Elevated Creatinine). KBIT's ability to produce psychometrically sound problem-specific, knowledge-based assessments of diagnostic capabilities made it possible to isolate and identify the knowledge base elements which characterize 'expertise'.

We subsequently demonstrated that two expert/KBIT derived instructional approaches significantly improved the diagnostic accuracy of treatment student groups when compared to a control group and a group of students trained with conventional instructional approaches. We believe that KBIT can serve as the foundation for the development of a new generation of psychometrically sound, 'intelligent' assessment and instructional tools.

Frank Papa DO, PhD
Jay Shores PhD
Department of Medical Education
Texas College of Osteopathic Medicine
3500 Camp Bowie
Fort Worth, TX 76107

Published Papers:

Papa FJ, Stone RC, Schumacher RE. Artificial Intelligence-Based Differential Diagnosis Assessment Procedures: Theoretical Advantages. In **Approaches to Assessing Clinical Competence**, Harden RM, Hart IR, Mulholland H eds. In Press.

Stone RC, Papa FJ, McIntosh, Aldrich DG. The Reliability of a Neural Network-Based Differential Diagnostic Assessment Instrument: A Pilot Study. In **Approaches to Assessing Clinical Competence**, Harden RM, Hart IR, Mulholland H eds. In Press.

Papa FJ, Young JI, Knezek G, Elieson B. Assessing Cognitive Strategies in Novice Learners. In **Proceedings of the Ninth International Conference on Technology and Education**, Estes N & Thomas M eds. Volume One, 213-216, 1992.

Papa FJ, Young JI, Knezek G, Bourdage R. A Differential Diagnostic Skills Assessment and Tutorial Tool. **Computers in Education**, 18, 45-50, 1992.

Papa FJ, Shores JH, Meyer S. The Use of a Pattern Recognition-based Prototype-Driven Research Tool to Study the Cognitive Constructs in Medical Decision Making. In **Current Developments in Assessing Clinical Competence**, Hart IR, Harden RM & Des Marchais J, eds. Ottawa, Can-Heal Publications, 210-213, 1992.

Shores JH, Papa FJ. The Effects of Experience, Pattern Recognition and Pattern Discrimination on Diagnostic Accuracy in Clinical Clerks. In **Current Developments in Assessing Clinical Competence**, Hart IR, Harden RM & Des Marchais J, eds. Ottawa, Can-Heal Publications, 317-320, 1992.

Papa FJ, Shores JH, Meyer, S. Effects of Pattern Matching, Pattern Discrimination and Experience in the Development of Diagnostic Expertise. **Academic Medicine**, 65,9,21-22, 1990.

Papa FJ, Shores J, Meyer S, O'Reilly B, Bourdage B. The Role of pattern Matching and Pattern Discrimination in Clinical Problem Solving. In **Teaching and Assessing Clinical Competence**, Bender W, Hiemstra RJ, Scherpbier, AJJA & Zwiestra RP, eds. Groningen, BoekWerk Publications, 322-331, 1990.

Papers in Review:

Papa FJ, Schumacker RE, Stone R, Young J. Pursuing Reliable and Meaningful Measures of Diagnostic Capabilities. **Teaching and Learning in Medicine.**

Papa FJ, Stone R. A Neural Network-based Differential Diagnosis Assessment Procedure. **Journal of Educational Computing Research.**

Papa FJ, Rusnak R, Meyer S. The Misdiagnosis of Acute Myocardial Infarction: Modeling a Cognitive Sciences-based Explanation for Diagnostic Errors. **Annals of Emergency Medicine.**

EXECUTIVE SUMMARY

Project Title:	Expert Systems Based Clinical Assessment and Tutorial Project
Grantee Organization:	Texas College of Osteopathic Medicine 3500 Camp Bowie Fort Worth, TX 76107
Project Directors:	Frank Papa DO, PhD Jay Shores PhD Department of Medical Education 817-735-2632

EXECUTIVE SUMMARY

Project Overview: In medicine, diagnostic expertise is problem-specific. Furthermore, diagnostic expertise appears to be 'knowledge base' and not 'cognitive skills' dependent. Unfortunately, conceptual and logistical problems associated with current medical assessment methodologies make it difficult to obtain reliable and valid, problem-specific/knowledge base dependent measures of diagnostic capabilities.

In the early 1980's, one author (FJP) demonstrated that an artificial intelligence-derived tool could be used to acquire a problem-specific knowledge base from medically trained individuals. This tool made it possible to draw reliable, valid, and logistically feasible inferences about diagnostic capabilities in a given problem area. With funding from FIPSE, we set out to determine: 1) if the AI tool (called KBIT - Knowledge Based Inference Tool) could provided reliable and valid measures of diagnostic capabilities across a number of problem areas (i.e., is KBIT generalizable?), and 2) if KBIT derived instruction could result in improved diagnostic capabilities.

We have recently demonstrated KBIT's generalizability by producing reliable and valid (diagnostic accuracy of experts > novices) measures of diagnostic performance in each of four distinctly different problem areas (Weakness, Red Eye, Papulosquamous Disorders and Elevated Creatinine). KBIT's ability to produce psychometrically sound problem-specific, knowledge-based assessments of diagnostic capabilities made it possible to isolate and identify the knowledge base elements which characterize 'expertise'.

We subsequently demonstrated that two expert/KBIT derived instructional approaches significantly improved the diagnostic accuracy of treatment student groups when compared to a control group and a group of students trained with conventional instructional approaches. We believe that KBIT can serve as the foundation for the development of a new generation of psychometrically sound, 'intelligent' assessment and instructional tools.

Purpose: For approximately the past forty years, 8-12% of all patients at autopsy are found to have died a premature death from missed diagnosis. Further, major missed illnesses with equivocal impact upon survival are present in another 20% of all autopsies. These findings reinforce the notion that the diagnostic process is an extremely difficult cognitive task.

Clearly, this less than optimal level of diagnostic performance must be derived in part, from deficiencies in the assessment methodologies and instructional approaches utilized during medical training. The purpose of this investigation was to determine if artificial intelligence-derived tools could improve diagnostic capabilities-related assessment methodologies and instructional interventions.

Backgrounds and Origins: In the early 1980's, one author (FJP) was intrigued with the notion that artificial intelligence (AI)-derived decision making tools could achieve levels of performance equal to experts in well defined problem areas across a variety of professions. Common to all of these AI tools was the fact that their performance depended almost exclusively upon the knowledge base with which it operated. Put simply, if the AI tool's knowledge base was acquired from an expert, then it's performance would be superior to the same AI tool operating with a knowledge base acquired from a less knowledgeable individual.

This author became interested in the notion that AI tools might serve as the basis of a new generation of assessment instruments in medical education. The advantages of AI-derived assessment instruments are as follows. One, performance assessments could be problem-specific and knowledge-based (given that expertise was problem-specific and knowledge-base dependent, it made sense to develop testing methodologies congruent with the nature of expertise). Two, once the subject's knowledge base for diagnosing a given problem were in the AI tool, their knowledge base could be challenged by hundreds to thousand of problem-specific test cases. This could solve the logistical problems which adversely affected the reliability of conventional methodologies (i.e., methodologies wherein only one to two test cases could be used rather than the tens of test cases needed to produce reliable problem-specific performance measures).

Three, if in fact the AI-based performance measures of experts were superior to novice performance measures, then this element of construct validity could legitimize the further use of AI tools as a means of exploring the knowledge base characteristics which distinguished experts and novices. Four, if the knowledge base elements which contributed to the experts superior diagnostic performance could be isolated via these tools, then these same critical knowledge base elements could be fashioned into instructional units designed to explicitly impart expertise. Explicitly structured, expert-derived instructional units could make it possible to improve the efficiency and effectiveness of the medical educational process and in-turn the novice's diagnostic accuracy.

Project Description: Originally, the authors had committed themselves to exploring the generalizability of KBIT as a assessment instrument via investigations involving six distinct medical problem domains. During the first two years of this project, the investigators were able to successfully demonstrate KBIT's generalizability (in terms of reliability and validity) over the first four problem domains (Weakness, Red Eye, Papulosquamous Disorders and Elevated Creatinine). These investigations involved over one hundred board certified experts in neurology, ophthalmology, dermatology and nephrology and over two hundred junior and senior medical students.

Given this level of success, the authors deferred activities related to the remaining two problem domains (both are now currently underway) and focused their primary efforts at identifying the knowledge base characteristics which distinguished experts from novices. Investigations into the four validated problem domains revealed that the experts knowledge bases not only achieved higher levels of diagnostic accuracy but that they also achieved higher pattern recognition measures (i.e., the pattern matching and pattern discrimination levels of experts > novices).

These pattern recognition measures reflected the decision making paradigm upon which the AI-derived assessment tool was based. That is, that diagnostic performance was a pattern recognition phenomena. The authors further hypothesized that this pattern recognition phenomena involved the dual processes of matching and discriminating a patient's constellation of signs and symptoms with internalized 'prototypical' disease patterns. The paradigm further purported that these disease prototypes were derived from the subject's knowledge base (i.e., a prototype was an abstracted, highly structured knowledge base (or disease template) consisting of ranked and weighted, disease-specific signs and symptoms).

The authors subsequently developed a means of extracting disease prototypes from expert knowledge bases. We developed various ways of extracting and describing expert-derived disease-specific knowledge bases and prototypes. We subsequently hypothesized that these expert-derived disease-specific descriptions could enable students to achieve higher levels of diagnostic accuracy than control (untrained) students and students trained via conventional medical educational approaches. The results of our pilot AI-derived instructional approaches demonstrated that explicitly structured problem and disease-specific knowledge bases, when imparted to novice medical students resulted in statistically superior levels of diagnostic accuracy than control or conventionally trained students.

Project Results: (See Table 1 & 2 on next page)

Summary and Conclusions: Moderate to highly reliable and valid, problem-specific assessments of diagnostic accuracy are logistically possible. KBIT-derived, explicitly structured problem and disease-specific knowledge base elements and prototypes (Table 2, groups 4 and 5), when imparted to novice medical students produce statistically superior levels of diagnostic accuracy than control (Table 2, group 1) or conventionally trained students (Table 2, group 2).

The availability of psychometrically sound, problem-specific measures of diagnostic capabilities and knowledge base acquisition techniques now makes it feasible to use KBIT as the foundation of a new generation of educationally sound, 'intelligent' assessment and instructional tools.

Project Results:

Table 1. Results of KBIT-based assessment instrument capabilities (generalizability) in terms of reliability and construct validity (student-t test) across multiple problem areas are as follows.

	Reliability Estimate (K-R 21) (students)	Student-t (one tailed) (Experts > Novices)
Weakness	.89	$p < .000$
Red-Eye	.95	$p < .011$
Papulosquamous Disorders	.71	$p < .001$
Elevated Creatinine	.96	$p < .000$

Table 2. Results of KBIT-derived instructional treatments (groups 3, 4 & 5) designed to produce diagnostic performance increases.

I. ANOVA: F ratio 5.8074 F Probability < .0006

II. Student-Newman-Keuls Procedure

	Groups				
	4	5	3	2	1
Groups					
4					
5					
3					
2	*	*			
1	*	*			

* Significantly different groups

Group 1 Untrained; Group 2 trained with conventional approaches
Groups 3, 4, & 5 trained with various KBIT-derived approaches.

BODY OF REPORT

Project Overview:

In medicine, diagnostic expertise is problem-specific and knowledge base dependent. Unfortunately, conceptual and logistical problems associated with current medical assessment methodologies preclude educators from achieving reliable and valid measures of problem-specific diagnostic capabilities. Furthermore, the autopsy literature clearly points out that instruction in the extremely difficult task of diagnosis must improve if physicians are to improve upon the persistently high levels of misdiagnosis.

While the medical education literature cries out for innovative assessment and instruction approaches as a means of solving these intractable problems, few if any researchers have investigated the potential of AI tools in these arenas. This project represents an effort primarily designed to use the latest knowledge and tools derived from the cognitive sciences to solve these long standing medical assessment and instructional problems.

For approximately twenty years, investigators in the field of artificial intelligence (AI) have been using a variety of computer-based tools to emulate and study human decision making. One area of fruitful activities has involved the use of an AI tools known as the Expert System (ES). These tools have characteristics, which on theoretical grounds, make them ideal as assessment instruments for measuring diagnostic capabilities. Some of these characteristics are as follows.

One, an ES is designed to solve problems or cases involving in a single problem area (i.e., to identify the most likely cause of a given problem from among a number of possible causes). Two, once the knowledge base needed to solve a given problem is acquired form a subject, the ES can use it to solve literally hundreds to thousands of problem cases. Three, an ES can usually solve numerous problem cases in literally seconds to minutes. Four, the criteria used to determine if a given problem case was solved correctly or incorrectly can be precisely and consistently applied to all problem cases in its case data bank.

Five, one inherent aspect of an ES is that a knowledge base acquired from an expert is likely to make the ES perform in a manner superior to an ES using a knowledge base acquired from an individual with intermediate or novice level knowledge. Six, the knowledge base of an ES can be investigated so as to determine why the knowledge base acquired from the expert performed in a manner superior to the knowledge base acquired from a novice.

From a medical educators perspective, the first five characteristics translate into the following assessment advantages. One, given that diagnostic expertise is

problem-specific and knowledge base (and not cognitive skills) dependent, then an ES-based assessment tool would appear to be an ideal means of acquiring, controlling for and assessing the diagnostic utility of an individual's knowledge base for solving (diagnosing) cases in a given problem area. Two, the logistical problems associated with the reliability of an assessment instrument are generally associated with the limited number of case challenges which an examine can physically pass through in a given time unit. Given that the ES contains the subject's knowledge base, it can literally be challenged with hundreds to thousands of problem-specific test cases and thereby achieve test reliability when supplied with the number of test cases sufficient for a given level of test reliability.

Three, this great number of test case challenges can be solved by an ES in essentially no more time than it took to acquire the subject's knowledge base to begin with thereby solving the logistical problems (lengthy test taking time) associated with traditional testing formats. Four, measurement error attributable to case presentation-related variance and examiner-related variance can be completely eliminated as the ES applies the same assessment criteria to all cases and for all subjects. Five, the long sought for, and infrequently if ever attained quest for 'construct validity' is likely to be achieved given the nature of the ES (i.e., the performance of expert-derived knowledge bases are likely to be superior to novice-derived performances).

The ability to acquire (with a fair degree of psychometric validity) the knowledge base elements which distinguish experts from novices makes it possible to extract and explicitly impart the precise knowledge base elements which could expedite the novice's transformation from novice to expert. Thus this ability, the sixth of the inherent advantages of ES tools, makes it conceivable to produce instructional interventions which could increase the efficiency and effectiveness with which the student's diagnostic abilities are developed.

In the early 1980's, one author (FJP) demonstrated that an artificial intelligence-derived tool could be used to acquire a problem-specific knowledge base from medically trained individuals and use it to draw reliable, valid, and logistically feasible inferences about their diagnostic capabilities in the problem area of Acute Chest Pain. With funding from FIPSE, we set out to determine: 1) if the AI tool (called KBIT - Knowledge Based Inference Tool) could provided reliable and valid measures of diagnostic capabilities across a number of problem areas (i.e., is KBIT generalizable?), and 2) if KBIT derived instruction could result in improved diagnostic capabilities.

We have recently demonstrated KBIT's generalizability by producing reliable and valid (diagnostic accuracy of experts > novices) measures of diagnostic performance in each of four distinctly different problem areas (Weakness, Red Eye, Papulosquamous Disorders and Elevated Creatinine). These studies

involved over one hundred board certified experts and two hundred medical students (novices).

KBIT's ability to produce psychometrically sound problem-specific, knowledge-based assessments of diagnostic capabilities made it possible to isolate and identify the knowledge base elements which characterize 'expertise'. We subsequently demonstrated that two expert/KBIT derived instructional approaches significantly improved the diagnostic accuracy of treatment student groups when compared to a control group and a group of students trained with conventional instructional approaches. We believe that KBIT can serve as the foundation for the development of a new generation of psychometrically sound, 'intelligent' assessment and instructional tools.

Purpose:

For approximately the past forty years, 8-12% of all patients at autopsy are found to have died a premature death from missed diagnosis. Further, major missed illnesses with equivocal impact upon survival are present in another 20% of all autopsies. These findings reinforce the notion that the diagnostic process is an extremely difficult cognitive task.

Clearly, this less than optimal level of diagnostic performance must be derived in part, from deficiencies in the assessment methodologies and instructional approaches utilized during medical training. The purpose of this investigation was to determine if artificial intelligence-derived tools could improve diagnostic capabilities-related assessment methodologies and instructional interventions.

Background and Origins:

The background regarding the theoretical advantages possible via the use of AI-derived assessment and instructional approaches has been briefly discussed. At this time we would like to discuss our emphasis upon the need to achieve construct validity.

It is very possible to use traditional assessment instruments and subsequently produce highly reliable test results via either Classical Test Theory or Generalizability Theory. However, high levels of test reliability do not mean that the test does reflect measures of the targeted construct. Therefore, all of our efforts have been designed to first attack head-on the issue of construct validity.

Specifically, we attempted to develop a means of measuring the 'diagnostic abilities' (in terms of diagnostic accuracy) of medically trained individuals. Therefore the construct under investigation and assessment was that of

diagnostic performance. Given that the medical education literature had determined that diagnostic capabilities were: 1) problem-specific and 2) knowledge base (and not cognitive skills) dependent, the investigators looked to develop an assessment instrument that involved the acquisition of a problem-specific knowledge base from test subjects.

The problem-specific/knowledge based nature of our testing methodology allows us to create a testing environment wherein all subjects are required to describe their knowledge base as related to the same pre-defined number of common/important diseases for the given problem area and the common/important signs and symptoms used to diagnose these same diseases. Thus the investigators have created a perfectly even playing field wherein all subjects must work within the same problem solving context. Therefore, any extraneous or hidden advantages which the expert may have or deficiencies of the novice are eliminated.

Subsequently, any differences in the AI-tools diagnostic performance must be related to differences in the expert or novice groups knowledge base. The boundaries of this knowledge base are explicitly delineated via the use of pre-defined problem area test boundaries (diseases and signs/symptoms). By demonstrating construct validity (i.e., that the diagnostic performance of experts is greater that the diagnostic performance of novices) the investigators can say that these differences can only be due to differences in their knowledge base as related to the specific problem at hand and as defined by the problem space boundaries.

The investigators subsequently felt less compelled to pursue very high levels of test reliability (> .80) in these pilot tests. It is important however, to keep in mind that via these AI tools, all that would be needed to achieve the needed level of test reliability in any given problem area (given that construct validity was demonstrated) is to acquire the number of test cases sufficient to attain a given reliability level and simply add them to the test case data bank. In reviewing the results of our investigations note that three of the problem areas achieved reliability estimates (K-R 21) of .89 to .96.

Project Description: This project had essentially two separate components. The first involved the determination of the generalizability of KBIT as a reliable and valid assessment instrument. To determine this we initially set out to investigate the reliability and validity parameters derived from studies involving six separate problem areas. The process of developing a problem-specific assessment instrument required that we first define the boundaries for a given problem area. This required in-turn that we identify the common/important diseases likely to cause a given problem and the common/important signs and symptoms that should be gathered in order to determine the cause of the given problem.

In defining these problem area boundary condition, we generally utilized the expertise of two experts in a given specialty. For example in the problem area of Red-Eye we meet with two board certified Ophthalmologists over three separate occasions with our KBIT tool. During each session we would refine the number of diseases and signs/symptoms to be included in the problem area. We used KBIT to help the physicians focus in on what represented the essence of the essential issues related to solving the given problem. We termed this component of the investigation Problem Space Boundaries Definition and utilized what we have termed Knowledge Engineering techniques to help the expert consultants to gradually refine the boundaries of the problem area.

Once we felt comfortable with the boundaries for a given problem area, we developed a questionnaire which allowed us to acquire the needed knowledge base form our targeted groups of students and novices. The knowledge base which we needed consisted of each subject's knowledge of the 'relationships' between each of the diseases in the problem area and the signs/symptoms included. These relationships can be viewed as representing the individual's understanding of the percentage of time a patient with a given disease was likely to have a given finding. The cognitive sciences literature refers to this type of knowledge as feature frequency estimates. The probability literature refers to this knowledge as conditional probability estimates.

We would generally send out our questionnaire to at least 100 board certified specialists per given problem area. We anticipated, and generally received a 25-35% response rate (with the exception of Elevated Creatinine which produced a poor response rate of less than 10%). We were able to obtain feature frequency estimates form students while on different clinical rotations via the cooperation of the Departments of Internal Medicine and Family Practice. We usually obtained questionnaires from 60 to 80 medical students per problem area.

All feature frequency estimates were entered into the KBIT software (the PI's own design (FJP)). Criteria test cases which were used to challenge the diagnostic accuracy of each subject were gathered via the cooperation of the specialists used to create the problem area's test boundaries. Generally we wanted to accumulate approximately 100 cases per problem area. This proved to be the most difficult aspect of the investigation as the consultants were not enthusiastic to collect test cases data in the specific manner as outlined in the problem area. Nonetheless, with much coaxing and persistence, we were generally able to acquire the number of test cases sufficient to produce highly reliable measures for each of the problem areas.

Much of the true research in this project involved the various ways in which the knowledge bases could be manipulated. The real objective of our deepest levels of research involved gaining new insights into how experts structured

their knowledge base and how these knowledge base structures supported the experts achieving their higher levels of diagnostic accuracy.

Ultimately we were able to determine that 'disease prototypes', that is abstracted representations of ranked and weighted signs and symptoms for a given disease appeared to be a parsimonious mechanism for storing and conveying 'expertise'. These prototypes proved to be the most efficient and effective means of conveying the 'hidden' knowledge of the expert to the novice.

Project Results:

Table 1. Results of KBIT-based assessment instrument capabilities (generalizability) in terms of reliability and construct validity (student-t test) across multiple problem areas are as follows.

	Reliability Estimate (K-R 21) (students)	Student-t (one tailed) (Experts > Novices)
Weakness	.89	$p < .000$
Red-Eye	.95	$p < .011$
Papulosquamous Disorders	.71	$p < .001$
Elevated Creatinine	.96	$p < .000$

Table 2. Results of KBIT-derived instructional treatments (groups 3, 4 & 5) designed to produce diagnostic performance increases.

I. ANOVA: F ratio 5.8074 F Probability < .0006

II. Student-Newman-Keuls Procedure

	Groups				
	4	5	3	2	1
Groups					
4					
5					
3					
2	*	*			
1	*	*			

* Significantly different groups
Group 1 Untrained; Group 2 trained with conventional approaches
Groups 3, 4, & 5 trained with various KBIT-derived approaches.

The results of this project are very encouraging. We have had a great deal of success in presenting and publishing our results and have at least as many potential papers and presentations yet to deliver. Within the context of our own institution the Dean has given us support in the establishment of problem-specific assessment instruments in each of the core clinical rotations.

We have received acknowledgment from the Association of American Medical Colleges/Research in Medical Education subgroup (i.e., the awarding of the Thomas Hale Hamm "New Investigator" award) and have been increasing asked by American and European medical school faculty members to provide additional unpublished information regarding the nature and scope of our AI-related activities. We believe that during the next five years, a small but significant number of educators interested in the use of AI assessment and instructional activities and approaches will come foreword.

Finally, we also believe that the medical education assessment establishment will balk at the widespread use of this assessment technology primarily because they have no to very little knowledge or understanding of the concepts surrounding artificial intelligence techniques. However, perhaps as the first decade of the twenty-first century ends there will be a number of medical training institutions using these AI-derived tools and techniques in an effort to truly prepare physicians for medical practice in the twenty first century.

Summary and Conclusions: Moderate to highly reliable and valid, problem-specific assessments of diagnostic accuracy are logistically possible. KBIT-derived, explicitly structured problem and disease-specific knowledge base elements and prototypes (see Results Table 2, treatment groups 4 and 5), when imparted to novice medical students produce statistically superior levels of diagnostic accuracy than control (see Results Table 2, group 1) or conventionally trained students (see Results Table 2, group 2).

Work in the two remaining problem areas is underway. One area involves the revisiting of the problem of Acute Chest Pain. In this investigation we intend to determine the degree to which this assessment instrument is capable of making fine discriminations between subjects. We have already acquired knowledge bases from a number of residents in training in the area of Emergency Medicine. Our preliminary results suggest that KBIT can in fact draw fine levels of discrimination from subject's with varying degrees of expertise (i.e., residents in their first few months to three years of residency training).

The second problem area involves the problem of Polyarticular Joint Pain. In constructing this problem area we have utilized a more sophisticated approach to the construction of the problem space boundaries. That is we are interested

in maximizing the amount of discrimination possible in terms of disease-specific diagnostic abilities per subject. We hope to begin the data collection for this problem area in January, 1993.

The availability of psychometrically sound, problem-specific measures of diagnostic capabilities and knowledge base acquisition techniques now makes it feasible to use KBIT as the foundation of a new generation of educationally sound, 'intelligent' assessment and instructional tools. We caution investigators in this area however, to play increasing attention to the care needed to produce efficient and effective problem space boundaries. That is, if the problem space definitions do not allow for the amount of discrimination needed to support the drawing of distinctions between experts and novices, and now more importantly to us, between one disease and another, then the results are likely to be disappointing. Much work needs to be done in this area, an area which we have called 'test construction knowledge engineering'.

The authors have recently submitted a FIPSE proposal designed to take the additional steps necessary to develop an "intelligent" assessment and instructional tool. We hope that this report substantiates the merit in further supporting this line of investigation. Clearly, the PI of this project is committed to continuing this line of investigation. Evidence of this is derived not only from the number of publications and presentations related to this project but also by the completion of post graduate training (Ph.D.) in the area of Computer Education and Cognitive Systems at the University of North Texas. Further evaluations of the KBIT system will continue.

Appendices: FIPSE assistance was in general very adequate. We were especially appreciative of the support of Saundra Newkirk.

In terms of reviewing future proposals, the investigators suggest that reviewers continue to place heavy emphasis upon the ability of AI tools to demonstrate elements of construct validity. Clearly these new tools can produce a tremendous amount of instructional material. The real question is whether any of this new material (or the AI-derived instructional approach) can produce efficient and effective changes in performance.

32

An expert program shell designed for extracting "disease prototypes" and their use as models for exploring the "strong problem-solving methods" employed in clinical reasoning

FJ Papa, Associate Professor, Division of Emergency Medicine, S Meyer, Department of Medical Education, Texas College of Osteopathic Medicine, Camp Bowie at Montgomery, Fort Worth, Texas, U.S.A.

Many contemporary medical educators focus on the perceived importance of the hypothetico-deductive process as the critical element of clinical reasoning. Recent evidence however, suggests that expert/novice reasoning performance differences may rest not so much upon the "process" but rather on the breadth, depth and richness of the "content" or knowledge base employed in problem resolution. A method for the explicit extraction and comparison of expert and novice knowledge bases must be developed, and their commonalities and distinctions compared and studied, before we can efficiently and effectively impart the critical elements of the knowledge base which characterize higher performance levels. Developments in the field of computer-based decision-making, particularly those software applications known as expert program shells, appear to be a viable tool for enabling the extraction and study of the critical knowledge base elements, which distinguishes the expert from the novice problem-solver. This presentation will report on our initial experience in applying an expert program shell in the study of clinical reasoning.

Introduction

The works of Elstein and Barrows characterize medical education's early formal attempts to study and define the nature of clinical reasoning. A primary purpose of these efforts was to identify the elements which enabled experts to achieve their higher levels of diagnostic accuracy and distinguished expert performance from that of the non-expert and novice clinician. Once these elements were captured, it was believed that the medical educational curriculum could be deigned to impart these critical elements to students earlier, more efficiently and effectively, and thereby ensure the production of more astute clinicians.

Newell (1973) has characterized the reasoning or problem-solving process as being comprised of two broad methodologies: weak or general heuristics and strong or specific techniques. The weak or general heuristics can be used in the initial problem formulation phase of problem-solving. They appear to be applicable to almost any problem domain but do not seem to lend themselves to expedite the phase of accurate problem resolution in the final stages of problem-solving. The strong techniques appear to be heavily content specific. As such, they appear to be designed for use within a particular problem domain. Furthermore, they appear to function as efficient and effective decision-making aids (perhaps are even most sensitive and specific) when applied in the later stages of problem-solving i.e., problem resolution.

Recently, Groen and Patel pointed out that medical educators appear in large to erroneously interpret the clinical reasoning literature, and emphasize the general or hypothetico-deductive process as the primary element which characterizes expert/novice performance distinctions. Simon and Hayes, and Elstein have pointed out that, in addition to an understanding of general problem-solving processes, the knowledge base is no less an important element for successful problem resolution. Others such as Greeno and Norman have however posited that knowledge specific to a particular problem domain is more important for correct problem resolution than the understanding and application of more generalized clinical reasoning processes.

While an understanding of both elements of clinical reasoning are important, traditional approaches have not readily aided medical educators to clearly identify the particulars of either the process or the content which lead to better clinical performance. Furthermore,

as Groen and Patel state, the real issue is not the primacy of general (weak) vs specific (strong) methods, but rather to identify when strong and weak methods are actually being used and how process relates to the doctor's knowledge base. Until the advent of computer-based decision-making tools, research into the specifics of expert/novice process and knowledge base distinctions and their interactions was at best a difficult task.

Computer-based expert shell and applications programs

In recent years there have been advances in the field of computer-based artificial intelligence which have significantly enhanced our understanding of the particulars of the fundamental elements of reasoning. One particularly important tool has been the development and use of a variety of "expert program shells" for extracting and eventually approximating the reasoning and decision-making accuracy of experts. While the specific means by which these expert program shells interact with an expert may vary, currently these shells' critical interaction with the expert is the extraction of the content or "knowledge base" which the expert utilizes in solving a given problem.

Once a problem-oriented knowledge base is extracted, the shell must combine the expert's knowledge base with the shell's inference engine (an inference engine is a software tool which draws upon a knowledge base in such a way as to draw a conclusion or infer a solution when challenged with a problem case). Together the knowledge base and inference engine form what is known as an "expert (applications) program". This "expert program" can be subsequently studied, tested and refined in an ongoing effort to both understand and approach the decision-making accuracy of the expert from whom the problem-oriented knowledge base was extracted.

With the advent of expert shells and expert applications programs, the authors considered the implications and potential applications for which these computer-based tools could be used. The following will discuss the basis upon which an expert system (both shell and applications program) was developed, and the preliminary results, impressions and conclusions drawn from work with this system.

Theoretical basis underlying the development of the described expert system

Information processing theory

Without much elaboration, the theoretical basis of the instrument which will be described shortly is rooted in the information-processing theory of Newell and Simon. Two fundamental concepts of this theory are those of "Task Environment" and "Problem Space". The task environment is defined as the structure of 1) facts (features and categories), 2) their inter-relationships and 3) concepts that define the elements of a problem. The problem space is the problem-solver's mental representation of the task environment.

1. Facts and relationships

With the information processing theory as a basis, the expert program shell was designed to allow the user (either student or expert) to describe and input aspects of his mental representation of the task environment (facts, relationships and concepts) into the instrument's software. Facts are represented in the software in both the forms of 1) historical or physical finding (features) and 2) the disease differentials (categories) which should be searched for and considered when solving cases in the given problem domain (e.g. "crushing" pain may be a feature and "myocardial infarction" a disease within the problem space known as "Chest Pain"). The relationships are the student's estimates of the frequency with which a given feature is likely to be present when a given disease is the etiology for cases likely to present within the problem domain (e.g. 80% of patients presenting with chest pain who have as their cause, myocardial infarction, will describe their pain as "crushing").

2. Concepts

It is the student's or expert's concepts which drive the individual's capacity to make estimates of each feature's frequency of association with each disease under consideration. These concepts cannot be directly extracted from the expert or student. However, the expert shell and applications system was designed to derive from the facts and relationships a machine-based representation of the expert's or student's concepts.

Within this expert system, a special tool known as an "inference engine" was designed to take the facts and relationships and from

them develop a machine-based pattern or "disease prototype" for each disease in the problem space. In essence, the individual's concepts (the qualities characteristic of, i.e. the features and their estimated frequency, each particular disease contained within the problem space) are now approximated or represented by these computer-derived "disease prototypes". Taken together, these now computer-based facts, interrelationships and disease prototypes represent the student's or expert's problem space or diagnostic paradigm. We refer to the student's or expert's computer-based end products as Problem-Oriented Diagnostic Protocol (PODP). They function (to varying degrees of success) as "expert programs".

The information processing theory applied: an expert program

With these disease prototypes in the inference engine's memory, the software is capable of taking input, either from a data bank containing a large number and variety of problem-related "criteria cases" or actual clinical case encounters. The machine-based disease prototypes are then compared with the data for each criteria case and an attempted match is made between each disease prototype and the case data. The machine then infers a "diagnosis" for the case based upon the machine's disease prototype which best matches the case data. The software then compares its diagnosis for the case against the diagnosis for the same case as originally determined by the faculty member who entered the case. In this manner, all cases in the data bank are used to challenge the student's or expert's PODP (and disease prototypes within the inference machine).

Guided by the use of these criteria cases, the PODPs which achieve higher levels of diagnostic accuracy are identified. The disease prototypes which are contained within these more accurate PODPs can then be further analyzed in an attempt to determine the commonalities which underlie accurate disease prototypes. In a similar fashion, weaker PODPs and their associated disease prototypes can be analyzed. Finally, both accurate and inaccurate disease prototypes can be compared and analyzed to help determine the elements which account for the performance distinctions.

General methods

Subjects

Thirty-two consecutive students who were taking a non-elective one month rotation on the Emergency Medicine service at the authors' institution were entered into the study.

Materials

Each student was provided with an IBM or IBM compatible personal computer, which they used to run the software provided. The software required 256K of memory and was operated by MS-DOS and an 8088 CPU.

Design

The authors determined that the problem of "Chest Pain" would be an appropriate pilot study for students to engage in while on rotation in Emergency Medicine. Nine disease categories were chosen for the students to include as the differentials in their Chest Pain PODP. These differentials were chosen because they all were either common and or important causes of "Chest Pain". These differentials were: Myocardial Infarction, Myocardial Angina/Ischemia, Pericarditis, Pneumonia, Pulmonary Embolus, Pneumothorax, Dissecting Thoracic Aortic Aneurism, Musculoskeletal Disorders and Esophagitis/Reflux Gastritis. During a one year pilot prior to this study, students were asked to develop Chest Pain PODPs which were based on these same nine disease categories, but were not predefined in terms of the historical or physical findings which they could use in their protocol. Following this one year pilot study, it was determined that the students used a total of sixty-seven different features in various combinations in compiling these PODPs. Thus, these sixty-seven findings served as a list from which the study group could choose any and all 67 findings for final inclusion into their own individual PODP.

Procedure

Prior to their arrival on the Emergency Medicine rotation, all students were given written notice that they would be required to develop a PODP for the problem of Chest Pain during the first three hours of their ER rotation. They were given a list of the nine disease categories and were requested to come to this first session prepared to develop a PODP based on their experience, knowledge and th

review of at least three current Chest Pain articles of their own choosing. Following a 60 minute introduction to the software, the computer and the specifics of the task which they were to perform (i.e. selection of features to be included, and estimations of each chosen feature's frequency in relation to each of the nine differentials), they were given 75 to 90 minutes to complete their own Chest Pain PODP. In a similar manner, three experts were asked to develop their own PODP for the problem of "Chest Pain".

At the completion of the study, each student's Chest Pain PODP was assessed in terms of their performance as measured by the PODP's overall level of diagnostic accuracy and disease-related sensitivity and specificity against a data bank of 72 criteria cases. Because of the small number of experts, a statistical analysis of the significance of the difference between expert/novice overall performance, and disease-related sensitivity and specificity will not be presented. However, the results derived from the preliminary expert sample suggest that there is a consistent trend and potentially significant difference in their parameters when compared with the students. The authors will report the results of these expert/novice comparisons when a larger number of experts have been included into the expert group.

Preliminary results

In our preliminary studies dealing with the problem of "Chest Pain", the initial trend suggests that expert PODPs and their associated disease prototypes achieved higher levels of diagnostic accuracy than the students. As previously stated, the small number of experts preclude us from reporting any specific comments regarding the experts' performance at this time.

Figure 1 contains the results compiled from this pilot study involving 32 students. The figure reports the students' compiled performance in terms of their PODP's overall diagnostic accuracy, and sensitivity and specificity as related to each of the nine individual disease categories among the 72 criteria cases.

Another general comment is that when critically evaluated, those student PODPs which approach the performance of expert PODPs have disease prototypes which more closely resemble the disease prototypes of the experts. Figure 2 demonstrates a disease prototype. This particular example demonstrates a portion of the Myocardial Infarction (MI) prototype taken from an expert. It consists of the dis-

ease's (MI) features and their relative value (i.e. features are plac in descending order of importance — note the feature's value weight as represented in the FEAT.VALU. column). Also note th distinctions are derived as to whether the feature is used by the e pert as either a confirming feature (Positive [POS] feature correl tion – as represented in the FEAT.CORR. column) or as a differe tiating feature (Negative [NEG] feature correlation — as also repr sented in the FEAT.CORR. column).

Conclusions

One preliminary impression derived from this work is that expe program shells can be used to extract and study the problem-solvi elements of both experts and novices. Furthermore, we believe th this expert program shell is capable of focusing upon and delineati the commonalities and distinctions among the knowledge base ar models of disease prototypes used by both expert and novice. This turn may provide a more concrete means of accounting for e pert/novice performance distinctions.

As expected, expert derived knowledge bases and disea prototypes resulted in higher levels of diagnostic accuracy again criteria cases than did the student's knowledge base and disea prototype. In addition, the disease prototypes of students who a proach the performance levels of experts more closely resemble t expert disease prototypes than do low level performers.

The expert system appears to be able to rapidly and objective challenge and assess the diagnostic accuracy of the knowledge ba of students with a large number and variety of criteria cases. Prior the advent of computer-based expert systems, such an assessment w essentially impossible. Such large sampling of the diagnostic accura of a student's knowledge base and disease prototype can possib serve as a significant improvement in the validity of assessment clinical competence.

Finally, these computer-derived disease prototypes, which a based upon the individual's knowledge base, appear to suggest th the knowledge components (i.e. features and their values) containe within the knowledge base are of unequal disease confirming and di ferentiating value. That is, the components of each knowledge ba and disease prototype vary from strong disease confirming and di ferentiating knowledge components to weak disease confirming an differentiating knowledge components. This has lead us to entertai

the thought that knowledge bases, just as problem-solving "processing heuristics", contain a gradation of stronger to weaker components. □

References

1. Barrows H, Tamblyn R: Problem-Based Learning, An Approach To Medical Education, Springer Publishing, New York, 1980
2. Bordage G, Zacks R: The structure of medical knowledge in the memories of medical students and general practitioners: categories and prototypes. J Med Educ 1984; 18: 406-416
3. Cantor N et al: Psychiatric diagnosis as prototype categorization. J Abnorm Psychol 1980; 89(2): 181-193
4. Doyle K: Evaluating Teaching, DC Heath and Company, Lexington, MASS, 1983
5. Frederiksen N: Implications of cognitive theory for Instruction in problem-solving. Review of Educational Research 1984; 3: 363-407
6. Greeno: Some examples of cognitive task analysis with instructional implications. In Snow RE, Federico PA, Montague WA (eds): Aptitude, Learning, and Instruction. Volume Two. Cognitive Process Analysis of Learning and Problem Solving, Erlbaum, Hillsdale, NJ, 1980
7. Groen G, Patel V: Medical problem-solving: Some questionable assumptions. J Med Educ 1985; 19: 95-100
8. Hasher L, Zacks RT: Automatic processing of fundamental information. Am Psychol 1984; 39(12): 1372-1388
9. Jason H, Westburg J: Teachers and Teaching in U.S. Medical Schools, Appleton-Century-Crofts, Norwalk, CONN, 1982
10. Jones JG, Cason GJ, Cason C: The acquisition of cognitive knowledge through clinic experiences. J Med Educ 1986; 20: 10-12
11. Kellog RT: Feature frequency and hypothesis testing in the acquisition of rule-governed concepts. Memory and Cognition 1980; 8: 297-303
12. Michie D: Introductory Readings in Expert Systems, Gordon and Breach Science Publishers, New York, 1982
13. Newell A, Simon H: Human Problem-Solving, Prentice Hall, Englewood Cliffs, 1972
14. Nitko AJ: Educational Test and Measurement: An Introduction, Harcourt, Brace, and Jovanovich, New York, 1983

15. Norman DA: Cognitive engineering and education. In Tuma DT, Reif F (eds): Problem Solving and Education: Issues in Teaching and Research, Erlbaum, Hillsdale, NJ, 1980
16. Norman GR: Objective measurement of clinical performance. J Med Educ 1985; 19: 43-47
17. Papa FJ, O'Reilly RP: An explicit analytical teaching and learning clinical problem-solving methodology: Implications of cognitive theory for medical school instructions and evaluation. In Hart IR, Harden RM, Walton HJ (eds): Newer Developments in Assessing Clinical Competence, Can-Heal Publications, Montreal, 1986
18. Simon HA, Hayes JR: Understanding complex task instructions. In Klahr D (ed): Cognition and Instruction, Erlbaum, Hillsdale, NJ, 1976

Figure 1
OVERALL DIAGNOSTIC ACCURACY
(AGAINST 72 CRITERIA CASES) 54%

	Sensitivity	Specificity
Myocardial Infarction (18 cases)	45%	92%
Myocardial Angina/Ischemia (26 cases)	39%	92%
Pericarditis (2 cases)	67%	94%
Dissecting Thoracic Aortic Aneurysm (0 cases)	–	94%
Upper GI Disorders (1 case)	94%	91%
Pneumonia (4 cases)	66%	99%
Pneumothorax (1 case)	88%	96%
Pulmonary Embolus (0 cases)	–	97%
Musculoskeletal Disorders (20 cases)	75%	91%

Figure 2
ANALYSIS OF MYOCARDIAL INFARCTION

	Feature	Feat. Valu.	Feat. Corr.	Diff. Total
14	< 20-30 minutes	0.06	neg	0.06
25	To neck/jaw/arm	0.05	pos	0.10
1	Patient > 40 years old	0.04	pos	0.15
21	Substernal/left precordial	0.04	pos	0.18
30	Movement/posturing	0.04	neg	0.22
18	Sharp/stabbing/fleeting	0.04	neg	0.26
23	Posterior thoracic	0.03	neg	0.29
61	Reproducible pain with movement/posturing	0.03	neg	0.33
15	> 20-30 minutes and < 24 hours	0.03	pos	0.36
19	Dull/pressure/squeezing	0.03	pos	0.39
43	Hemoptysis	0.03	neg	0.42
22	Lateral to costochondral junction	0.03	neg	0.44
4	Prodromal flu-like symptom complex	0.03	neg	0.47
7	Fever	0.03	neg	0.50
64	Cool/pale/moist	0.03	pos	0.53

THE USE OF A PATTERN RECOGNITION-BASED, PROTOTYPE-DRIVEN RESEARCH TOOL TO STUDY COGNITIVE CONSTRUCTS IN MEDICAL DECISION MAKING*

F.J Papa, J.H Shores , S. Meyer, Department of Medical Education
Texas College of Osteopathic Medicine

Introduction

Barrows and Tamblyn have forwarded what represents perhaps the most widely accepted characterization of medical decision making. They characterize the clinician as a scientist who uses what is fundamentally, a hypothetico-deductive process.

The medical decision making process proposed by Barrows and Tamblyn contains five components. They are as follows: 1) Initial conceptualization (problem identification), 2) Hypothesis generation, 3) Inquiry strategy, 4) Problem formulation (case building), and 5) Diagnostic/Therapeutic decision making (closure). Initial conceptualization or problem identification occurs within the first few seconds to minutes in most patient encounters. Within three to four minutes of the onset of the patient encounter the physician has generated several possible hypothesis or a cause list pertinent to the patients problem.

The physician then begins to develop a line of questions designed to rule in and rule out the list of possible causes (hypothesis). The physician's strategy quickly turns to case building (recognition of the disease state present). Finally, the physician determines that there is evidence sufficient for rendering a diagnosis and the initialization of therapeutic approaches.

College of Emergency Physicians (ACEP).

While the work of Barrows and Tamblyn, Elstein et al and many others has done much to increase our understanding of the medical decision making process, there remains many unasked and unanswered questions. This presentation will deal with those questions which arise from what has been termed the physician's "recognition" of the disease present in a given case.

Given that the physician uses some form of hypothetico-deductive reasoning process (a macro or weak cognitive process), how is it that the physician comes to recognize (diagnose) the signs and symptoms in a given case presentation as being representative of a specific disease process? In other words, what constitutes the micro or strong cognitive processes not addressed by the hypothetico-deductive process?

* This research is supported, in part, by SmithKline Beckman/FOCUS and the U. S. Department of Education Fund for the Improvement of Postsecondary Education (FISPE). We wish to thank the generous support of the members and governing body of the Texas chapter of the American

Pattern Recognition

Without elaborating, there is a large body of cognition research which has, as it fundamental principle, the notion that much of both perception and cognition is a pattern recognition process (Pao). There is growing evidence within the medical decision making literature which also suggests that diagnosis, a categorization task, utilizes a pattern recognition process (Norman).

The authors accept as a working premiss that pattern recognition is a plausible micro or strong cognitive process utilized in medical decision making. The authors, as have several other authors, pose as their next question "What symbolic, internalized form of knowledge representation is used as the template against which external data (case data) is compared in pattern recognition?"

Within the medical decision making literature Cantor et al and Bordage and Zacks have forwarded evidence which suggests that the clinician's pattern recognition processes (diagnosis) employes a single internalized "prototypical" representation of a disease class concept as the template against which case data is compared. These authors have added to the macro (weak) cognitive process paradigm (represented by the hypothetico-deductive process), a micro (strong) cognitive process of pattern recognition via prototypes.

The authors set out to extend this combined macro/micro (weak/strong) medical decision making paradigm. Specifically, they sought to develop a better understanding of the cognitive constructs underlying the pattern recognition component of the paradigm. To this end they developed a pattern recognition-based, prototype-driven research tool. The purpose of this tool is to model and explore the micro or strong cognitive constructs which underlie a pattern recognition-based, prototype-driven medical decision making paradigm.

The research tool (currently referred to as a Knowledge Base Inference Tool) (KBIT) and its methodological properties has been previously described (Papa and Meyer). KBIT is an artificial intelligence-derived, expert system-based research tool. It is designed to measure a subjects diagnostic accuracy against a data bank of criteria cases. KBIT is also capable of measuring at least two cognitive constructs underlying diagnostic accuracy, i.e. pattern matching and pattern discrimination (Papa et al).

In previous investigations, KBIT demonstrated that it is capable of reliable measures of diagnostic accuracy, pattern matching and pattern discrimination (unpublished work KR-20, .89). The study which is to be presented in this paper was designed to measure the construct validity of the measures derived from KBIT. Specifically, the authors set out to determine:

If there a statistically significant difference between experts (board certified practitioners) and medical novices (medical students) in terms of KBIT derived measures of diagnostic accuracy, pattern matching and pattern discrimination.

Evidence of construct validity in conjunction with assessment reliability would provide the authors with a research tool capable of reliable and vaild assessments of the cognitive constructs underlying a pattern recognition-based, prototype-driven model of medical decision making.

Methods

A total of 34 experts (board certified Emergency Medicine practitioners) and 122 third and fourth year clinical clerks at the Texas College of Osteopathic Medicine comprise the study subjects.

Using the methods previously described (Papa and Meyer and Papa et al), an acute chest pain knowledge base was extracted from each subject. This knowledge base was transformed by KBIT into a set of disease prototypes representing nine common and/or important diseases known to cause acute chest pain.

These nine diseases were myocardial infarction, myocardial angina, pericarditis, pneumonia, pneumothorax, pulmonary embolus, dissecting thoracic aortic aneurysm, esophageal/upper intestinal disorders and musculoskeletal disorders. Each subject's prototypes were used to diagnose 18 myocardial infarction criteria test cases.

Three measures were made of each subject's performance against the 18 cases. The measures were: 1) Diagnostic Accuracy, the number of myocardial infarction cases correctly diagnosed, 2) Pattern Match, the degree to which a correctly diagnosed case matched the subject's prototype, and 3) Pattern Discrimination, the degree to which a correctly diagnosed criteria case was differentiated from the second most likely diagnosis.

Results

Student-t test was used to measure the primary construct of diagnostic accuracy. The results demonstrated significant differences between the experts and novices with the experts demonstrating higher levels of diagnostic accuracy (p .001).

Student-t test was also used to measure the constructs underlying diagnostic accuracy, i.e. correct pattern recognition via pattern matching and pattern discrimination. The results once again demonstrated significant differences between the experts and novices with the experts demonstrating higher levels of pattern matching and pattern discrimination (p .001).

Discussion

Cognition literature in general and medical decision making literature both suggest that categorization tasks such as those which occur in differential diagnosis utilize a pattern recognition (diagnosis) via prototype paradigm. The authors set out to develop a pattern recognition-based, prototype-driven research tool expressly designed to test the validity of this paradigm.

The authors designed an advanced, artificial intelligence derived tool called KBIT, to measure three pattern recognition related constructs, i.e. diagnostic accuracy, pattern matching and pattern discrimination. If the paradigm and the research tool were to demonstrate construct validity then there should have been significant differences between experts and novices in terms of these three constructs.

KBIT demonstrated statistically significant differences between experts and novices along the three constructs tested. The ability of a research tool to demonstrate valid assessments in the constructs underlying medical decision making appears to be a breakthrough in medical cognition research. The authors are currently investigating the ability of KBIT to provide reliable and valid assessments of expert and novices decision making constructs across a variety of acute chest pain test cases (i.e. cases of pneumonia, angina, pneumothorax, etc).

If these results also demonstrate significant expert/novice differences in these three constructs across a specific problem area, then tools such as KBIT may some day be able to provide reliable and valid assessments of a subject's medical decision making skills across a variety of specific medical problems. The development of a reliable and valid, problem-specific assessment tool would be a significant advance in medical education.

References

1. Barrows HS, Tamblyn RM. Problem-based Learning: An approach to Medical Education. New York, Springer Publishing, 1980.

2. Pao YH. Adaptive Pattern Recognition and Neural Networks. Reading MA, Addison-Wesley Publications, 1989.

3. Norman GR, Rosenthal D, Brooks LR, Allen SW, Muzzin LJ. The Development of Expertise in Dermatology. Arch Derm 1989; 125:1063-1068.

4. Cantor N, Smith EE, French RD, Mezzich J. Psychiatric Diagnosis as Prototype Categorization. J Abnl Psyc 1980; 89:181-193.

5. Bordage GR, Zacks R. The Structure of Medical Knowledge in the Memories of Medical Students and General Practitioners: Categories and Prototypes. Med Ed 1984; 18:406-416.

6. Papa FJ, Meyer S. An Expert Program Shell Designed for Extracting "Disease Prototypes" and their use as Models for Exploring the "Strong Problem-Solving Methods" employed in Clinical Reasoning. In: Hart IR, ed. Further Developments in Assessing Clinical Competence. HEAL Publications, Montreal, Canada. 1987; 354-364.

7. Papa FJ, Shores JH, Meyer S, O'Reilly R, Bourdage R. The Role of Pattern Matching and Pattern Discrimination in Clinical Problem Solving. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP, eds. Teaching and Assessing Clinical Competence. Groningen, BoekWerk Publications, 1990; 317-322.

Computers Educ. Vol. 18, No. 1-3, pp. 45-50, 1992
Printed in Great Britain.

0360-1315/92 $5.00 + 0.00

A DIFFERENTIAL DIAGNOSTIC SKILLS ASSESSMENT AND TUTORIAL TOOL

FRANK J. PAPA,[1] JON I. YOUNG,[2] GERALD KNEZEK[2] and ROBERT J. BOURDAGE[1]

[1]Department of Medical Education, Texas College of Osteopathic Medicine, 3500 Camp Bowie Blvd, Fort Worth, TX 76107-2690 and [2]Texas Center for Educational Technology/University of North Texas, Texas, U.S.A.

Abstract—This paper reviews the progress made towards the development of an Intelligent Computer Assisted Instructional tool designed to function in a medical education setting. The tool, called KBIT (Knowledge Base Inference Tool) is an expert system-based instrument principally consisting of an assessment and a tutorial module. KBIT's sole purpose is to support the development and refinement of the differential diagnostic (DDX) knowledge and skills of medical students. The objective of the assessment module is to provide psychometrically reliable and valid measures of several DDS skills. The objective of the tutorial module is to create a learning environment wherein students make refinements in knowledge base (KB) constructs which result in progress towards the next level of DDX skills. KBIT's proposed educational approach is comprised of an iterative two-step process consisting of the assessment of several DDX skill performance parameters, followed by individualized formative instruction.

INTRODUCTION

This paper reviews the progress made towards the development of an Intelligent Computer Assisted Instructional (ICAI) tool designed to function in a medical education setting. The ICAI tool, called KBIT (Knowledge Base Inference Tool) is an expert system-based instrument principally consisting of an assessment and a tutorial module. KBIT's sole purpose is to support the development and refinement of the differential diagnostic (DDX) knowledge and skills of medical students.

DDX is the keystone intellectual skill of the medical practitioner. The objective of DDX is to determine which class of diseases best accounts for the patient's signs and symptoms. Medical practitioners initially use only the data obtained at the patient's bedside (i.e. historical and physical findings, not laboratory data) to reach a "clinical" diagnosis. However, diseases are rarely confidently diagnosed with such data. This is because disease states in general lack explicitly defined criteria for bedside-based diagnosis, i.e. a list of necessary and sufficient historical and physical signs and symptoms. Rather, the practitioner uses soft or fuzzy criteria to formulate a clinical diagnosis at the bedside. The practitioner subsequently attempts to confirm the clinical diagnosis with laboratory data. In short, the clinical (bedside) component of the diagnostic process represents decision making under uncertainty.

COMPUTATIONAL MODELS OF INFORMATION PROCESSING UNDER UNCERTAINTY

At least three general computational models of information processing under uncertainty have evolved[1]: probability, possibility (fuzzy logic or set theory), and certainty theory. The most widely utilized models are probabilistic, with Bayes' as perhaps the best recognized. Consequently, many researchers in cognition are not as aware of possibility and certainty theories as potentially useful information processing models in inherently uncertain decision-making domains. These alternative theories are sometimes referred to as deterministic theories. From the perspective of deterministic theories, the likelihood with which an exemplar is a member of a given class has nothing to do with the *a priori* occurrence of the given class in the population (as typified by probabilistic theories). Rather, a given exemplar is assigned, or determined to have, a "grade of membership" for each of several competing classes without consideration of each class's *a priori* occurrence. Without elaborating, Cohen[2] and Jungerman[3] have argued that probabilistic theories should not be unquestionably accepted as the only valid criterion for measuring the rationality or correctness of human decision making under uncertainty. Shortliffe and Buchanan[4] have gone

further to suggest that probabilistic models such as Bayes' are not appropriate methods in inherently uncertain decision-making domains such as medicine. A deterministic computational model functions as a critical component within KBIT's DDX paradigm.

THE TENDENCY TOWARDS A DETERMINISTIC APPROACH TO DECISION MAKING

Deterministic models are theoretically and mathematically viable information-processing models. However, evidence that one actually attacks uncertain classification tasks from a deterministic rather than probabilistic approach is supported by the work of Kahneman and Tversky[5]. Their frequently referenced study suggests that people perform classification tasks based upon the extent to which an exemplar is typical of, or a member of, a class. This is contrary to mathematically correct or "normative" probabilistic theories. This deterministic approach to classification, i.e. classification via recognition of the degree to which an exemplar is similar to the typical class representation, is frequently termed the "Representative Heuristic".

CLASSIFICATION AND PATTERN RECOGNITION

The medical cognition literature embraces two primary theories of classification; exemplar and prototype theories. These two theories attempt to describe, with detail greater than the representative heuristic described above, the type of knowledge used and how knowledge is used to perform classification tasks.

In exemplar theories[6] a clinician performs DDX (disease classification) by recalling the specific previously experienced disease exemplar which best matches the presenting case. The diagnosis associated with the best matching, previously experienced exemplar, provides the clinician with the diagnosis for the presenting case.

In prototype theories[7, 8] the clinician performs DDX by comparing the presenting case to an abstracted representation of each of the possible disease classes likely to account for the case presentation. The disease class prototype which best matches the case presentation is the diagnosis that will be made by the clinician.

The types of knowledge (exemplars and prototypes) used in classification is different in the two theories. However, it is important to note that both of these classification theories clearly express (while the Representative Heuristic implicitly suggests) that classification is accomplished via the use of a pattern recognition mechanism. The importance of pattern recognition in KBIT's DDX paradigm will be discussed later.

ASSESSMENT ISSUES

The medical education literature contains research sufficient to question the psychometric properties (reliability and validity) of DDX assessment instruments. The realization of truly efficient and effective DDX-related ICAI tools will not occur unless their developers can first resolve these psychometric concerns, for which there are at least three prerequisites. First, there is a need to create an explicitly defined and cognitively sound DDX paradigm for modeling a DDX assessment instrument. Second, because expertise in general, and DDX skills in particular, are problem and disease-specific, medical educators will need to create an assessment format which is capable of measuring competency at the problem and disease-specific level. Third, these assessment instruments must provide reliable and valid disease- and problem-specific measures for DDX skills. We have already described a cognition-based DDX paradigm. Possible solutions to the second and third prerequisites are now described.

The reliability problem

The reliability problem stems from the following two notions. First, the lack of disease criteria for clinical diagnosis speaks to the variability with which a disease class will manifest itself in different individuals. Second, there are a number of common and important diseases that are likely to cause a given medical problem. Subsequently, students' skills for disease and problem-specific

DDX can be reliably assessed only by having them solve a number and variety of test cases (perhaps six or more) for each of the diseases relevant to the given problem.

For a medical problem such as "acute chest pain", for which there are nine common or important different causes, it appears that a student would need to be tested with approx. 54 test cases (six different cases for each of the nine diseases in the problem area). With conventional DDX assessment instruments, a test case takes approx. 5–15 min to work through. With these assessment instruments, a prohibitively large amount of time would be required to reliably test each student's DDX skills in this area.

Utilizing conventional assessment formats, medical educators almost universally utilize only 1 or 2 test cases per disease class, or worse, per problem area. By lumping a large number of different test cases and question formats together, a respectable reliability coefficient of 0.70–0.80 might be achieved. However, in reflecting upon the notion that competency is at very least, problem-specific, one must ask "What is it that their conventional assessment approaches are measuring?" The simple answer is that they are not reliable estimates of competency with problem-specific skills.

With little elaboration, the promise of KBIT as a reliable, problem- and disease-specific instrument for assessment, comes from three sources. First, expert systems are, by definition, problem-specific in application. Second, once a knowledge base (KB) has been input into an expert system, there is almost no limit to the number and variety of test cases that it could be given to solve. Third, the DDX performance levels achieved by the expert system would reflect the diagnostic utility and soundness of the KB from whom the KB was extracted. Problem- and disease-specific test reliability would be, theoretically, a relatively easy psychometric property to achieve.

The development of an instrument for expert system-based assessment of DDX skills would require the creation of an expert system shell capable of extracting a subject's KB in a time-efficient manner. The approach to KB extraction taken by the authors has been described elsewhere[9] but will be briefly reviewed later. However, in studies conducted in two separate problem areas ("Acute Chest Pain" and neurological "Weakness"[10]) KBIT produced KR-21 reliability coefficients >0.89 with only 100 cases per problem area.

The validity problem

When experts outperform novices in a test of DDX skills then the test is said to have "construct validity". Perhaps the most critical psychometric concern confronting medical educators has been that experts do not necessarily perform better than novices with conventional DDX testing instruments.

An inherent capability of an expert systems-based assessment instrument is the potential to achieve construct validity. Put simply, a knowledge base extracted from an expert should outperform the knowledge base of a novice. KBIT has provided valid assessments at the disease-specific level[11]. KBIT has also provided valid assessments at the problem-specific level in two distinct problems areas ("Acute Chest Pain" and neurological "Weakness" [10]).

KNOWLEDGE BASE EXTRACTION AND DDX SKILLS ASSESSMENT

The process to extract a knowledge base in KBIT utilizes a single, predefined, "bounded" problem–space matrix. The matrix columns represent a list of x common or important diseases known to cause the problem, while the rows represent a list of y common signs/symptoms associated with each of the diseases in the problem space. The KB extraction routine requires each subject to fill in the empty cells of the matrix. That is, the student's task is to declare their understanding of the percentage of patients with a given disease who exhibit a given finding. These feature frequency estimates define their knowledge of the relationship between each disease and sign/symptom (see Fig. 1).

Via a series of manipulations, KBIT transforms these relationships into a highly structured representation of the subject's KB, which contains four interrelated, yet distinct, cognitive constructs. The first construct is a one-to-one representation of the subject's simple declarative KB, i.e. the original feature frequency estimates. The second construct is a more complex declarative KB construct termed a disease prototype (one prototype is created for each disease in the problem

	Disease # 1	Disease # 2	Disease # ...	Disease # ...	Disease # x
Feature # 1	65	50	20	35	40
Feature # 2	75	20	90	20	50
Feature # 3	30	40	35	15	80
...	90	20	20	10	40
...	90	95	30	05	20
...	85	95	80	10	20
Feature # y	10	50	75	90	20

Fig. 1. Subject's estimates of feature frequencies.

space). The third type of construct represents a form of procedural knowledge referred to as weighting rules. The declarative (prototype) and procedural (weighting rules) knowledge constructs are integrated into a fourth construct called a problem-specific DDX schema.

The purpose of these tranformations is to enable KBIT to use weighting rules and a fuzzy set theory-like inferencing mechanism based on pattern recognition to diagnose a collection of test cases. Diagnosis is conducted by having KBIT determine the degree to which each test case resembles or matches each of its internalized disease prototypes. Thus, KBIT's DDX information-processing paradigm emulates prototype-based classification theories. A test case is said to be correctly diagnosed when the disease class which has accumulated the greatest weight, i.e. highest degree of "prototype match", is the same disease class actually diagnosed for the test case. Three DDX skills measures are made for each subject. These are diagnostic accuracy, pattern matching and pattern discrimination. Diagnostic accuracy is defined as the number of test cases correctly diagnosed. Pattern matching is defined as the degree to which each of the subject's disease prototypes correctly matched the findings associated with all test cases representative of the same disease. Pattern discrimination is defined as the distance between a correctly diagnosed test case and the next most highly weighted disease class, i.e second leading hypothesis. Diagnostic accuracy, pattern-matching and pattern-discrimination values can be produced for disease-specific and overall problem areas.

CORRELATIONS BETWEEN SKILLS AND CONSTRUCTS

KBIT's assessment parameters represent measures of three different levels of DDX skills. Diagnostic accuracy represents a coarse DDX skills measure (both for a disease and for the general problem level) while pattern matching and pattern discrimination represent two finer, yet distinct, DDX skills measures. The authors have attempted to determine the degree to which refinement at one level of DDX skills might impact another DDX skill. Preliminary investigations suggest that diagnostic accuracy is more dependent upon pattern discrimination skills then pattern matching skills[8]. However, because of KBIT's design, each of the three DDX skills parameters represent estimates of the utility of each subject's four cognitive constructs. Given these inter-dependencies between skills and constructs, the finding that diagnostic accuracy is more dependent upon pattern discrimination than pattern matching suggests that it is the distinctiveness between an individual's prototype constructs which best accounts for diagnostic accuracy. This hypothesis represents the beginning of efforts to define more precisely the correlations between diagnostic skills and KB constructs.

ADVANTAGES OF AN EXPLICIT COGNITIVE AND INTEGRATED PARADIGM FOR ASSESSMENT AND INSTRUCTION IN DDX

There are several potential advantages to an assessment instrument based on pattern recognition. First, there appears to be the capability to provide reliable and valid measures of three different problem- and disease-specific DDX skills. Second, there is the potential to correlate these three DDX skills performance measures with each of the four distinct yet interrelated cognitive constructs which, within the KBIT DDX paradigm, are responsible for the DDX skills performance levels achieved. Third, there is the possibility of predicting (in background) how modifications not just at a given construct level (e.g. weighting rules), but more so, at a specific aspect of a particular construct (e.g. the weighting rule which relates the feature of "fever" and the disease class called

pneumonia), would lead to x% improvement in, for example, the subject's diagnostic accuracy for pneumonia. Fourth, KBIT can use the prototypes and weighting rules derived from an individual expert or composite group of experts as the basis for modeling particular constructs or performance activities in novices.

CURRENT AND PROPOSED INSTRUCTIONAL FEATURES

The immediate challenge is to determine how to integrate KBIT's current assessment capabilities with an instructional module which optimizes learning. The approach taken thus far has been to base the construction of the instructional module on the work of Burton[12], who used a seven-stage strategy for the development of instructional aids. This approach is illustrated in Table 1.

Level 1 (Help—the lowest level), the student is provided with the tools and information necessary to navigate through the system via help through built in cues and instructions. The students are informed of the information they need to provide, and, how to perform specific tasks. In the KBIT program this option is fully implemented.

Level 2 (Assistance) and level 3 (Empowering tools), KBIT is rather weakly implemented. There is no context sensitive help nor is there an historical summary of the student's performance. However, there currently is an tool which allows the student to modify feature frequency estimates, transforms them into new weight rules and offers the student an opportunity to view the new levels of diagnostic accuracy resulting from these changes.

Level 4 (Reactive learning), permits the student to propose diagnostic strategies [i.e. determine the specific feature(s) to be used, the number of features to be used and their order] and test their strategy against the test case data bank. KBIT provides feedback concerning the accuracy of the strategy against a specific test case or all test cases in the data bank. Level 4 will support reiterative interactions with the subjects via a repetitive process of strategy changes and skills re-assessments.

Level 5 (Modeling), allows the student to observe an expert perform diagnosis on a given case and indicates why the expert selected a particular feature in solving the case. Eventually, as additional experts are entered, it is possible that a student could choose a specific expert to watch or KBIT could match a student with an expert based on similarities between expert and student over a number of cognitive constructs or DDX skill performance levels.

Level 6 (Coaching), is the process of assisting the student with suggestions as to which learning options would provide the most valuable information. This is currently planned as being done in two ways. First, as the student faces a learning decision (e.g. which construct changes to make), he/she can ask for help from the coach. Second, if the student makes an inappropriate selection the coach can interrupt and offer an explanation as to why that choice is not the best and even provide the student with a better learning option. Coaching will interact with the subjects at the construct levels of prototype and weighting rules modifications. An iterative process of KB modification and skills re-assessments is envisioned.

Table 1. Burton's categories of software aids

	Burton's examples	KBIT's tutoring possibilities
1 Help	On-line documentation	Glossary of terms Program navigation
2 Assistance	On-line Calculator	On-line calculator Context sensitive help
3 Empowering tools	Decison tree history for self review	Structured log of %/accuracy of results
4 Reactive learning	Challenge system with hypothesis Get feedback on consistency of choices	Current "simulation" plus possible tutoring excerpts
5 Modeling	System trouble-shoots fault while student watches	Student "watches" expert diagnose case
6 Coaching	System recognizes suboptimal behavior and breaks in	System recognizes lack of progress and suggests alternative activities
7 Tutoring	Teaches and test mastery	Teaches fundamental concepts

Level 7 (Tutorial), has not been implemented. However, the intention is to provide the student with free form access to all prior levels so that the individual style of the student can be taken into consideration.

CONCLUSION

The authors have made significant progress towards the development of an expert system-based ICAI tool whose single purpose is to support the development and refinement of the DDX knowledge and skills of medical students. The majority of the work to date has involved: (1) the development of an explicitly defined and sound DDX paradigm which serves as the cognitive foundation of the ICAI tool, and (2) the development of a psychometrically reliable and valid instrument for problem-specific assessment which measures DDX skills levels in a manner consistent with an explicitly defined DDX-skills paradigm. The authors are in the early phases of modeling the instructional phases of the ICAI tool.

The most exciting findings involve those which suggest that the assessment tool has provided a robust research environment for exploring the correlations between the DDX skills performance levels achieved and the constructs responsible for the DDX skills performance levels. These findings suggest that ICAI projects have great potential utility not as ends in themselves but also as research tools to be used to actively model and test information-processing hypotheses.

Acknowledgements—This research was funded in part by the Fund for the Improvement of Post Secondary Education (FIPSE) and SmithKline Beecham Foundation.

REFERENCES

1. Parsaye D. and Chignell M., *Expert Systems for Experts*. Wiley, New York (1988).
2. Cohen L. J., Can human irrationality be experimentally demonstrated? *Behav. Brain Sci.* **4**, 317–370 (1981).
3. Jungerman H., Two camps of rationality. In *Decision Making Under Uncertainty* (Edited by Scholz R. W.). Elsevier, Amsterdam (1983).
4. Shortliffe E. H. and Buchanan B. G., A model of inexact reasoning in medicine. *Math. Biosci.* **23**, 251–279 (1975).
5. Kahneman D. and Tversky A., On the psychology of prediction. *Psychol. Rev.* **80**, 237–251 (1973).
6. Norman G. R., Rosenthal D., Brooks L. R. and Allen S. W., The development of expertise in dermatology. *Archs Dermat.* **125**, 1063–1068 (1989).
7. Bordage G. and Zacks D., The structure of medical knowledge in the memories of medical students and general practitioners: categories and prototypes. *J. med. Educ.* **21**, 92–98 (1985).
8. Papa F. J., Shores J. H. and Meyer S., Effects of pattern matching, pattern discrimination and experience in the development of diagnostic expertise. *Acad. Med.* **65**, S21–S22 (1990).
9. Papa F. J. and Meyer S., An expert program shell designed for extracting "Disease prototypes" and their use as models for exploring the "Strong problem solving methods" employed in clinical reasoning. In *Further Developments in Assessing Clinical Competence* (Edited by Hart I. R.), pp. 354–364. Heal, Quebec, Canada (1987).
10. Papa F. J., Test of the generalizability of KBIT (an artificial intelligence-derived assessment instrument) across medical problems. Unpublished Ph.D. dissertation, University of North Texas (1991).
11. Papa F. J., Shores J. H. and Meyer S., The use of a pattern recognition-based, prototype-driven research tool to study cognitive constructs in medical decision making. *Proceeding of the 4th International Conference on Assessing Clinical Competence*, Ottawa, Canada. In press.
12. Burton R. R., The environment module of intelligent tutoring systems. In *Foundations of Intelligent Tutoring Systems* (Edited by Polson M. C. and Richardson J. J.). Erlbaum, Hillsdale, N.J. (1988).

STRUCTURE OF MEDICAL COGNITION

Moderator: Christine McGuire

Effects of Pattern Matching, Pattern Discrimination, and Experience in the Development of Diagnostic Expertise

FRANK J. PAPA, JAY H. SHORES, and STEVE MEYER

Pattern recognition, prototypes, and experience play significant roles in medical decision making. To study the role that these factors play in the development of diagnostic expertise, a pattern-recognition–based, prototype-driven model of medical decision making was created. The model, an assessment tool derived from artificial intelligence (AI), provides valid measures of diagnostic accuracy and two prototype-related contributors to pattern recognition, this is, pattern matching and pattern discrimination.

In this study an AI assessment tool used disease-by-feature frequency estimates from each subject to create disease prototypes for each of 9 common causes of acute chest pain. The AI tool then used each subject's 9 prototypes and a pattern-recognition–based decision-making mechanism to diagnose 18 myocardial infarction cases. The data were analyzed to describe the role of pattern matching, pattern discrimination, and experience in the development of diagnostic expertise for myocardial infarction. The following questions are addressed:

1. Is there a statistically significant relationship between diagnostic accuracy and measures of pattern matching and pattern discrimination?
2. Is the effect of pattern matching and pattern discrimination on diagnostic accuracy independent of experience?

Researchers have attempted to determine whether expert/novice diagnostic performance differences were primarily due to differences in the formation or use of declarative or procedural knowledge. Elstein and colleagues[1] and Barrows and Tamblyn[2] among others attempted to describe expert/novice differences with comparisons of procedural knowledge. Despite numerous efforts, they did not account for expert/novice differences on the basis of procedural knowledge.

Grant and Marsden[3] and Bordage and Zacks[4] presented evidence supporting the existence of expert/novice declarative differences in knowledge-base content and knowledge-base structure. Their studies did not tie expert/novice differences in content-related and structure-related declarative knowledge to differences in diagnostic accuracy. Norman and colleagues[5] successfully related diagnostic accuracy to a pattern recognition process derived from knowledge of multiple past instances or examples.

Medical decision making is a categorization task. To carry out this task required for clinical diagnosis, some cognitive scientists believe that clinicians use declarative and procedural knowledge to form a structured knowledge base. Within this framework, a physician's knowledge base contains many elements, some of which are conceptual representations of disease classes. A given disease-class concept is internalized as a structured set of weighted, disease-related features (signs/symptoms). This structured set of disease-related weighted features is often referred to as a pattern or prototype. These disease-class concepts are used by clinicians to classify a patient's signs and symptoms as being due to a specific disease. It is suggested furthermore that diagnostic (class categorization) performance is based upon a prototype-to-example comparison.[4,6] The physician compares findings in the patient with a mental catalogue of disease prototypes.

Papa and Meyer[7] designed an AI-derived tool to model this explanation of medical decision making. This framework has been extended to suggest that the physician's ability to correctly diagnose (recognize) cases depends upon two underlying constructs, that is, the degree to which the patient findings match a prototype (pattern matching) and the extent to which that prototype is distinct from alternative prototypes (pattern discrimination). Measures of diagnostic accuracy, pattern matching, and pattern discrimination derived from this tool have demonstrated construct validity.[8] In the present study, the role that the two prototype-related constructs play in the development of diagnostic expertise is explored.

Methods

A total of 173 subjects at varied levels of clinical experience participated in the study (121 third-year and fourth-year medical students at the Texas College of Osteopathic Medicine, 18 emergency medicine residents, and 34 board-certified emergency medicine physicians).

A "problem space" for acute chest pain was created. It consisted of 67 historical and physical findings commonly associated with the clinical diagnosis of acute chest pain and a list of 9 common or important diseases known to cause acute chest pain. The 9 diseases were myocardial infarction, myocardial angina, pericarditis, pneumonia, pneumothorax, pulmonary embolus, dissecting thoracic aortic aneurysm, esophageal–upper intestinal disorders, and musculoskeletal disorders. The 67 features have been previously described.[7] These features included history findings (e.g., age > 40, male, sudden dyspnea) and physical findings (e.g., wheezes, rales, S4 gallop). By predefining the differentials and features to be used by all subjects, possible differences in the knowledge-base content among subjects were eliminated.

The program required that the subjects declare their knowledge concerning the relationship between each of the 9 diseases and the 67 features. Their knowledge base took the following form: "Within the context of Acute Chest Pain, what percentage of patients with <disease> have <finding>"? All subsequent performance measures were directly related to differences in the subjects' knowledge of disease-by-feature relationships.

The AI tool was written in structured BASIC. It used the subjective disease-by-feature relationship matrix and a non-Baysian mathematical model to transform each subject's knowledge base into a set of 9 disease prototypes. These prototypes were used to infer a diagnosis upon each of 18 confirmed myocardial infarction cases. The performance of each subject's prototypes was recorded. Measures of diagnostic accuracy, pattern matching, and pattern discrimination against each of the 18 criteria cases were recorded. Diagnostic accuracy was the number of myocardial infarction cases correctly diagnosed. Pattern matching was

the degree to which a correctly diagnosed criteria case matched the subject's derived prototype. Pattern discrimination was defined as the degree to which a correctly diagnosed criteria case was differentiated from the second most likely diagnosis.[9]

Results

Diagnostic accuracy was the dependent variable. Estimates of pattern matching, pattern discrimination, and experience (months of clinical exposure) were treated as predictor variables. An initial correlation matrix established that the measures were highly interrelated ($r = .31$ to $.62$). A preliminary regression analysis confirmed the individual strength of the predictor variables. They each accounted for a significant ($p < .001$) proportion of variance in diagnostic accuracy.

To describe the effects of the cognitive constructs (pattern matching and pattern discrimination) on diagnostic accuracy, it was necessary to remove variance due to experience. Thus, experience was forced into the regression in the first step. In the second step, measures of both pattern matching and pattern discrimination were entered. Experience and discrimination were both significant ($p = .0157$ and $p < .0001$, respectively); pattern matching did not account for a significant proportion of the variance ($p = .51$). Pattern discrimination independently accounted for approximately 39% of the variance in diagnostic accuracy, while pattern matching independently accounted for some 19%, and experience independently accounted for some 12%. Jointly, these three variables accounted for 42% of the variance in diagnostic accuracy.

Discussion and Conclusions

Pattern discrimination is a primary predictor of diagnostic accuracy in MI cases. The present findings suggest that the likelihood that an individual will be able to recognize correctly an example as belonging to a given class is dependent upon the relative distinctiveness of competing classes. In cognitive terms, this distinctiveness represents the psychological space between classes or prototypes. The more distinctive a given prototype, in comparison with competing prototypes, the more likely it is that cases will be correctly classified.

The importance of pattern matching in pattern recognition was not supported by the findings of this study. Pattern matching may be, in part, the basis for the development of pattern discrimination or may help to account for the development of diagnostic accuracy as individuals gather experience. Further study of this variable is needed.

Clinical experience is significantly related to the development of diagnostic accuracy. In a previous study, as clinical clerks saw more cases, their ability to match patterns increased in a slowly rising curve.[10] In the same population, pattern discrimination had a steep linear growth. It is possible that one must see many cases in a problem area in order to match patterns effectively. Perhaps the rules that govern pattern discrimination require less exposure for learning to occur.

This study supports the findings of prior investigators[3,4,6] that differences in declarative knowledge-base structures affect medical decision making. It also supports the assumption that pattern matching and pattern discrimination are constructs related to pattern recognition in the context of clinical diagnosis.

The ability to diagnose correctly is more dependent on the robustness (distinctiveness) of prototypes than the degree to which prototypes are matched to cases. Physicians with higher diagnostic accuracy have more distinct prototypes. Even though pattern matching was overshadowed by pattern discrimination in this study, it may play a significant role in the development of pattern discrimination.

The authors acknowledge the support of the membership and governing board of the Texas Chapter of the American College of Emergency Physicians, and Dr. Gloria Kuhn's assistance in collecting resident data.

This study was supported in part by The Fund for the Improvement of Postsecondary Education and Smithkline Beckman/FOCUS.Correspondence: Dr. Frank J.

Correspondence: Dr. Frank J. Papa, Department of Medical Education, Texas College of Osteopathic Medicine, 3500 Camp Bowie Blvd., Fort Worth TX 76107.

References

1. Elstein, A. S., Shulman, L. S., and Sprafka, S. A. *Medical Problem Solving: An Analysis of Clinical Reasoning.*, Cambridge, Massachusetts: Harvard University Press, 1978.
2. Barrows, H. S., and Tamblyn, R. M. *Problem-Based Learning: An Approach to Medical Education.* New York: Springer, 1980.
3. Grant, J., and Marsden, P. The Structure of Memorized Knowledge in Students and Clinicians: an Explanation for Diagnostic Expertise. *Med. Educ.* **21**(1987):92–98.
4. Bordage, G., and Zacks, R. The Structure of Medical Knowledge in the Memories of Medical Students and General Practitioners: Categories and Prototypes. *Med. Educ.* **18**(1984):406–416.
5. Norman, G. R., Rosenthal, D., Brooks, L. R., Allen, S. W., and Muzzin, L. J. The Development of Expertise in Dermatology. *Arch. Derm.* **125**(1989):1063–1068.
6. Cantor, N., Smith, E. E., French, R. D., and Mezzich. J. Psychiatric Diagnosis as Prototype Categorization. *J. Abnorm. Psychol.* **89**(1980):181–193.
7. Papa, F. J., and Meyer, S. A Computer-Assisted Learning Tool Designed to Improve Clinical Problem Solving. *Ann. Emer. Med.* **18**(1989):269–273.
8. Papa, F. J., Shores, J. H., and Meyer, S. The Use of a Pattern Recognition-Based, Prototype-Driven Research Tool to Study Cognitive Constructs in Medical Decision Making. Paper presented at the 4th International Conference on Assessing Clinical Competence, Ottawa, Canada, July 1990.
9. Papa, F. J., Shores, J. H., Meyer, S., O'Reilly, R., and Bourdage, R. The Role of Pattern Matching and Pattern Discrimination in Clinical Problem Solving. Paper presented at the 3rd International Conference on Teaching and Assessing Clinical Competence, Groningen, The Netherlands, May 1989.
10. Shores, J. H., and Papa, F. J. The Effects of Experience, Pattern Recognition, and Pattern Discrimination on the Development of Diagnostic Accuracy in Clinical Clerks. Paper presented at the 4th International Conference on Assessing Clinical Competence, Ottawa, Canada, July 1990.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☐ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☒ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").