Journal of Medical Systems, Vol. 15, No. 1, 1991

Iliad Training Enhances Medical Students' Diagnostic Skills

Michael J. Lincoln, Charles W. Turner, Peter J. Haug, Homer R. Warner, John W. Williamson, Omar Bouhaddou, Sylvia G. Jessen, Dean Sorenson, Robert C. Cundick, and Morgan Grant

Iliad is a computerized, expert system for internal medical diagnosis. The system is designed to teach diagnostic skills by means of simulated patient case presentations. We report the results of a controlled trial in which junior students were randomly assigned to received Iliad training on one of two different simulated case mixes. Each group was subsequently tested in both their "trained" and "untrained" case domain. The testing consisted of computerized, simulated patient cases for which no training feedback was provided. Outcome variables were designed to measure the students' performance on these test cases. The results indicate that students made fewer diagnostic errors and more conclusively confirmed their diagnostic hypotheses when they were tested in their trained domain. We conclude that expert systems such as Iliad can effectively teach diagnostic skills by supplementing trainees' actual case experience with computerized simulations.

INTRODUCTION

The Iliad System

Iliad is a medical expert system designed both to teach medical decision-making and to provide consultations on actual clinical problems. Iliad represents the culmination of over two decades of research in expert systems at the University of Utah's Department of Medical Informatics. The system is composed of an "inference engine" (a collection of rules and procedures for making decisions) and a "knowledge base" (a collection of medical facts and relationships). At the time this research was performed, Iliad's internal medicine knowledge base contained approximately 980 individual diagnostic logic units,

From The University of Utah Department of Medical Informatics; The University of Utah Department of Psychology; The University of Utah Department of Internal Medicine, Salt Lake City, Utah 84132; The Salt Lake City Veterans Administration Medical Center.

0148-5598/91/0200-0093$06.50/0

or "frames," which covered about 470 medical diagnoses. At that time, the system's dictionary recognized over 6000 medical terms.

Iliad functions in three modes: consultation, simulation, and simulation-test. In consultation mode, users begin by entering a real patient's findings. Iliad interprets the findings and produces a differential diagnosis, ordered by the relative probability of each diagnosis. If the patient work-up is incomplete, Iliad contains various learning tools which can suggest how to proceed. For instance, Iliad's "Most Useful Information" function can indicate which potential patient findings could provide the most useful diagnostic information at the least cost. In the simulation mode ("training mode"), Iliad presents the user with the chief complaint of a simulated patient. The user must "work-up" the patient by "questioning," "examining," and ordering lab tests and procedures. Iliad responds by providing the simulated patient's answers, examination findings, and test results. The user is required to postulate diagnostic hypotheses that explain the findings, and then test and refine these hypotheses. Iliad evaluates the user's performance and provides tailored feedback at each step in the work-up. In simulation-test mode ("test mode"), the user also works-up a simulated patient, but the user feedback and learning tools are withheld. In this mode, Iliad silently tracks and evaluates the user's performance.

The program is designed to teach junior medical students about important diseases that they are unlikely to see in real patients. However, the quality of the training Iliad provides depends not only on the teaching tools, but also on the quality of Iliad's knowledge base. This knowledge base was developed in an ongoing series of subspecialty-oriented "knowledge engineering sessions" which occurred twice weekly for each of nine subspecialties (e.g., cardiology, pulmonary, gastroenterology). Participants in these sessions included knowledge engineers (PhD and MD personnel trained in medical informatics), domain content experts (clinicians and professors at the University of Utah and elsewhere), and medical information specialists (e.g., medical librarians). We developed a priority schedule to ensure that each subspecialty session completed certain key frames which represented important teaching objectives.[1]

The task of organizing and storing the large amounts of medical knowledge created in the knowledge engineering sessions was managed by a Knowledge Engineering Support System (KESS).[2] We also developed compiling tools to convert the knowledge stored by the KESS into a format that the computer could process. These tools were adapted from early versions of the mainframe-based expert system compiler we developed for the HELP hospital information system.[3] The new tools are designed to allow a short cycle of modification and recompilation. This short cycle allows the knowledge engineers to quickly and efficiently modify frames in response to evaluation and validation procedures. Validation of the knowledge base is an important part of the knowledge engineering process. The frames in the internal medicine knowledge base have been validated by means of statistical analyses,[4] case entry of approximately 500 cases, and extensive expert review of completed frames.[5]

The quality of Iliad's training also depends on how accurately the system makes inferences from the knowledge base. The purely Bayesian inference strategies we used in previous work proved to produce overly confident, unreliable diagnostic results.[4] If similar degrees of overconfidence were allowed in a system used for teaching, trainees might be led into falsely or prematurely concluding diagnoses that were not actually supported. The overconfidence we found arose because the frames included medical case

findings which were not completely independent (they tended to occur together in cases: e.g., "fever" and "chills"). One of our early solutions to this problem was to trim the nonindependent findings from the frames, leaving only "key" findings. However, this approach resulted in "sparse" frames which could not provide the rich and diverse simulated case presentations required for effective training. Therefore, we rejected the trimming approach and tried a new strategy of "clustering," or grouping, the conditionally independent findings.[6,7] Such clusters can be represented as non-Bayesian frames in which the inferencing is handled by decision rules. These clusters often describe pathophysiologic concepts. For example, "Lung Consolidation" can be described by a frame containing the relevant physical findings (e.g., rales, bronchial breath sounds, egophony) and an appropriate decision rule to arbitrate the findings. These clusters can then be cited as "findings" in Bayesian frames. The Bayesian frames do not directly contain the dependent findings, but only an evaluation of their net information value, as provided by the cluster. Previous research by other investigators[6,7] as well as our own research[4] has demonstrated that this strategy improves the inferencing and significantly reduces the overconfidence.

How Iliad Trains

Iliad trains by providing students with carefully constructed opportunities to practice applying new knowledge in realistic, simulated patient cases. As the trainee makes sequential decisions during the work-up, Iliad provides many types of constructive feedback to help identify and train potential improvements in case management and outcomes.[8] Our experimental design assumes that students must receive domain specific, problem-based practice in order to become competent diagnosticians. This assumption is consistent with recent developments in the fields of cognitive psychology and medical decision analysis. These developments indicate that a physician's skills in medical problem solving are highly domain specific and must be maintained by constant practice which is focused on particular problem domains.[9] In particular, physicians solving a particular problem may be highly dependent upon the availability of domain specific knowledge relating to that problem.[10–12] The observation that domain specific knowledge is required for the successful solution of diagnostic problems has been documented for physicians evaluating live simulated patients[12] as well as for computer simulated patients.[13,14] Iliad can provide an inexpensive source of problem-based, simulated case material, focused in specific domains.

Iliad accomplishes the training by identifying and correcting specific types of diagnostic errors which can occur during the sequential decision-making involved in the work-up of the simulated cases. Kassirer and Kopelman[15] have proposed a model for recognizing the types of cognitive errors and biases that can influence medical decision making. They identified certain errors, including (1) improper hypothesis triggering, (2) improper data gathering and interpretation, and (3) failure to adequately verify diagnoses. Improper hypothesis triggering, an error we propose Iliad can correct, occurs when physicians fail to activate or generate appropriate hypotheses to explain the patient findings.[15] For example, a student may fail to think of "spontaneous pneumothorax" as a possible explanation for sudden shortness of breath. If this student has learned that pulmonary embolus patients may present with sudden shortness of breath, the

"availability" of the recent training[16] in pulmonary embolus may cause the student to overlook the pneumothorax hypothesis. During a work-up of a simulated case of pneumothorax, Iliad's "Explain Finding" and "Explain Diagnosis" functions help remedy this error. If the student neglects to consider the pneumothorax hypothesis, these learning tools can suggest that pneumothorax may provide an alternative explanation for the patient findings.

Kassirer and Kopelman[15] also propose that practitioners may interpret data incorrectly. For instance, they may utilize faulty estimates of disease prevalence. A student practitioner who overestimates the frequency of a rare disease may mistakenly ignore a more common disease which better explains certain key patient findings. Iliad's "Browse" function can display the *a priori* prevalence of any disease, to remind the student of the actual disease prevalence relationships. In addition, Iliad's "Show Differential" function indicates how particular findings should be interpreted as modifying the posterior estimates of disease probability. These tools remind the student to work-up the diseases which are most likely to be present in the patient. If students do not receive this feedback, they may mistakenly order tests for unlikely diseases. When they make this mistake, they will tend to obtain a high percentage of negative test results and pursue fruitless, cost-ineffective lines of diagnostic inquiry. Iliad's "Most Useful Information" function reminds students to pursue cost-effective inquiries by measuring the relative information gain and cost of the diagnostic inquiries students make and comparing the student's choices to alternative choices. Students who receive Iliad's structured feedback will learn to pursue more likely diseases in a more cost-effective manner. These students will tend to obtain positive test results which advance appropriate diagnostic hypotheses.

Kassirer and Kopelman also report that practitioners may fail to adequately confirm or verify their diagnoses.[15] Iliad can also potentially correct this error. For instance, a student could commit a verification error by failing to collect sufficient findings to document a hypothesized diagnosis. A student who commits this kind of error may make premature and unsupported diagnostic conclusions. In this case, Iliad's Show Differential function can indicate that the probability associated with the student's hypothesis is low compared to that of alternative hypotheses. In addition, Iliad's Explain Diagnosis function can indicate how the case findings support the alternative hypotheses. Finally, Iliad's Most Useful Information function can indicate the most cost-effective findings which might be collected to confirm the alternative diagnoses.

EVALUATING ILIAD

The present study was designed to evaluate whether Iliad could improve students' medical decision-making skills in areas where they would otherwise be unlikely to receive adequate case-based training. Our goals were to examine the acceptability and teaching potential of the Iliad system among these students. For this study, students participated in the experiment during one 6-week clerkship in internal medicine which are conducted during the junior year. Our experiment occurred in the second semester clerkship. Each week, every student performed one weekly training simulation and one test-mode simulation. Students were randomly assigned to one of two training domains (different sim-

ulated case mixes). Regardless of their training domain, all students subsequently received simulation-test cases covering both domains. The present study focused primarily on examining the students' data interpretation and hypothesis verification skills.

Hypotheses

In this research, we examined the effects of Iliad training (simulated patient cases) on students' diagnostic skills. We designed this work to test several hypotheses. First, we propose that students who receive Iliad training on a particular diagnosis will perform better on several clinically relevant measures of diagnostic performance than students not trained on that diagnosis. Second, we propose that this training effect will be stronger for uncommon clinical cases, which the students do not frequently encounter in their training, than for more common cases. Third, we propose that increases in student performance are relatively domain specific. Students who receive Iliad training in one type of case will not necessarily perform better on cases from an Iliad-untrained area.

METHOD

Subjects

The subjects were all of the third year medical students ($n = 100$) in the 1989–1990 class at the University of Utah who participated in a six-week internal medicine clerkship. The data were obtained from four rotations (of approximately 25 students each) which occurred during the spring semester in 1990. The student clerkships were conducted at the LDS Hospital, the University of Utah Medical Center, and the Salt Lake Veterans Administration Medical Center.

Experimental Design

The experimental design was a 2 × 2 × 2 (Simulation Training Set × Simulation Test Set × Time) mixed factorial design. The first two factors were between subjects (uncorrelated) factors, while the Time factor was within subjects. The Simulation Training Set (Uncommon-Common) independent variable refers to the type of training cases that the students were randomly assigned to received during their simulation training. These cases either had a relatively low prevalence or high prevalence in our teaching hospitals. The Simulation Test Set independent variable refers to the types of test cases assigned to the students. Each student received test cases which either resembled (Trained level) or did not resemble (Untrained level) the previous week's training case (see below in "Independent variables" for definition of "resemblance"). The Time variable refers to whether the student was completing the first or the second replication of the experiment. The first replication of a Trained and Untrained pair of test cases occurred in weeks 2 and 3 and the second replication occurred in weeks 4 and 5. Three different dependent variables were collected for each test case. The first dependent variable, Final Diagnostic Errors, assessed whether the student's final diagnostic hypothesis was correct or not. A second variable, Posterior Probability, measured how adequately the student confirmed the correct diagnosis. The third dependent variable, the Average Hypothesis Score, mea-

sured how closely the student's differential diagnosis matched Iliad's during each step in the case work-up.

Iliad System

Iliad's inference engine was written in the C programming language for the Apple Macintosh computer. On the Macintosh, Iliad requires two megabytes of random access memory and approximately 1.5 to 5 megabytes of free hard disk space (depending on whether or not supplemental medical literature is desired). Iliad runs on any Macintosh computer (an IBM-PC version is under development) which meets these memory requirements. However, a 68020 or higher processor with math coprocessor is highly recommended. In this experiment, all students were trained on Macintosh SE-30 computers. All students used the version 3.0 of the Iliad software and knowledge base during this experiment.

Student Procedure

Students received 2½ hr of training in the use of Iliad during the first day of their clerkships. In addition to this basic orientation, weekly group and individual sessions were held by the medical faculty to provide technical support for the students. All students were required by the Clerkship Director to complete one simulated training patient and one simulated test case each week. To minimize potential student anxiety and reactivity, students were told that the exact results of their Iliad testing would not be revealed to their clinical supervisors. However, students had to achieve a minimum performance on each of the cases in order to pass their clerkship. Students on each of the medical wards of our three teaching hospitals were provided with Macintosh SE-30 computers and printers loaded with the Iliad system. A special program controlled access to the training and test simulations according to the conditions specified for each student in the experimental design. This program also made it impossible for one student to access or alter another student's results. In addition, students working on the same ward team received a different, counterbalanced order of case presentation. All students were instructed to work alone when using Iliad. These precautions were designed to prevent contamination across the experimental conditions. The patient test case in the first week for all students was Tuberculosis. The first week's test served as a practice case, because all of the students had been trained on Tuberculosis during the orientation session.

The student training experience with Iliad has been described in our previous work.[17] In summary, when the students experienced a training simulation, Iliad first presented the chief complaint. The student then pursued additional patient findings (history, physical exam, and laboratory data), and, in response, Iliad provided the simulated patient's responses. During the work-up, students were required to formulate a differential diagnosis. When requesting new patient information, the students were required to indicate both their best hypothesis and which hypothesis was being pursued. The Iliad training tools we have described (e.g., Show Differential, Explain Diagnosis) were available in the training mode, but not in the test mode. In test mode, Iliad collected the dependent measures relating to student performance. Students were eventually given performance

feedback on the test cases. However, this feedback did not occur until 2 days after all students had completed that week's testing.

Independent Variables

We created ten different simulated cases that were based on diagnostic learning objectives established for the junior clerkship by the Clerkship Director and medical faculty. The simulations were created using one of two methods. First, real cases were entered into Iliad via the consultation mode and then converted to simulation mode. When real cases were not readily available, Iliad's random simulation procedure was used to create a case, based on the probabilistic information contained in the knowledge base. Regardless of the simulation development procedure, two independent internal medicine faculty reviewed and validated the medical accuracy of each simulated case. Five cases represented relatively prevalent diseases in medical inpatients (Common: Congestive heart failure; Myocardial infarction; Insulin dependent diabetes mellitus; Duodenal ulcer; and Urinary tract infection) and five cases represented relatively uncommon diseases (Uncommon: Addison's disease; Multiple Sclerosis; Gonorrhea; Gastric cancer; Analgesic nephropathy).

Each student completed four training mode and four test mode cases during their clerkship. During week 2, every student received a test case consistent with their assigned level of the Test Set independent variable. Students assigned to the Trained level of this variable received a test case which resembled the previous week's training case (had many of the same diagnostic findings and had the same final diagnosis). However, to ensure that the students did not simply recognize the test case's resemblance to the training case, the test case was constructed so that the patient's age, sex, and initial complaints were different. Students assigned to the Untrained level of the Test Set independent variable received an unrelated, untrained test case during week 2. These cases also did not appear to be the same as the previous week's training case, and in fact proved to represent an entirely different diagnosis. In the successive weeks of the rotation (weeks 3 through 5), the Trained and Untrained conditions were alternated for each student. Therefore, as explained above, each replication of the experiment consisted of one Trained and one Untrained test case. The first replication occurred during weeks 2 and 3 of the experiment (Early level of the Time independent variable) and the other replication occurred during weeks 4 and 5 (Late level of the independent variable). The actual case sequence (Trained-Untrained) was presented in different, counterbalanced, random orders during each replication (i.e., a Latin square design).

The training and testing approach we adopted was designed to take advantage of the lack of substantial generalization expected to occur between unrelated training and test cases. Furthermore, this approach allowed all students to experience apparently equal training. We anticipated that equal appearing training would minimize potential student reactivity to overt experimental manipulation. Finally, this approach allowed all the students to have an equal opportunity to become experienced using the computer during the clerkship and score well on the test cases.

Testing Procedure

The students were instructed to complete the test cases without any assistance. On average, each test case required approximately 30 min for completion. Students were

instructed to reach a degree of diagnostic certainty that would be equivalent to a posterior prevalence of 0.95 (during training cases, they had been instructed to reach this same certainty level). The students received written feedback regarding the correctness of their final diagnostic hypothesis and the completeness of their work-up. However, this feedback was delayed until all students had finished that week's test cases. In order to reduce student anxiety, individual test results were not disclosed to the medical faculty.

Primary Dependent Variables

Three different dependent variables were collected for each test case.[8] The first dependent variable, *Final Diagnostic Errors,* assessed the correctness of each student's final diagnostic hypothesis. For each case, the student's response was defined as being either correct or incorrect. The dependent variable measured the percentage of students, in a particular condition, who obtained incorrect diagnoses. 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 had assigned to the correct diagnosis when the test case was finished. Therefore the range of this score was 0.0 to 1.0 for each student. Higher scores indicated that the student had elicited the appropriate findings to confirm the correct diagnosis. Across an experimental condition, we averaged the individual student scores on this variable. A third 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 simulated case work-up. At each stage in the work-up, the individual score was based on a comparison of Iliad's best hypothesis at that stage to the student's best hypothesis. These individual 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 that, halfway through a theoretical case, Iliad assigns a probability of 0.50 to its best diagnosis, Pneumonia. At that same point, the student's best diagnosis might be Chronic bronchitis, to which Iliad assigns a probability of only 0.20. Iliad's calculated individual hypothesis score at this stage would be 40% (i.e., $0.2 \div 0.5 \times 100\% = 40\%$). The Average Hypothesis Score is simply the average of these individual scores. Therefore, these scores ranged from 0 to 100%. Four students had incomplete cases because of computer failures or scheduling conflicts which prevented them from completing the replications.

RESULTS

Final Diagnostic Errors

In summary, we found that students who were Trained in Uncommon diseases committed significantly fewer Final Diagnostic Errors than students who were Untrained in Uncommon diseases. These results were supported by the students' Final Diagnostic Error scores, which indicated that neither the Simulation Test Set main effect [$F(1,91) = 2.16$, $p < 0.145$] nor the Simulation Training Set main effect [$F(1,91) = 1.12$, $p < 0.292$] were significant. The Time main effect was nearly significant [$F(1,91) = 2.86$, $p < 0.09$]. However, the Test Set by Training Set by Time interaction was significant

[$F(1,91) = 10.41$, $p < 0.002$]. This effect appeared to be due to improved student performance on the Final Diagnostic Error variable on Uncommon test cases during the second replication of the experiment (Late level of the Time variable). To test this hypothesis, we performed a planned comparison of the students' mean performance in the Uncommon, Untrained condition ($m = 21.7\%$) against the average of the other three conditions ($m = 7.7\%$).[18] Our analysis demonstrated that the students committed significantly more Final Diagnostic Errors when they worked-up Untrained, Uncommon test cases [$F(1,91) = 5.88$, $p < 0.02$]. The mean student performance on the Uncommon, Trained condition ($m = 10.9\%$) was not significantly different from the average of the Trained ($m = 6.0\%$) and Untrained ($m = 6.2\%$) Common conditions [$F(1,91) = 0.61$, ns]. The students' performance in these four conditions is depicted in Figure 1.

Posterior Probability

We also found that students who were Trained in Uncommon diseases attained a significantly higher Posterior Probability score than students who were Untrained in Uncommon diseases. The students' scores on this variable indicated that neither the Simulation Test Set main effect [$F(1,92) = 0.55$, $p < 0.461$], nor the Simulation Training Set main effect [$F(1,92) = 2.55$, $p < 0.137$], nor the Time main effect [$F(1,92) = 0.07$, $p < 0.794$] were statistically significant. However, the Test Set by Training Set by Time interaction was significant [$F(1,92) = 5.66$, $p < 0.019$]. This effect appeared to be due to improved performance on the Posterior Probability variable when Trained students experienced Uncommon test cases during second replication of the experiment. To test this hypothesis, we performed a planned comparison of the students' mean performance in the Uncommon, Untrained condition ($m = 73.6$) against the average of the other three conditions. Our analysis demonstrated that the mean Posterior Probability score for stu-

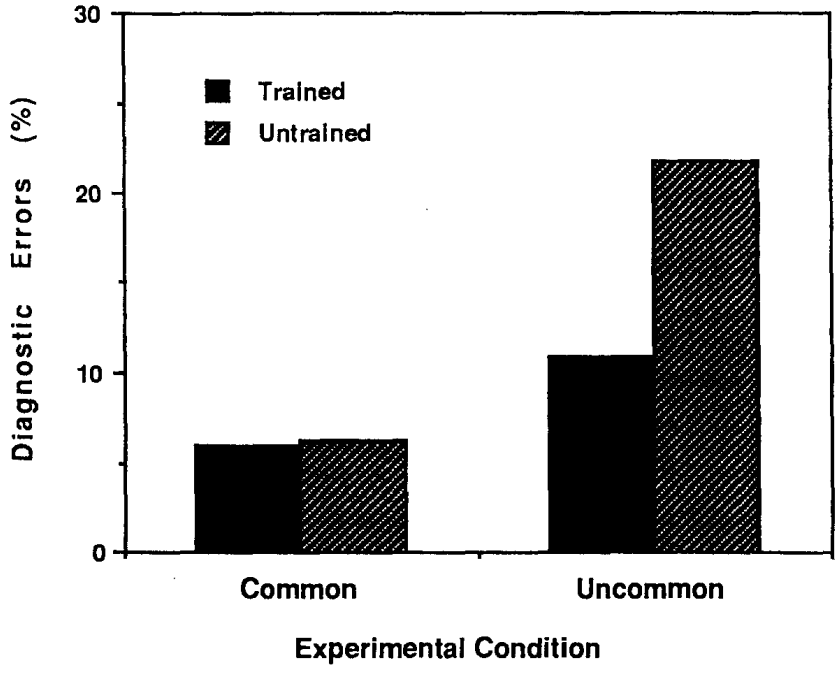

Figure 1. The students' ($n = 94$) mean performances on the Final Diagnostic Errors variable in the four experimental conditions. Lower scores reflect better performance.

dents in this condition was significantly lower than in the other conditions [$F(1,91) = 5.49$, $p < 0.025$]. The mean student performance on the Uncommon, Trained condition ($m = 85.2$) was not significantly different from the average of the Trained ($m = 84.9$) and Untrained ($m = 91.3$) Common conditions [$F(1,91) = 0.22$, ns]. The students' performance in these four conditions is depicted in Figure 2.

Average Hypothesis Score

For the Average Hypothesis Score variable, we found the effects of training were small and that the students scored much higher on the Common test cases, as compared to Uncommon test cases, regardless of their training condition. Specifically, the analysis of the students' Average Hypothesis scores indicated that the Simulation Test Set main was not significant [$F(1,94) = 0.20$, $p < 0.659$]. Also, the results indicated that the Simulation Training Set main effect was not significant [$F(1,94) = 0.74$, $p < 0.391$]. In addition, the Time main effect was not significant [$F(1,94) = 0.27$, $p < 0.603$]. However, the triple interaction between Test Set, Training Set, and Time was significant [$F(1,94) = 79.49$, $p < 0.001$]. The results indicated that students performed much better on the two Common conditions ($m = 64.0$) than the two Uncommon conditions ($m = 33.7$), [$F(1,94) = 48.43$, $p < 0.001$]. However, the pattern of results for the Average Hypothesis Scores was somewhat different than for the first two dependent measures. Specifically, the difference between the means in the Trained ($m = 65.6$) and Untrained ($m = 62.4$) Common conditions was not statistically significant. In the Uncommon condition, the difference between the Trained ($m = 35.8$) and Untrained ($m = 31.6$) means also was not statistically significant.

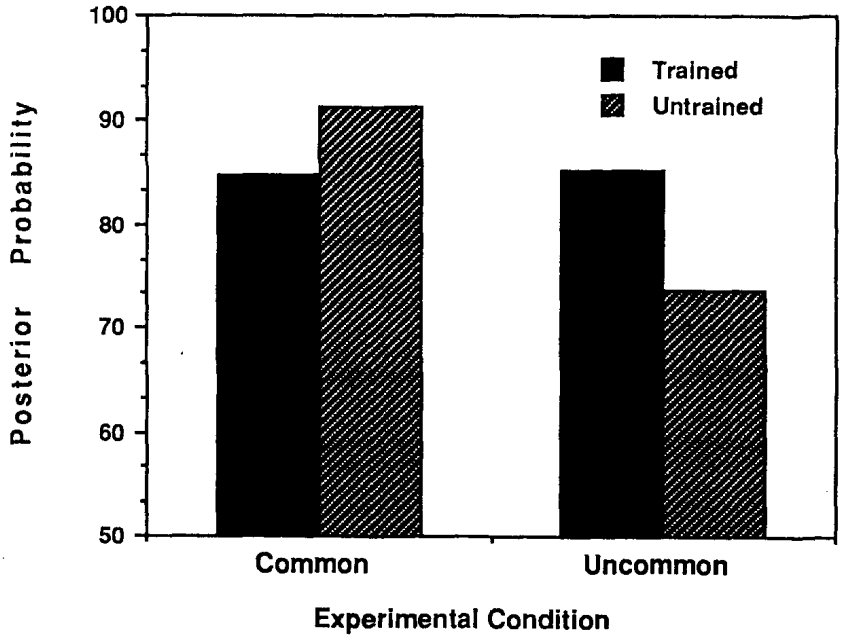

Figure 2. The students' ($n = 94$) mean performances on the Posterior Probability variable in the four experimental conditions. Higher scores reflect better performance.

Correlations between the Dependent Measures

We calculated the Pearson correlations among the dependent variables in each of the four Training and Testing combinations. The correlations among the student performances on different dependent variables for the same case can indicate whether or not the variables provide convergent measures of student skill. On the other hand, correlations among the student performances on the same dependent variable between different cases can indicate whether or not student training on one case generalizes to unrelated cases.

The comparison procedure involved taking the mean of the individual correlations across the four combinations of two Training Sets and two Test Sets (e.g., Uncommon Training, Common Testing). However, because r values are not normally distributed, we first converted the r values to normal (Z) scores, averaged the values, and then reconverted to r values. The mean correlations for each pair of dependent variables was tested for significance using Student's t test.

The dependent variables were significantly correlated with each other (on the same cases) within an individual experimental replication (replication one: weeks 2 and 3; replication two: weeks 4 and 5). For instance, the correlation between the Final Diagnostic Errors and Posterior Probability variables was $r = 0.706$ [$t(92) = 9.56$, $p < 0.002$]. The correlation between the Final Diagnostic Errors variable and the Average Hypothesis Score variable was $r = 0.391$ [$t(92) = 4.07$, $p < 0.002$]. Finally, the correlation between the Posterior Probability and Average Hypothesis Scores was $r = 0.400$ [$t(92) = 4.19$, $p < 0.002$].

We also found that these dependent variables were significantly correlated with themselves between different cases (across the two different replications). However, these correlations were substantially lower than the observed correlations among different dependent variables within a replication. For instance, the correlation between the Final Diagnostic Errors in replications one and two was $r = 0.224$ [$t(92) = 2.20$, $p < 0.025$]. The correlation between the Posterior Probability in these replications was $r = 0.236$ [$t(92) = 2.33$; $p < 0.02$]. Finally, the correlation between the Average Hypothesis Score across the replications was $r = 0.215$ [$t(92) = 2.11$; $p < 0.05$]. While these correlations are statistically significant, they explain only about six percent (e.g., $r = 0.236$, $r^2 = 0.056$) of the shared variance between the scores.

DISCUSSION

We have proposed that students who receive Iliad training on a particular diagnosis will perform better on several clinically relevant diagnostic measures than untrained students. In addition, we have proposed that this effect is stronger when students are trained on uncommon cases, for which they currently receive the least clinical training. To test these hypotheses, we examined students' performances when they were trained or untrained in diagnostic conditions representing particular training domains. We have also proposed that Iliad's training produces relatively domain specific increases in student performance. To examine this hypothesis, we correlated the student performance on the dependent measures within and between the experimental replications. The dependent measures we used to examine our hypotheses included: diagnostic correctness (Final

Diagnostic Errors), adequacy of diagnostic confirmation (Posterior Probability) and how well the student's differential diagnosis compared to Iliad's throughout the case (Average Hypothesis Score).

Before we discuss the specific training results, we must discuss the data relating to the validity of the dependent measures. If the dependent variables assess the same underlying quantities (various dimensions of diagnostic skill), student performance on these variables should be correlated within individual test cases. For example, a student experiencing a specific Trained test case should score highly across all of the dependent variables. When the same student experiences a specific Untrained test case, scores should be uniformly lower. As we expected, we found high correlations between the dependent measures within individual test cases within a given replication. For example, the correlation between the Final Diagnostic Errors and Posterior Probability variables was substantial and significant [$r = 0.706; t = 9.56, p < 0.002$]. Therefore, about 50% [$r^2 = (0.706)^2 = 0.498$] of the variance between student scores on these measures was correlated. Similar results were obtained within the test cases for the other dependent measures. These results indicate the convergent validity of the dependent measures.

The correlational data suggest that these three dependent measures are assessing three interrelated processes. For example, the Final Diagnostic Errors variable measures the student's ability to identify the correct diagnosis, while the Posterior Probability variable measures the student's ability to adequately verify the correct diagnosis. Students who collect an adequate amount of information to verify each potential diagnosis are likely to reach the correct diagnosis. These students tend to reach the correct diagnosis because they obtain sufficient information to rule-out reasonable (but incorrect) competitors. The Average Hypothesis Score measures how well the student's hypothesis is correlated with Iliad's at each stage in the work-up. Students must identify a plausible hypothesis (often, the correct final diagnosis) early in the case in order to earn a high Average Hypothesis Score. When students can recognize appropriate, plausible hypotheses, they are more likely to elicit appropriate patient findings to verify their diagnosis.

Efficacy of Iliad Training

The results indicate that students performed better on the dependent measures in the second replication of the experiment (second pair of Trained–Untrained test cases). However, this improvement only occurred when students experienced the Uncommon Training and Test Sets. When students experienced the Common Training and Test Sets, their mean rate of Final Diagnostic Errors was quite low regardless of whether they had been Trained (6.0%) or Untrained (6.2%). These low Final Diagnostic Error rates in the two Common conditions (Trained and Untrained) were not significantly different from the mean rate for Uncommon, Trained condition (10.9%). However, when the Untrained students experienced Uncommon tests, the mean rate of Final Diagnostic Errors was significantly higher (21.7%). Similar results were obtained for the Posterior Probability dependent variable. Students who were Untrained for the Uncommon tests received significantly lower Posterior Probability scores than the students in the other three conditions. For the Average Hypothesis Score, the results were not as clear. While the Test Set by Training Set by Time (replications) interaction was significant, the mean performances of the students were not clearly significantly different between the Trained and the Untrained conditions within the Uncommon or the Common Test Cases.

These results indicate the Iliad's training effects appear to be significant when the Uncommon cases were trained and tested. However, these training effects were not observed for the Common test cases. There are several potential explanations for these findings. First, the student performance on the Common test cases was quite good whether the students were Trained or Untrained (e.g., Final Diagnostic Errors: Trained 6.0%; Untrained 6.2%). These students could have been previously well trained on the subject matter relating to the common cases. This might have occurred because the students' preclinical training focuses on common conditions, and this tendency is reinforced on the clerkships. These students might have received enough training on these common conditions that the additional Iliad training was relatively ineffective. A second potential explanation is that our Common test cases were too easy. If the cases were too easy, the uniformly high student scores could prevent us from adequately distinguishing the students' performances. A third explanation might be that the dependent variables represent relatively coarse outcome measures which do not detect subtle, but important training effects. For instance, a student might arrive at and verify the correct final diagnosis, but do so in a very cost-ineffective manner. We are now developing and evaluating finer-grained assessment measures which can provide a more detailed evaluation of the work-up process.

We also observed that Iliad's training effects appeared to be significant only in the second replication of the experiment. This effect may occur because students must first gain experience with the software during the initial phase of the clerkship. The weekly support sessions in the first half of the clerkship do focus on reinforcing student use of the learning tools. Therefore, students may not reap the maximum teaching benefit from Iliad until later in the rotation (the second replication). Another potential explanation is that generalization of learning occurs across the domains, and that generalization from the first replication reinforces performance on the second replication. However, other aspects of the data (see below) make this explanation seem unlikely.

Domain Specificity

We used the correlational data within the dependent measures, across the replications, to examine whether domain specificity was observed. Previous research suggests that practitioner performance is quite domain specific.[10–12] Training or experience in one domain does not predict performance in another domain. Therefore, we would expect low correlations for scores on a particular dependent measure across different Iliad test cases. The correlational data provides support for this reasoning. For instance, the correlation between the Final Diagnostic Errors in replications one and two (different cases in each replication) was weak [$r = 0.236$; $t(92) = 2.33$, $p < 0.02$). While this correlation and the others were still statistically significant, they explain only about 6% of the shared variance between the scores [$r = 0.236$, $r^2 = 0.056$]. This shared variance is a full order of magnitude less than the correlation noted between the different measures *within* a replication. Similar results were obtained for correlations between the Posterior Probability and Average Hypothesis Score variables in replications one and two.

The small correlations within the specific dependent variables that we observed between the replications could indicate a weak generalization across case domains. However, these correlations could also be spurious. For instance, a few students might have

been relatively uninterested in Iliad. These students might have tended to perform poorly when they experienced the Trained as well as the Untrained conditions. Their uniformly poor performance would have been pooled with data from other students who performed well when Trained and less well when Untrained. This pooling could result in the small, but significant, correlations we observed. We believe this was not a major effect, because we assessed students' attitudes about Iliad following the training. In a postclerkship evaluation, most students indicated that they enjoyed using the software and rated it as equivalent to or superior than standard teaching modalities (e.g., books, lectures, rounds). However, just a few disinterested students could produce the small correlations that we documented. Whatever their cause, such correlations (up to $r = 0.30$, which would explain about 10% of the variance) are consistent with the findings of other researchers who have examined this domain generalization issue.[11]

Potential Limitations of the Research

One potential limitation of this research is that we did not have a computer-untrained control group. Instead, we provided experimental control by placing the students in different training domains. This strategy will provide adequate experimental control if training is relatively domain specific (as indicated by both previous research and the current results). There are three reasons why we adopted this strategy instead of simply using a computer-untrained control group. First, all students must become equally proficient with the computer in order to score to the best of their abilities on the test cases. If control students perform no training cases, their testing proficiency will decay after the initial orientation. Second, student reactivity to the experiment is minimized when all students receive interesting training. In early pilot studies, some students were randomized to train with a version of Iliad that did not contain learning tools. These students had a very negative reaction because they perceived that the software did not help them learn. Their negative reaction threatened to contaminate the good will we had created among the faculty as well as among the experimental students who had received a fully functional version of Iliad. Based on this experience, we concluded that medical students are more likely to accept training when it is outwardly nonmanipulative and appears beneficial. Third, we adopted apparently "equal" training for all students in order to minimize any potential "Hawthorne" or placebo effects.[19]

Another potential limitation of the research is that we did not analyze the students' exposure to real patient cases during their training. However, because the average student's actual case experience is so small during the clerkship, we do not believe this contaminated the results. At our institution, the average student is assigned to take primary responsibility for only two or three cases each week during their clerkship (they follow other cases peripherally, perhaps assisting with procedures or attending patient rounds). A previous analysis of our students' case logs indicated that fewer than 50% of the students would be expected to see more than two patients who resembled Common cases during a single clerkship. For Uncommon cases, the students' exposure to real cases is even more limited. For example, we found that substantially fewer than 1% of our students will see a real case of Addison's Disease. Even when students do see real cases which duplicate training or test cases, this experience is randomly distributed among the students in each experimental condition. This random experience with real cases would

actually be expected to increase the experimental variation in the test results (''noise'') and thus reduce our chances of finding a significant training effect.

Adopting Expert Systems Training

Based on our work, we have identified several factors we believe influence the successful adoption of an innovation such as Iliad. First, the support of key faculty must be obtained. At Utah, we have long enjoyed a tradition which supports the early clinical implementation of innovative new developments in medical informatics. For example, the Chief of Medicine and Clerkship director were strong early supporters of Iliad. While achieving the support key faculty could be difficult in other settings, we have noticed that faculty at other institutions who are now adopting Iliad tend to be pleased when someone offers to enhance the student clerkships. However, these faculty must be convinced that the new approach is meritorious before they will endorse the new training procedures.

A second key technique to promote successful adoption is to inform the trainees' clinical supervisors about the proposed training. These supervisors (e.g., faculty and residents) work with and evaluate the students on a daily basis, and must approve and support any significant commitments of student time. We adopted several strategies to convince these supervisors of the merit of the student training. For the residents, we gave ''noon conferences'' which explained Iliad and the goals of the training. We made similar presentations to the faculty at research and clinical seminars. When these clinical supervisors and teachers understand and support the training, they provide signals to the students which indicate that the training is likely to be worth their time and effort.

A third key strategy which ensures more successful adoption is to place the computers where the students perform the majority of their work. For junior students, this location should be the medical wards, not necessarily a computer learning center (even if it is relatively close by). Computers should be placed on the wards because students have many competing demands on their time, and are most likely to use the computer when it is conveniently situated and immediately available.

A fourth important strategy is to provide students with an adequate initial training and ongoing technical support. We have found that a 2- to 3-hr orientation on the first day of the rotation is adequate to get students started using the software. We supplement this initial orientation with brief (20-min) weekly meetings, following ''grand rounds.'' We also provide students with faculty advisers who conduct ''office hours'' for individual students.

A fifth and final strategy we have adopted is to *require* that all trainees use Iliad during their clerkship. To ensure that our junior students had enough time to use the software, we first reduced their lecture requirements. Because Iliad was required, we monitored each student's compliance with the training requirements and provided feedback when individual students fell behind. We found in our pilot work that students tend to focus on those activities which are clearly required their performance evaluation.

Directions for Future Research

Our current research did not find a training effect for the Common cases. However, our research employed relatively coarse outcome measures (such as the Final Diagnostic

Errors or Posterior Probability) to quantify the training effect. We are now exploring whether finer-grained assessment measures could detect important training effects for the Common cases. The measures we are now designing and evaluating will be used to assess the work-up process as well as the outcomes of the diagnostic work-up. For instance, algorithms now being evaluated in new versions of Iliad can measure the cost and information gain associated with each patient finding or test which the student orders during the work-up. In the simulation training mode, these algorithms can provide the student with tailored, formative feedback regarding the cost-effectiveness of the work-up. In the simulation-test mode, these algorithms can provide the basis for calculating dependent measures which characterize the cost-effectiveness of the student's work-up.

The current research also does not examine the durability of the training effects. We examined student performance following a relatively short (1 week) lag period after the training. The lag period was designed to be long enough to prevent students from divining the experimental design, but short enough to minimize any long-term forgetting. We are now designing a new student experiment intended to examine the durability of the learning over longer periods.

Finally, while our current research measures how student performance improves for simulated patients, it does not indicate how student performance might improve on real patients. We are now planning research to determine how patient care might be affected by simulated case training. This research will measure trainees' performance on "standardized patients." Standardized patients are actors who simulate clinical conditions. Simulated patients have been used extensively to analyze student and practitioner problem-solving performance.[12,20] Future research will also attempt to explain student performance on computerized simulated patients by conducting "stimulated recall" sessions following each student's computerized testing. These sessions are designed to permit faculty analysis of the student's test case performance.[21]

SUMMARY

We implemented a training regime for junior year medical students in which we used an expert system to provide tailored feedback on simulated patient cases which were administered in a learning mode. This regime was integrated into the University of Utah's standard internal medicine clerkship for junior year students. Our analysis of the students' performance on computerized test cases revealed a significant training effect for students who were trained in uncommon cases. These training effects were strong enough to boost the performance of these students to approximately the same level as that demonstrated by students who experienced common test cases. One important assumption of our research was that the training effects would be relatively domain specific. This assumption was confirmed by our results, as is consistent with previous research in the area.

Microcomputer-based expert systems can provide an effective method of supplementing case-based training programs for students or residents. In this study, we found that the expert system was especially effective for training students in uncommon diagnostic conditions. Students and other trainees may not receive experience with the complete set of diagnostic conditions with which they are expected to become familiar. In these cases, the expert system training could be especially effective.

ACKNOWLEDGMENTS

Michael J. Lincoln was supported in part by a grant (1 R29 LM-05260-01) from the National Library of Medicine and in part by the Veterans Administration. This work was also supported by a grant (5 RO1 LM-04604) from the National Library of Medicine, which was awarded to Homer R. Warner. The authors would like to thank especially the University of Utah Medical Class of 1991, who participated in this research; Dr. William Odell, the Chief of Medicine; and Dr. G. Richard Lee, the Clerkship Director.

REFERENCES

1. Bouhaddou, O., Lepage, E., Warner, H.R., and Warner, H.R., Jr., An approach to evaluating the completeness of a medical knowledge base. *Proceedings of the 13th Annual Symposium of Computer Applications in Medical Care*. IEEE Computer Society Press, Los Alamitos, California, pp. 110–115, 1989.
2. Ben-Said, M., Dougherty, N., Anderson, C., *et al.*, KESS, Knowledge Engineering Support System. *Proceedings of the 11th Annual Symposium of Computer Applications in Medical Care*. IEEE Computer Society Press, Los Alamitos, California, pp. 56–59, 1987.
3. Warner, H.R., *Computer-Assisted Medical Decision Making*, Academic Press, New York, 1979.
4. Yu, H., Haug, P.J., Lincoln, M.J., *et al.*, Clustered knowledge representation: Increasing the reliability of computerized expert systems. *Proceedings of the 12th Annual Symposium on Computer Applications in Medical Care*. IEEE Computer Society Press, Los Alamitos, California, pp. 126–130, 1988.
5. Lau, L.M., *Performance of a Diagnostic Expert System (Iliad) as a Tool for Quality Assurance*. (In press, 15th Annual Symposium of Computer Applications in Medical Care, IEEE Computer Society Press, Los Alamitos, California, 1991).
6. Norusis, M.J., and Jacquez, J.A., Diagnosis I. Symptom nonindependence in mathematical models for diagnosis. *Comput. Biomed. Res.* 8:156–172, 1975.
7. Norusis, M.J., and Jacquez, J.A., Diagnosis II. Symptom nonindependence in mathematical models for diagnosis. *Comput. Biomed. Res.* 8:173–188, 1975.
8. Turner, C.W., Williamson, J.W., Lincoln, M.J., *et al.*, The effects of Iliad on medical student problem solving. *Proceedings of the 14th Annual Symposium on Computer Applications in Medical Care*. IEEE Computer Society Press, Los Alamitos, California, pp. 478–482, 1990.
9. Elstein, A.S., Shulman, L.S., and Sprafka, S.A., Medical problem solving—A ten year retrospective. *Eval. Health Prof.* 13:5–36, 1990.
10. Schmidt, H.G., Norman, G.R., and Boshuizen, H.P.A., A cognitive perspective on medical expertise: Theory and implications. *Acad. Med.* 65:611–621, 1990.
11. Norman, G.R., Tugwell, P., Feightner, J.W., *et al.*, Knowledge and clinical problem-solving. *Med. Educ.* 19:344–356, 1990.
12. Norman, G.R., and Tugwell, P., A comparison of resident performance on real and simulated patients. *Med. Educ.* 57:708–715, 1982.
13. Norcini, J.J., Swanson, D.B., Grosso, L.J., and Webster, G.D., A comparison of several methods for scoring patient management problems. *Proceedings of the 22nd Conference on Research in Medical Education*. Washington DC; 1983.
14. Skakun, E.N., Taylor, W.C., Wilson, D., Taylor, T., Grace, M., and Fincham, S.C., Preliminary investigation of computerized patient management problems in relation to other examinations. *Educ. Psychological Measure*. 39:303–310, 1979.
15. Kassirer, J.P., and Kopelman, R.I., Cognitive errors in diagnosis: Instantiation, classification, and consequences. *Am. J. Med.* 86:433–441, 1989.
16. Tversky, A., and Kahneman, D., Judgment under uncertainty. *Science* 185:1124–1131, 1974.
17. Cundick, R., Turner, C.W., Lincoln, M.J., *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*. IEEE Computer Society Press, Los Alamitos, California, pp. 902–906, 1989.
18. Keppel, G., *Design and Analysis: A Researcher's Handbook* (second edition), Prentice-Hall, Englewood Cliffs, New Jersey, 1982, pp. 160–165.
19. Cook, T.D., and Campbell, D.T., *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Rand McNally, Chicago, 1979.
20. Norman, G.R., Neufeld, V.R., Walsh, A., *et al.*, Measuring physician performances by using simulated patients. *J. Med. Educ.* 60:925–934, 1985.
21. Piemme, T.E., Yamamoto, W.S., Tidball, C.S., *et al.*, Medical informatics in the curriculum at George Washington University. *Proceedings of the 14th Annual Symposium on Computer Applications in Medical Care*. IEEE Computer Society Press, Los Alamitos, California, pp. 478–482, 1990.