Iliad Training Effects: A Cognitive Model and Empirical Findings

Charles W. Turner PhD [2,1,4], Michael J. Lincoln MD [4,3,1], Peter Haug MD [1], John W. Williamson MD [4,3], Sylvia Jessen MS [1], Kirt Cundick BS [1], Homer Warner MD PhD [1]

The University of Utah Departments of: Medical Informatics[1], Psychology[2] Internal Medicine[3], Salt Lake City Veteran's Administration Medical Center[4]

Abstract

Iliad is a diagnostic expert system consisting of an "inference engine" (collection of rules and procedures for making decisions) and a "knowledge base" (collection of medical facts). Iliad's internal medicine knowledge base recognizes 5000 medical findings and covers 1150 diagnostic conditions in 10 subspecialty fields. We used Iliad's simulator mode to train diagnostic skills in junior-year medical students. The results corroborate previous findings documenting Iliad's teaching efficacy. Recent developments in cognitive psychology provide a framework for explaining Iliad's training effects.

Introduction

The Iliad expert system Iliad is a medical expert system which can provide simulated case training in the domain of internal medicine[1,2,3]. Faculty members create and validate the simulated cases, which are presented to students in a controlled fashion. The users "work-up" the cases by "questioning" and "examining" the simulated patients. Iliad "replies" with the simulated patient's answers, physical exam findings, and test results. During the simulated case work-up, numerous teaching tools can identify errors in the diagnostic work-up.

Some of the teaching tools compare the matching of the student's differential diagnosis to Iliad's optimal differential. For example, at each step in the work-up, the user's top diagnosis is compared to Iliad's top diagnosis and a "Hypothesis Score" is generated. Also, the "Explain Disease" function can provide a diagnostic explanation for any diagnosis on the differential. This explanation indicates how strongly each finding contributes to the current diagnostic probability. Additional teaching tools demonstrate how certain findings evoke consideration of specific diagnostic hypotheses. For example, the "Explain Finding" function creates a differential diagnosis for any selected patient finding. This function alerts students to trigger certain plausible hypotheses when they encounter this finding. Iliad alerts students to use these tools when they do not pursue the correct diagnosis or when they pursue a cost-ineffective work-up.

Iliad also allows assessment of sequential problem-solving in a testing mode. The test mode tracks and evaluates the student's work-up, but does not provide any training or feedback. Our prior work demonstrated that Iliad provides a valid means of training and assessing students [2]. Iliad's knowledge base of 1150 diseases and pathophysiological entities covers 90% of the diseases typically encountered on the medical wards. In addition, Iliad's diagnostic accuracy has been validated by entry of 500 actual patient cases.

The Iliad training model The Iliad training model is consistent with recent developments in the fields of cognitive psychology and medical decision analysis. These developments indicate that medical problem solving and decision making are highly domain or content specific. Therefore, the ability of an individual to solve a particular problem may be highly dependent upon the availability of domain specific knowledge relating to that problem [4]. This conclusion was documented for the work-up of live simulated patients [5,6] and computer simulated patients [7,8]. A result of domain specificity is that physicians may commit serious diagnostic errors when functioning in unfamiliar domains.

Elstein reasoned that physician skill in a particular domain (e.g., chest pain diagnosis in emergency rooms) is closely related to the amount of case experience in that domain. Because domain experience is critical, students should participate in a variety of diagnostic exercises focused in appropriate domains. Ideally, the training should provide incremental feedback at the sequential decision points in the work-up. This feedback allows students to learn by adaptively modifying their problem-solving repertoire. However, obstacles now limit the feasibility of this sort of training. First, the students experience severe time limitations and may be exposed to a rather limited case mix of patients. Second, faculty time for directed student feedback is limited. As a result, students experience inadequate training feedback and diagnostic supervision. We propose that these problems can be overcome by providing an appropriate diversity of simulated case experiences and guided training feedback.

Iliad training remedies diagnostic errors Iliad's learning tools are designed to train students to avoid the sorts of common diagnostic errors identified by researchers. Kassirer and Kopelman [9] proposed a model for recognizing the types of cognitive errors and biases that can influence medical decision making. Some errors they identified include (1) improper hypothesis triggering, (2) improper data gathering and interpretation, and (3) failure to adequately verify diagnoses.

0195-4210/91/$5.00

Kassirer and Kopelman proposed that improper hypothesis triggering occurs when physicians fail to activate or generate appropriate hypotheses to explain the patient findings. For example, a physician may fail to think of "spontaneous pneumothorax" as a possible explanation for sudden shortness of breath. If this physician has recently treated several pulmonary embolus patients who presented with sudden shortness of breath, the "availability" of the recent embolus experience may cause the physician to overlook the pneumothorax hypothesis [10]. Iliad's Explain Finding and Explain Diagnosis functions help remedy this error by reminding the physician of the relationship between sudden shortness of breath and other hypotheses, such as spontaneous pneumothorax.

Kassirer and Kopelman also propose that physicians err by using faulty estimates of disease prevalence, especially by overestimating the frequency of rare diseases. Because most patients present with common diseases, this can be a serious error. Iliad's Browse function can display the *a priori* prevalence of any disease, reminding the physician of which diseases are most likely to be present. This training reminds the physician to work-up diseases that are likely to be present. When physicians order tests for unlikely diseases, they will tend to obtain a high percentage of negative test results and pursue fruitless lines of diagnostic inquiry. A consequence of these actions is that the diagnostic work-up is not cost-effective. In contrast, physicians who pursue likely diseases tend to obtain positive test results which advance appropriate diagnostic hypotheses.

Kassirer and Kopelman also report that clinicians fail to adequately verify some diagnoses. Verification errors can occur when the clinician fails to collect sufficient findings to document a hypothesized diagnosis. These errors can result in premature or unsupported diagnostic conclusions. Such faulty conclusions could lead the physician to prescribe potentially harmful treatments. Iliad alerts physicians and students to recognize unsupported diagnoses and indicates which findings would be most useful and cost-effective to confirm the diagnostic conclusions.

This project investigated Iliad's ability to teach medical students better diagnostic and problem solving skills. Iliad simulations were used to both train and test junior students' problem-solving abilities. The experiment presented each student with a simulated training case followed one week later by a selected test case. Students alternated on a weekly basis between test cases which were similar to the previous weeks training case (same diagnosis, different presenting complaints) or dissimilar (different diagnosis, different presenting complaints). Students received cased in a counter-balanced order so that the testing sequence or week of the clerkship was not confounded with the specific medical diagnosis to be evaluated.

Method

Subjects The subjects were third year medical students (n = 75) in the 1990-1991 class at the University of Utah who participated in a six-week internal medicine clerkship. The student clerkships were conducted at the University of Utah Medical School: the LDS Hospital, the University of Utah, and the Salt Lake VA Medical Centers.

Experimental Design The experimental design was a 2 x 2 x 2 (Simulation Training Set x Simulation Test Set x Replication) mixed factorial design. The first and second factors were between subjects (uncorrelated) factors. A third independent variable was a within subjects (correlated, repeated measures) factor. The Simulation Training Set (Common-Uncommon) independent variable refers to the type of cases that the students randomly received during their simulation training. The cases either had relatively low prevalences in our teaching hospitals (Uncommon: Addison's; Multiple Sclerosis; Gonorrhea; Gastric cancer; Analgesic nephropathy) or relatively high prevalences (Common: Congestive heart failure; Myocardial infarction; Insulin dependent diabetes mellitus; Duodenal Ulcer; and Urinary tract infection). The Simulation Test Set independent variable refers to the types of test cases assigned to the students. Each student completed four test cases. They had been been Trained on two cases and Untrained on two cases. The Replications independent variable refers to whether the Test case was the first or the second instance of the case in each training set.

All students received a Trained and an Untrained test case during weeks 2 and 3 and again during weeks 4 and 5. The actual diagnosis and the sequence of Trained-Untrained cases were presented in different, counterbalanced random orders (i.e, a Latin square design). The patient test case in the first week for all students was tuberculosis; this case served as a baseline assessment. Four different dependent variables were collected for each test case. The first dependent variable assessed the errors of the student's final diagnostic hypothesis (Final Diagnostic Errors). A second variable measured the completeness of the student work-up (Posterior Probability). A third variable assessed the Cost of the student work-up. A fourth dependent variable was the Average Hypothesis Score.

Student procedure Students received a two hour Iliad orientation on the first day of their clerkship. Medical faculty members at each hospital were available to assist students with Iliad on a daily basis. The students met once a week with a faculty member who provided ongoing training and user support. All students were required to complete at least one simulated patient and one test case using Iliad for each week of the clerkship. Iliad computers (Macintosh SE-30) and printers were located on each student's medical ward [1,2,3]. Computers were placed on the wards so that the students do not need to leave the ward in order to use the

program. All simulations and test cases were completed while students were on the wards.

Simulation training procedure 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 Finding). In the test mode, the student is not alerted and the teaching tools are not available. In test mode, Iliad silently tracks the student's strategy and generates scores for the dependent variables. The student does not receive performance feedback until two days after all students have completed that week's testing.

Independent variables Faculty created ten different simulated cases with diagnostic problems defined as medical student clerkship objectives. Two independent internal medicine faculty reviewed each simulated case for validity. Five cases represented relatively prevalent diseases (Common level of the Training independent variable). Students may be be coincidentally exposed to actual patient examples of these relatively prevalent diagnoses at during their clerkship. The other five cases represented diagnoses (Uncommon level of the Training independent variable) that students were unlikely to see in their clerkship. One training case was presented during weeks 1 to 4 of the clerkship.

Each week all students received a test case consistent with the assigned level of the Testing Set independent variable. Students assigned to the Trained level of this variable received a test case which had the same final diagnosis as the previous week's learning case. To ensure that the cases did not initially appear similar, the test case was constructed so that the presenting complaints were different. Students assigned to the Untrained level of the Test Set independent variable received unrelated, untrained test cases. These cases also did not initially appear to be the same as the previous week's training case. In the successive weeks (2-5), conditions were reversed so that students alternately received Trained and Untrained test cases.

Test Procedure The students were instructed to complete the test cases without any assistance. On average, each test case required approximately 30 minutes for completion. Students were instructed to reach a degree of diagnostic certainty that would be equivalent to a posterior prevalence of 0.95. The students received prompt written feedback regarding the correctness of their final diagnostic hypothesis and the completeness of their work-up. In order to reduce student anxiety, individual test results were not disclosed to the medical faculty.

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 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 case was finished [1]. The range of this score was 0.0 to 1.0. Higher scores indicated that students 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 for the tests and procedures that the student ordered. A fourth dependent variable was the Average Hypothesis Score. This score was an average of the individual hypothesis scores that Iliad assigned at each stage in the case work-up. At each work-up step, an individual hypothesis score was generated when students chose a diagnosis most consistent with the current findings. These scores were calculated by dividing the probability that Iliad assigned to the student's best hypothesis by the probability that Iliad assigned to its own best hypothesis. For example, suppose Iliad's best diagnosis was pneumonia (probability 0.5) and the student's best diagnosis was chronic bronchitis (Iliad probability = 0.2). Then, the individual hypothesis score at this stage would be 50% (i.e., 0.2/0.4 * 100%). An Average Hypothesis Score is the average of individual scores which range from 0 to 100%.

Results

Final Diagnostic Errors The results indicated that students experienced significantly fewer diagnostic errors on trained than on untrained cases. The Final Diagnostic Error scores were analyzed using a 2 x 2 x 2 (Simulation Training Set x Simulation Test Set x Weeks) factorial analysis of variance. Since the findings indicated that the performance on one test case was independent of performance on another test case, all three independent variables were treated as uncorrelated factors in the design. The results indicated that the Simulation Test Set (Trained vs Untrained) main effect was statistically significant [$F(1,295) = 8.08$, $p < .005$] while the Simulation Training Set (Common vs Uncommon) main effect was not significant, $F < 1.37$. The Replications main effect and interactions were not statistically significant, F's < 1.4. The Training Set x Test Set interaction was statistically significant, $F(1,295) = 9.44$, $p < .002$. The means for this interaction are presented in Figure 1. A Neuman-Keuls multiple range procedure indicated that the highest rate of

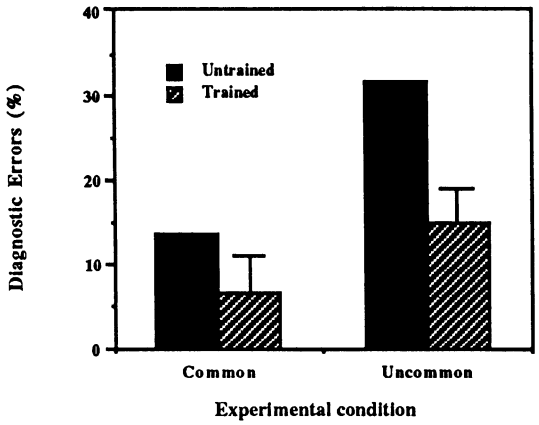

Figure 1: Effects of Simulation Training Set and Simulation Test Set on Diagnostic Errors. Error bars reflect standard error units.

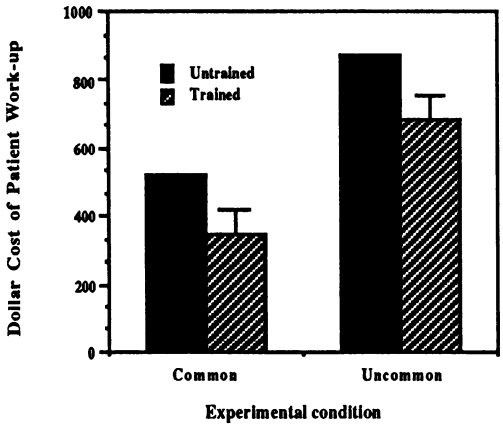

Figure 2: Effects of Simulation Training Set and Simulation Test Set on total Cost of the diagnostic work-up. Error bars reflect standard error units.

errors occurred for the Untrained, Uncommon diagnosis (M = .316). This value was significantly higher than the Untrained, Common (M = .137) or the Trained, Uncommon (M = .147). These two means were significantly higher than the Trained, Common (M = .067) condition.

Posterior Probability Results for the Posterior Probability scores indicated that students achieved significantly higher posterior probabilities in the Trained than the Untrained conditions. A 2 x 2 x 2 factorial analysis of variance indicated that the Simulation Test Set main effect was statistically significant, $F(1,295) = 3.75$, $p < .05$. The Training Set x Test Set interaction also was significant, $F(1,295) = 23.85$, $p < .001$. A multiple range comparison procedure of the means indicated that the Untrained, Uncommon mean (M = .632) was statistically significantly lower than the Untrained, Common (M = .89), the Trained, Uncommon (M = .76) means. These latter two means were significantly lower than the Trained, Common (M = .933) condition. The Simulation Training Set and Replications main effects and other interactions were not statistically significant, Fs < 1.5.

Cost Results for the Cost dependent variable indicated that students completed the patient work-ups with lower cost in the Trained than the Untrained conditions. The Cost scores also were analyzed using a 2 x 2 x 2 factorial analysis of variance. The results indicated that the Test Set main effect was statistically significant, $F(1,295) = 6.09$, $p < .01$. The Training Set x Test Set interaction also was significant, $F(1,295) = 22.54$, $p < .001$. The means for this interaction are reported in Figure 2. A multiple comparison procedure of the means indicated that the Untrained, Uncommon mean (M = $873.58) was significantly higher than the Untrained, Common (M = $517.53), and the Trained, Uncommon (M = $683.71) means. These latter two means were significantly lower than the Trained, Common (M = $347.22) condition. The Training Set and Replications main effects and other interactions were not statistically significant, $F < 1.6$.

Average Hypothesis Score The results for the Average Hypothesis Score indicated that students achieved higher scores in the Trained than Untrained conditions. A 2 x 2 x 2 factorial analysis of variance indicated that the Test Set main effect was statistically significant, $F(1,295) = 5.66$, $p < .02$. The Training Set x Test Set interactions also was significant, $F(1,295) = 47.27$, $p < .001$. A multiple comparison procedure of the means indicated that the Untrained, Uncommon mean (M = 43.89) was significantly lower than the Untrained, Common (M 71.24), the Trained, Uncommon (M = 54.33) means. Both Untrained means were significantly lower than the = Trained, Common (M = 78.71) condition. These effects indicate that untrained students perform better when diagnosing common disease conditions. The Training Set and Replications main effects and other interactions were not significant, $F < 1.4$.

Discussion

The present findings provide support for the hypothesis that the problem solving performance of medical students can be improved through experience with Iliad's simulated patients. The results indicate that students experiencing the Trained condition incurred fewer Final Diagnostic Errors and reached a more appropriate final Posterior Probability. Students also had higher Average Hypothesis Scores and incurred a lower Cost during the patient work-ups. The Final Diagnostic Errors and Posterior Probability variables measured whether or not the student was able to come to a correct

diagnosis with an appropriate degree of certainty. The Cost and Average Hypothesis Score variables measured the quality of the sequential decision-making choices involved in the work-up strategy.

Failure to trigger appropriate disease hypotheses can be a key reason for many errors in diagnosis [9]. Several Iliad functions remedy these triggering errors. The "Explain Finding" function provides a differential diagnosis (ordered by relative probability) of the various hypotheses which might explain the finding. For instance, the finding "elevated jugular venous pressure" is associated with the pathophysiologic process called systemic venous congestion. Systemic venous congestion is in turn a manifestation of right sided heart failure. These associations are represented in the Iliad's Browse mode. Iliad can teach the students to recognize these associations in subsequent encounters with similar cases, enabling the students to trigger appropriate hypotheses more readily.

Iliad trains students to verify diagnoses more accurately and reach a more appropriate final Posterior Probability. By examining Iliad's differential diagnosis ("Show Differential"), students learn to recognize whether they have completed the case work-up. If the student's top hypothesis is not the same as Iliad's, the student learns to pursue another hypothesis. When this learning is reflected in the simulation-test mode, Trained students achieve higher Average Hypothesis Scores. Experience gained by using Show Differential allows the students to better estimate whether they have reached an appropriate Final Posterior Probability. Trained students also learn to recognize which combinations of findings are adequate to achieve an appropriate Final Posterior Probability.

The Explain Diagnosis function also teaches students to recognize the base rates (*a priori* prevalences) associated with various diseases. These functions all improve the ability of students to gather and interpret data accurately for the trained disease domains. This training also improves the Average Hypothesis Score and the Final Posteriori, because students' differential diagnoses are better matched with Iliad's at all stages in the work-up. Iliad can help students to recognize cost-effective strategies. Iliad calculates the information value as well as the cost of the competing approaches. Iliad alerts students who have selected relatively cost-ineffective strategies and suggests the most cost-effective approach as an alternative.

The present findings corroborate our previous research [2] as well as other cognitive research indicating that clinicians can benefit from training in an adequately diverse range of diagnostic domains [4-10]. The 175 students contained in the current and previous research have completed approximately 2400 cases in Iliad's learning mode and 1800 cases in Iliad's testing mode. Our studies show that the Trained students commit fewer Diagnostic Errors, reach a more appropriate Final Posteriori, and attain higher Average Hypothesis Scores. The current study includes the Cost variable, which was not available for the previous research. A total of 30 diagnostic categories have now been trained using Iliad. Students (90%) report on anonymous surveys that the simulations are valuable educational tools[11].

References

[1] 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.

[2] Turner CW, et al. The effects of Iliad on medical student problem solving. Proceedings of the 14th Symposium on Computer Applications in Medical Care, Washington, DC: IEEE Computer Society Press, 1990; 14:478-482.

[3] 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.

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

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

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

[7] Norcini JJ, Swanson DB, Grosso LJ, Webster GD. A comparison of several methods for scoring patient management problems. Proceedings of the 22nd conference on research in medical education, Washington DC, 1983.

[8] Skakun EN, et al. Preliminary investigation of computerized patient management problems in relation to other examinations. Education and Psychological Measurement, 1979; 39:303-310.

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

[10] Tversky A, Kahneman D. Judgment under uncertainty. Science 1974; 185:1124-1131.

[11] Lincoln MJ, et al. Evaluating student acceptance of an expert system for teaching medical students. Proceedings of the American Medical Informatics Association, 1990; 1:84.

Acknowledgements

Grant number 5R01-LM-046043 from the National Library of Medicine supported the research. Drs. Richard Lee, William Odell, and Dean Sorenson provided valuable support in implementing the research.