ILIAD'S ROLE IN THE GENERALIZATION OF LEARNING ACROSS A MEDICAL DOMAIN*

Michael J. Lincoln MD, Charles W. Turner PhD, Peter J. Haug MD, John W. Williamson MD, Sylvia Jessen MS,
Robert M. Cundick, Ph.D., Kirt Cundick BS, Homer R. Warner MD PhD,
The University of Utah Department of Medical Informatics

*Supported in part by NLM grants 5R01-LM-05202 and 5 R29 LM-05260 and by the Veteran's Administration

ABSTRACT

Medical Informatics could facilitate more effective analysis and use of clinical knowledge by means of expert systems. To be most effective, such systems should be constructed in a manner which is consistent with physicians' cognitive processes. Our past five years' work with a system called Iliad indicates that it provides effective medical training and education. The current research extends our previous work by using a wider array of training and test cases. We also evaluated whether training on specific cases could generalize to improved testing performance on related cases, which featured similar complaints and pathophysiologic mechanisms, but different final diagnoses. In their junior internal medicine clerkship, students (n = 100) completed 1300 Iliad training cases covering 48 diagnoses. The findings indicated improved problem solving on the specifically trained cases as well as the generalization cases. We discuss a possible training model for expert systems such as Iliad.

INTRODUCTION

Medical outcomes could be improved if clinicians had better tools to manage clinical knowledge. Such tools should be designed to complement clinicians' natural cognitive processes. Designing such tools requires an understanding how medical experts and novices differ in their thinking. The key element in expert problem solving appears to be domain-specific knowledge and experience, rather than use of any general strategy. Elstein showed that physicians' abilities to work-up one type of medical problem do not predict their abilities in unrelated areas [1]. He also found that clinicians performed best on particular medical problems when they had recent case experience with expert-level performance feedback. Norman confirmed that medical expertise was highly variable across different clinical problems. He demonstrated that problem-solving expertise generalizes somewhat within a domain of medical knowledge (e.g., rheumatology), but does not necessarily generalize to other domains (e.g. pulmonary or cardiology). His work confirms Elstein's finding that the expert does not possess any innate or learned advantage in problem-solving techniques, but rather solves problems better "because he knows more in his domain than the novice." [2,3,4]. This work has led us to conceptualize "domains" of training as sets of medical problems which share major pathophysiologic features and chief complaints.

A logical conclusion from this work is that expertise in a particular domain might be developed and trained through systematic exposure to a series relevant clinical cases, followed by appropriate performance feedback. This conclusion underlies the structure of traditional medical education, particularly student clerkships and housestaff training. However, if clinical performance is very case experience dependent, then training on one medical problem area will not necessarily improve performance on others. This problem could be particularly important for training in tertiary-care hospitals, where the practice settings and case mixes are substantially different from those trainees will encounter in their future practices. One solution is to attempt to provide trainees with exposure to all the relevant diagnoses they may later encounter in practice. However, this approach is both expensive and difficult to accomplish, because faculty, students, and appropriate training cases must be brought together at the right time. An alternative approach, based on the concept of domain generalization, implies that selected cases could be used to train across a broader domain of related conditions. This approach could be cost-effectively implemented by using medical expert systems to provide simulated case experience and expert-level feedback.

The Iliad expert system Iliad is one medical expert system which has been evaluated for training students [5,6,7]. Version 4.0 of Iliad covers 1350 diagnostic conditions, and the system's data dictionary recognizes over 6300 types of disease manifestations. Iliad can present simulated patient cases as well as function as a diagnostic consultant or medical textbook [5,6,7]. In simulation mode, the user must "question" and "examine" the simulated patient. Iliad provides the simulated patient's results, and evaluates the user's diagnostic hypotheses and work-up strategy. The program now runs on IBM-compatible computers (with Microsoft Windows) and Macintoshes.

Iliad training model Clinicians are more likely to make errors in domains for which they have little recent, relevant case experience [1]. When these errors occur during simulated training cases, Iliad provides the domain-specific information necessary to correct these performance errors. Kassirer and Kopelman have identified three important types of cognitive errors [8]: (1) faulty hypothesis triggering, (2) faulty information gathering and processing, and (3) faulty verification of diagnoses. Iliad has been designed to remediate these errors.

Iliad can potentially correct improper hypothesis triggering, which occurs when physicians fail to generate appropriate hypotheses to explain key findings. For example, a student may think of pulmonary embolus as a possible explanation for sudden shortness of breath, but

0195-4210/92/$5.00 ©1993 AMIA, Inc.

fail to consider spontaneous pneumothorax. Iliad's Browse and Explain functions both can correct this error by indicating the link between "sudden shortness of breath" and both diagnoses. Students can also access Iliad's differential diagnosis, which indicates that both diagnoses should be considered.

Iliad can correct improper data interpretation. When students mistakenly estimate disease prevalence, Iliad's Browse function can show the correct *a priori* for any disease in the knowledge base. Iliad can also help when students mistakenly interpret test sensitivity and specificity information. The student may explore, using a "what-if" mode, the effect of the presence or absence of any relevant finding on the probability of this disease. Iliad enables the student to quickly select different combinations of findings, and explore exactly how combinations of findings actually modify the initial (a priori) prevalence estimate of the disease likelihood.

Iliad also helps students learn to adequately verify (confirm) their diagnoses. Our previous work indicates that junior students often make this mistake [7]. They overestimate the probability of their hypothesis and fail to obtain enough corroborating information, leading to premature and unsupported diagnostic conclusions. Iliad's Show Differential function can indicate that the hypothesis is not adequately confirmed. Then, using Iliad's Explain Disease function, the student can determine which diagnostic information could be obtained to complete the work-up.

The primary purpose of the current research is to determine the degree to which Iliad training on one problem is likely to generalize across a medical problem area to related diagnoses. As we have indicated, these related diagnoses may be on the differential diagnosis of the original condition, and share common features. Therefore, this research attempted to train students on the differential diagnosis surrounding specific training cases. In other words, the research was designed to determine whether simulation-based training can generalize to related conditions. By determining the degree of generalization which can be expected, the research could allow educators to select an appropriate case mix of simulated patients to span the range of required educational objectives.

METHOD

Subjects The subjects were third year medical students in the 1991-1992 (n = 100) class at the University of Utah who were participating in their internal medicine clerkship. These clerkships were conducted at the University of Utah Medical School: the LDS Hospital, the University of Utah Medical Center, and the Salt Lake VA Medical Center. All students were required to complete training and test cases with Iliad as part of their regular clerkship assignments.

Experimental Design The experiment was a 2 x 2 x 2 (Simulation Training Group x Simulation Testing Domain x Generalization Test Set) mixed factorial design. The first independent variable, which describes the training domain for each student, was a between subjects (uncorrelated) factor. The second and third independent variables describe the types of test cases assigned to each student. These later variables were repeated measures for each subject. This factorial design allows us to manipulate the independent variables while randomly counterbalancing the order of case presentations. Otherwise, differences in the difficulty of training and testing in the various case domains might confound the results. The primary dependent variables were Diagnostic Errors, Final Posterior probability, Cost of workup, and Findings Scores.

Independent variables The Simulation Training Group independent variable refers to the medical training domains assigned to each student (*Domain A/B* vs. *Domain C/D*). *Domain A/B* consisted of simulated patients who presented with a chief complaint of chest pain (A) or upper abdominal pain (B). *Domain C/D* consisted of patients who presented with a chief complaint of diarrhea (C) or hematuria (D). Each student completed two training cases per week during weeks 2-6 of their clerkship. *Domain A/B* students received both a chest pain and an abdominal pain case each week. *Domain C/D* students received a hematuria and a diarrhea case each week. We used two training domains (either A/B or C/D) for each student to increase the number and heterogeneity of training and test cases assigned to the student.

The Testing Domain variable refers to whether or not the student has received previous training on the specific disease being tested. Assume that *Domain A/B* students were trained on Pulmonary Embolus, while *Domain C/D* students were trained on Crohn's Disease. Now, *Domain A/B* students might be tested on either Pulmonary Embolus (*Trained* level of Testing Domain variable) or Crohn's Disease (*Untrained* level of this variable). For *Domain C/D* students, the Crohn's Disease case would be the *Trained* case and the Pulmonary Embolus case would be the *Untrained* case. This strategy permits us to balance the presentation of cases across the students, thereby controlling for differences in the inherent difficulty of the cases.

The Generalization Domain independent variable refers to whether or not the student's previous training is included in the same domain as the specific disease being tested. Again, assume that *Domain A/B* students were trained on Pulmonary Embolus, while *Domain C/D* students were trained on Crohn's Disease. When *Domain A/B* students receive a Spontaneous Pneumothorax test case, they receive the *Generalization* level of this variable. Similarly, when the *Domain C/D* students receive an Ulcerative Colitis case, they also receive a *Generalization* case. However, when students receive a case from the opposite domain, they receive a *Non Generalization* case. For example, a *Domain A/B* student might receive a test case of Ulcerative Colitis.

Student procedure Students received a two hour Iliad orientation on the first day of their clerkship, and continuing user support by medical faculty members [7].

The Clerkship Director required each student to complete two simulated training cases and one test case each week. Iliad computers were located on each student's medical ward so that students could conveniently complete their work [5,6,7].

When the students experience a simulation, Iliad first presents the chief complaint. Then, the student pursues additional patient findings (history, physical exam, and laboratory data). After each query, Iliad provides the simulated patient's responses. With each query, the student must indicate which hypothesis is being pursued and which hypothesis is currently most likely to account for the prior findings. In the learning mode, the student is alerted when possible diagnostic errors occur. In this mode, the student is also able to use the teaching tools (e.g., Browse, Explain). However, when working in test mode, the student is not alerted and the teaching tools are not available. Iliad also tracks the student's strategy and generates scores for the dependent variables. Students receive automatic feedback only after the test cases are completed.

A total of 48 training cases and 8 domain generalization cases were created from actual patient charts and validated by medical experts [6]. Each simulation is created in training and test versions (except for generalization cases, which only exit in test format). For each case, we identify two findings that have approximately the same sensitivity for the diagnosis. Then, we randomly assign one of these complaints to the training case and the other finding to the test version of the case. We also change the patients' age and sex between the training and test cases so that the two versions do not initially appear to be the same case. Students were instructed to complete the test cases without any assistance, reaching a degree of diagnostic certainty that would be equivalent to a posterior prevalence of 0.95.

Dependent variables Four different dependent variables were collected for each test case. The first dependent variable, Final Diagnostic Errors, assessed the correctness of the student's final diagnostic hypothesis. For each case, the student received a score for this variable of either 1.0 if they had the wrong final diagnosis, or 0.0 if they had the correct final diagnosis. A second dependent variable, Posterior Probability, measured the completeness of the student work-up. Each student received a score for this variable equal to the final posterior probability Iliad assigned to the correct diagnosis when the test case was finished [9]. Therefore the range of this score was 0.0 to 1.0. A higher score indicated that the student had elicited the appropriate findings to confirm the correct diagnosis. A third variable assessed the Cost of the student work-up. The value of this variable was the actual hospital charge the simulated patient would have accumulated at the University of Utah Hospital for the tests and procedures that the student ordered. A fourth dependent variable, Findings Score, reflects the students' choices of findings for a given diagnosis. We compare the student's choice of a finding to the most useful finding that could have been pursued at that point in the workup. These scores were calculated separately for History, Physical Exam, and Laboratory findings.

RESULTS

When we compared the students' performance across their repeated cases, we found that the performance on one case was not significantly correlated with performance on the other cases. This finding is consistent with the results in our previous studies, which supported the domain specificity hypothesis. Because the students' performances were not correlated, we treated all three independent variables as uncorrelated factors in the design. In the following analyses, we standardized each student's performance on each case. For each dependent variable, this standard score was equal to the student's actual score for the variable minus the mean performance on that variable for all students who completed the case. The analysis of variance was performed on these standardized scores. This procedure allowed us to remove the variability within each condition, due to case difficulty, without influencing the variability among the experimental conditions. This standardization procedure removes the variance due to Simulation Training Group, so this factor was not included in the analysis.

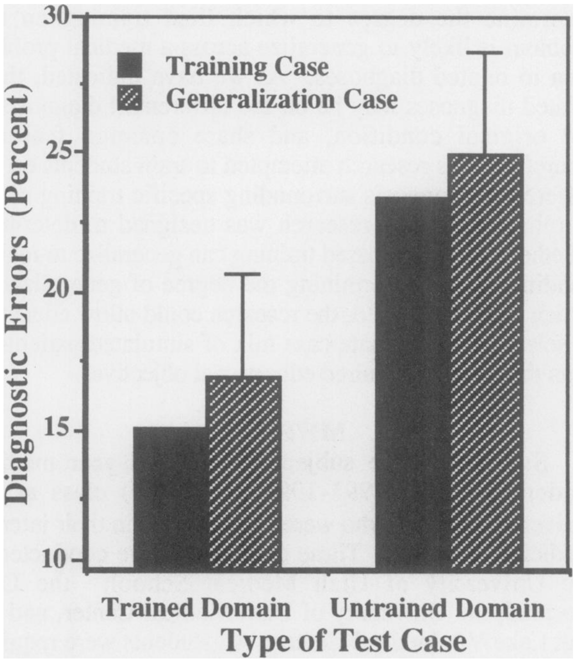

Figure 1. Effects of Testing Domain and Generalization Test Set on the students diagnostic errors. The error bars are based upon standard errors.

Final Diagnostic Errors The student's Final Diagnostic Error scores were analyzed with a 2 x 2

(Testing Domain x Generalization Set) factorial analysis of variance. The results indicated that the Testing Domain [$F(1,571) = 8.58$, $p < .004$] main effect was statistically significant. The Generalization Test Set [$F(1,571) < 1.00$, $p > .10$] main effect and the Testing Domain by Generalization Test Set [$F(1,571) < 1.00$, $p > .10$] effects were not statistically significant.

The means for the Final Diagnostic Errors variable are reported in Figure 1. For the Training Set test cases, (i.e., cases involving direct learning), students made fewer errors on the *Trained* (M = 15.1%) than the *Untrained* (M = 23.4%) diagnoses. For the *Generalization Set* cases, students made fewer errors on cases within the *Trained Domain* (M = 16.9%) as compared to cases in the *Untrained Domain* (M = 24.9%) cases. In other words, the performed better when their test case was similar to one they had previously experienced (but not the same diagnosis) than when the case originated outside their training domain. These findings replicate our previous work in showing that Iliad training did improve performance on cases for which students had been previously trained. The results also showed that students performed better on a *Generalization Case* when they had received training on a similar case within the same domain rather than training in another domain.

Posterior Probability The students' Final Posterior Probability scores were analyzed with a 2 x 2 (Testing Domain x Generalization Set) factorial analysis of variance. The results indicated that Testing Domain main effect was significant, [$F(1,571) = 6.89$ $p < .009$]. The Generalization Set main effect and the interaction effects were not statistically significant, $F < 1.0$. A comparison of the means for this effect indicated that students had higher posterior probabilities in the *Trained Cases* (M = 0.852) than on the *Untrained Cases* (M = 0.765). For cases at the *Generalization* level of the Generalization Set independent variable, students had higher posterior probabilities for cases within the *Trained Domain* (M = 0.820) as compared to cases in the *Untrained Domain* (M = 0.769) cases.

Cost of workup The Cost of the student's workup was analyzed using a 2 x 2 (Testing Domain x Generalization Set) factorial analysis of variance. The results showed that the main effects and interaction were not statistically significant, [$F < 1.00$]. The present findings do not replicate our previous findings. This work showed that training could reduce student workup costs [6,7]. However, some of the new test cases used in this experiment required expensive diagnostic work-ups. In other words, better performance on these cases was associated with higher rather than lower costs.

Quality of Findings For each finding requested, Iliad computes a finding score which compares the result to the best possible finding of the same class that could have been pursued. For example, a selected history finding is compared to the best alternative history finding. In order to interpret the significant generalization effects on the other dependent variables, we analyzed the students' performance on the *Generalization* level of the tests cases. The score reflecting the Quality of the Finding was analyzed with a 2 x 2 x 3 (Testing Domain x Type of Finding) repeated measures factorial analysis of variance. The Type of Finding independent variable (*History, Physical Exam, Laboratory)* was treated as a repeated measures factor. The results revealed significant main effects for the Type of Finding [$F(2,492) = 148.94$, $p < .001$] and the Testing Domain [$F(1,246) = 4.11$, $p < .04$] independent variables. The Testing Domain X Type of Finding interaction was not statistically significant [$F < 1.0$]. The means for this measure are reported in Figure 2.

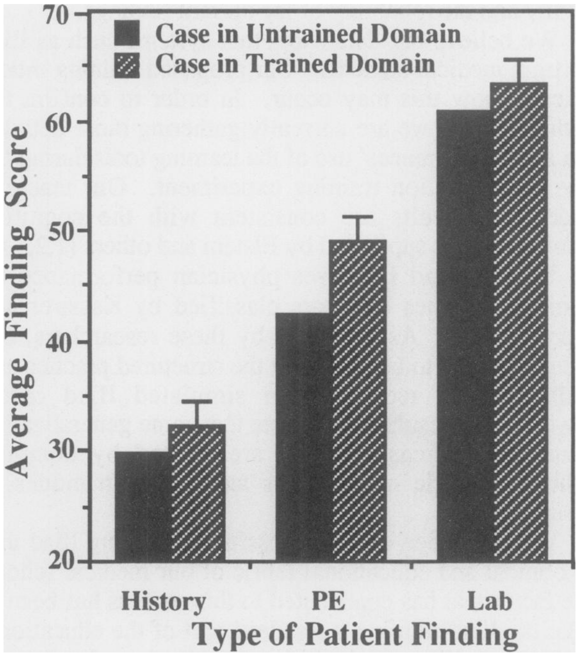

Figure 2. Effects of Testing Domain on students' Finding Scores for History, Physical Examination, and Laboratory. The results are only for the *Generalization* test cases. The error bars are based upon standard errors.

The cell means of this interaction were compared using a Neuman Keuls multiple range test. The result indicated that the *Laboratory* findings scores (*Trained* M = 63.38, *Untrained* M = 60.87) were significantly higher than both of the *Physical Examination* finding means (*Trained* M = 49.09, *Untrained* M = 42.33) which were also significantly higher than both of the *History* finding means (*Trained* M = 32.35, *Untrained* M = 29.92). The *Trained* and *Untrained* means for *Physical Examination* findings were also significantly different from each other. However, the two *Laboratory* and the two *History* findings scores were not significantly different from each other. These findings suggest that the beneficial effects of training for the *Generalization* cases might have resulted because students were able to select more appropriate physical examination findings in working up their cases.

DISCUSSION

The findings extend our earlier work with Iliad using a much larger sample of training (48) and test cases (56). The findings indicated that students were more likely to obtain the correct diagnosis and to achieve a higher posterior probability when their cases came from a *Trained* rather than an *Untrained* domain. These results suggest that the training effect generalized from the specific cases on which the student was trained to related cases on the same differential diagnosis. Additional findings in Figure 2 suggest that the improved effects on the *Generalization* cases resulted in part because students were able to select more appropriate physical examination findings in working up the case. Further research is needed to test the validity and the reliability of the current findings.

We believe our data show that systems such as Iliad can train medical students. Our proposed training model indicates how this may occur. In order to confirm the training model, we are currently gathering more detailed data about the trainees' use of the learning tools during the new generalization training experiment. Our teaching model and results are consistent with the cognitive learning models supported by Elstein and others [1,2,3,4]. We believe Iliad improves physician performance by training the types of errors classified by Kassirer and Kopelman [8]. As proposed by these researchers, our students appear to benefit from the structured practice and feedback they receive from simulated Iliad cases. However, our results do indicate that some generalization occurs among cases which are related by common pathophysiologic mechanisms and common modes of presentation.

We have been quite successful in weaving Iliad into the clinical and educational fabric of our medical school. One factor that has contributed to this success has been to make the Iliad training a standard part of the educational curriculum. The students have little time for "extra-curricular" activities, and must therefore often pass up potentially beneficial activities which are not required. The required nature of the training also minimizes any reactivity on the part of the students. They may object if only some students are required to do the work, as part of an "experiment". A second factor that has been important is obtaining the support of the faculty. They must be convinced that the students are spending their time on training which is both beneficial and relatively inexpensive (especially in terms of their limited time). To win over the faculty, we have pointed out that the cost of machines and software is quite small compared to the costs of providing extra case experience by hiring more doctors, building additional hospital beds, and employing the necessary ancillary personnel. A third important factor is to locate the computers in a convenient, natural location. For third year students, who spend most of their time caring for patients, we believe the computers should be located on the medical wards rather than in distant learning laboratories.

Our final observation is that a small, core group of clinical faculty should help manage the project and provide user support. At our hospital, we have created series of long-lasting, important relationships between the Department of Medical Informatics and the clinical faculty. However, other institutions now successfully using Iliad have been able to forge similar relationships by involving a core of well regarded, active, clinical faculty (often general internists). These faculty are required because they are highly visible and because they validate the clinical relevance of the project to young students and residents. We have found their presence provides a remarkable catalyst for student and resident support. Naturally, these faculty should work together with key non-clinicians, such as medical educators and medical librarians.

Acknowledgements: The authors are indebted to Drs. Odell, Williams, Lee, Sorenson, Bouhaddou and Yu who helped in implementing the experiment.

REFERENCES

[1] Elstein AS, Shulman LS, Sprafka SA. Medical Problem Solving - A Ten Year Retrospective. Eval and the Health Professions, 1990; 13:5-36.

[2] Norman GR, Tugwell P. A comparison of resident performance on real and simulated patients. Medical Education, 1982; 19:43-47.

[3] Schmidt HG, Norman GR, Boshuizen HPA. A cognitive perspective on medical expertise: Theory and implications. Acad Med, 1990; 65:611-621.

[4] Norman GR, Tugwell P, Feightner JW, et al. Knowledge and Clinical Problem-Solving. Medical Education 1985; 19: 344-356.

[5] Warner HR, et al. Iliad as an expert consultant to teach differential diagnosis. Proceedings of the 12th Annual Symposium on Computer Applications in Medical Care, Washington, DC: IEEE Computer Society Press, 1988; 12:371-376.

[6] Turner CW, et al. Iliad training effects: A cognitive model and empirical findings. Proceedings of the 15th Symposium on Computer Applications in Medical Care, Washington, DC: IEEE Computer Society Press, 1991; 15:68-72.

[7] Lincoln, MJ, et al. Iliad training enhances medical students' diagnostic skills. Journal of Medical Systems, 1991; 15:93-110.

[8] Kassirer JP, Kopelman RI. Cognitive Errors in Diagnosis: Instantiation, Classification, and Consequences. The Am J of Med, 1989; 86:433-440.

[9] Cundick R, et al. Iliad as a patient case simulator to teach medical problem solving. Proceedings of the 13th Annual Symposium on Computer Applications in Medical Care, Washington, DC: IEEE Computer Society Press, 1989; 13:902-906.