The Influence of a Decision Support System on the Differential Diagnosis of Medical Practitioners at Three Levels of Training

by Gwendolyn C. Murphy, Ph.D.[a], Charles P. Friedman, Ph.D.[a], Arthur S. Elstein, Ph.D.[b], and Fredric M. Wolf, Ph.D.[c], Tom Miller, M.D.[a], Judith G. Miller, M.I.L.S.[c]
[a]University of North Carolina at Chapel Hill, Chapel Hill, N.C.
[b]University of Illinois at Chicago, Chicago, IL.
[c]University of Michigan at Ann Arbor, Ann Arbor, MI

As computer-based diagnostic consultation systems become available, their influence and usefulness need to be evaluated. This report, based on partial data from a larger study, examines the influence of Iliad, a diagnostic consultation system, on the differential diagnosis of fourth year medical students, residents in medicine, and attendings in general internal medicine. Our results show that when faced with difficult diagnostic cases, medical students add significantly more diagnoses from Iliad's differential than do residents or attendings. However, the quality of Iliad's diagnostic advice in terms of the presence of the correct diagnosis, is no better for consultations done by students or residents compared to attendings.

INTRODUCTION

Medical diagnosis is a complex problem-solving process that often requires judgments based on incomplete knowledge. Over the past several decades computer-based diagnostic consultation systems have been developed to aid physicians in this process.[1] As these systems become available to the medical community, evaluation of their influence and usefulness becomes necessary.[2-5] According to Lundsgaarde[6], systems should be evaluated for efficiency, practical feasibility, user acceptability, impact, and cost-effectiveness. Clinical trials should be by unbiased observers who have no direct personal or professional stake in the evaluation outcome. Studies up to this point generally have involved evaluation of cases by an expert user of the system.[7] In one study that did involve students, the students worked in groups using Iliad to solve clinical cases from New England Journal of Medicine; the only measure was whether the correct diagnosis appeared in the top ten diagnoses on Iliad's list.[8] The present study compares each subject's pre- and post-consultation lists of diagnostic hypotheses to assess the impact of the advice from Iliad on the subjects' lists. The subjects are non-expert users at three levels of medical training.

Iliad is a decision support system (DSS) that was designed to teach medical decision-making and to provide assistance in differential diagnosis across the domain of internal medicine. The system was developed at the University of Utah, and much of its knowledge base derives from patients admitted to University of Utah Medical Center and the Latter Day Saints Hospital. Iliad uses Bayes theorem, conditional probabilities, and clusters of conditionally independent findings that are governed by Boolean logic. When findings are entered into the system, a sequential Bayesian inference algorithm generates a ranked list of diagnoses and assigns a posterior probability to each one.

Research Questions

This study explores the influence of a Iliad on the diagnostic hypotheses generated by medical clinicians at three levels of experience. Specifically:

1) Are medical students more influenced (in terms of number of diagnoses added, confirmed, or moved) by Iliad's differential diagnosis list than residents and attending physicians?

2) Does the quality of the subject's differential diagnosis (in terms of the presence of the correct diagnosis) vary depending on the subject's level of medical experience?

3) Does the quality of Iliad's advice (in terms of the presence or rank of the correct diagnosis) depend on the subject's level of medical experience?

METHODS

The data for this report were from a larger, three-site study involving cases developed at three sites with subjects from all three sites. The sites are: University of Illinois at Chicago, University of Michigan at Ann Arbor, and University of North Carolina at Chapel Hill. This report uses a portion of the data collected from subjects at the University of North Carolina.

Subjects

The subjects were 12 fourth-year medical students; 12 second- and third-year medicine residents; and nine attending physicians from UNC. The attendings were faculty members in general internal medicine, with at

0195—4210/96/$5.00

least two years out of residency and who saw patients at least two half-days per week. All subjects were volunteers.

Cases

Thirty-six diagnostically challenging cases (12 from each site) were selected and abstracted by a general internist at each of the three participating institutions (UIC, UM, UNC). For all cases, the correct diagnosis is in Iliad's knowledge base. The correct diagnoses for each case was known from definitive tests, but the case author was instructed not to include the "gold standard" test that would verify the correct diagnosis. This was done so that the cases would be challenging. Each case was abstracted from a patient chart without consideration to how well it would fit the disease frames of Iliad. The cases were field-tested with subjects from each experience level at each institution. To make the task of reasonable length, the 36 cases were divided into four clusters with nine cases each, based on the body system involved, the institution that authored the case, and estimates of the difficulty of each case (which were done by the clinicians at each site).

Training

Each subject was given individual instruction for 1.5-2 hours on use of Iliad. Three elementary cases were used in this instruction and a check-list of system features was utilized to standardize the training. The instruction for Iliad included the use of the consultation and critique modes, and showed the subject the use of all the floating menu and header menu items.

The Decision Support System

The study employed Iliad 4.2, developed in the Department of Medical Informatics at the University of Utah. The current version has over 920 disease frames (diagnoses and intermediate decisions) and over 10,000 findings in its knowledge base.[9]

Protocol

Each subject completed one cluster of nine cases. The order of case presentation was random except that the two easiest cases were always presented first. For each of the nine cases, the subject read the case and listed up to six diagnostic hypotheses, in descending order of likelihood. Then, the subject reworked the case with access to Iliad. The subject was free to choose any of the features of Iliad and to enter as may or as few findings as desired. S/he decided when the end-point of the consultation was reached, at which point s/he provided a second list of hypotheses in descending order of likelihood. A research assistant familiar with the program was always immediately available to the subject to answer procedural questions about the program.

Design

Level of training is the independent variable. For each case, the differential diagnosis was compared from the pre-consult, post-consult, and Iliad advice lists. The following dependent variables were analyzed for each subject :

Diagnoses Added. The number of diagnoses not on the subject's pre-consultation list, but on Iliad's list and on the subject's post-consultation list.

Diagnoses Confirmed. The number of diagnoses on the subject's pre-consultation list, on Iliad's list, and on the post-consultation list.

Diagnoses Moved. The number of diagnoses on the pre-consultation list and raised or lowered in rank on the post-consultation list in concert with a high or low rank on Iliad's list.

Pre-consultation correct diagnosis. The number of cases for which the correct or near-correct diagnosis is on the pre-consultation list. Where "near correct" is very close cousin of the correct diagnosis (example: Polymyalgia Rheumatica for Temporal Arteritis) as judged by the clinical investigator (TM).

Post-consultation correct diagnosis. The number of cases for which the correct or near-correct diagnosis is on the post-consultation list.

Iliad correct diagnosis displayed. The number of cases for which the correct or near-correct diagnosis is anywhere on Iliad's list.

Iliad correct diagnosis in top six. The number of cases for which the correct or near-correct diagnosis is in the top six diagnoses of Iliad's list.

The first three variables (diagnoses added, confirmed, and moved) relate to the first research question about the influence of Iliad's diagnostic list on the subject's post-consultation list. The next two variables (correct diagnosis pre-consultation and correct diagnosis post-consult) relate to the second research question about the quality of the subject's diagnostic list pre- and post-consultation. The last two variables (Iliad correct diagnosis, and Iliad correct diagnosis top six) relate to the third research question regarding the quality of the diagnostic advice. In this report we have analyzed the data from 33 subjects with nine cases each for a total of 297 cases. The data were analyzed by group with the subject as the unit of analysis, using analysis of variance, with a posteriori tests for those variables which showed significant group differences.

Table 1
Means and Statistical Analysis of the Dependent Variables for Three Groups of Clinicians

Variables Analyzed (per 9 cases)	MS4 mean ± s.e.m. n=12	Resident mean ± s.e.m. n=12	Attending mean ± s.e.m. n=9	Overall ANOVA
Diagnoses Added	21.58*† ± 3.51	7.67 ± 1.66	4.89 ± 0.63	F(2,30)=13.25 (p=.0001)
Diagnoses Confirmed	1.83* ± 0.47	5.83 ± 0.71	3.78 ± 0.57	F(2,30)=12.11 (p=.0001)
Diagnoses Moved	2.75 ± 0.86	3.83 ± 0.68	2.44 ± 0.56	F(2,30)=0.97 (p=0.39) N.S.
Pre-consultation Correct diagnosis	2.00*† ±0.35	4.50 ± 0.47	4.89 ± 0.65	F(2,30)=10.89 (p=.0003)
Post-consultation Correct diagnosis	2.75*† ± 0.39	4.75 ± 0.58	5.33 ± 0.62	F(2,30)=6.55 (p=.004)
Iliad Correct diagnosis	3.67 ± 0.36	4.17 ± 0.30	4.44 ± 0.53	F(2,30)=1.03 (p=.37) N.S.
Iliad Top 6 Correct diagnosis	1.67 ± 0.26	1.58 ± 0.29	2.22 ± 0.52	F(2,30)=0.91 (p=.41) N.S.

*MS4 significantly different from Residents
†MS4 significantly different from Attendings
N.S.: not significant

RESULTS

All subjects completed the protocol of nine cases. Table 1 gives the means for each variable studied.

Iliad Influence

Our first research question was whether medical students are more influenced (in terms of number of diagnoses added, confirmed, or moved) by Iliad's differential diagnosis list than residents and attending physicians. On average, the students added 21.58 diagnoses per nine cases from Iliad to their post-consultation differential diagnosis, while the residents and attendings added 7.67 and 4.89 diagnoses respectively. The residents had confirmation of their diagnoses by Iliad an average of 5.83 diagnoses per nine cases, while the attendings had 3.78 and the students had 1.83 diagnoses per nine cases. An ANOVA with follow-up analyses showed that the medical students were significantly different from the residents and attendings for the number of diagnoses added from the Iliad list and significantly different from the residents, but not from the attendings on the number of diagnoses that were confirmed by the Iliad list. There was no significant difference between the groups on the number of diagnoses moved.

Quality of pre- and post-consultation differential diagnosis lists

To evaluate the extent that the quality of the pre- and post-consultation differential diagnosis lists (in terms of the presence of the correct diagnosis) depend on the subject's level of medical experience, we tallied the number of times the subject had the correct or near correct diagnosis on the differential list pre-consultation and post-consultation. The students had the correct or near-correct diagnosis on the pre-consultation list in 2.00 of the nine cases, while the residents and attendings were correct in 4.50 and 4.89 cases respectively. The students had the correct or near-correct diagnosis on the post-consultation list in 2.75 of nine cases, while the residents and attendings had 4.75 and 5.33 respectively. The correct or near correct diagnosis on both the pre- and post-consultation lists were significantly different between the medical students and both the residents and attendings, but there was no significant difference between the residents and attendings on either of these measures.

Quality of Iliad's diagnostic advice

Our third question asks to what extent the quality of Iliad's advice (in terms of the presence or rank of the correct diagnosis) depends on the subject's level of medical experience. To answer this we compared the presence of the correct or near-correct diagnosis on Iliad's differential diagnosis list, and within the top six diagnoses of Iliad's list. There was no significant difference between the groups in the number of cases in which Iliad listed the correct or near-correct

diagnosis either on its entire list, nor within its top 6 diagnoses.

DISCUSSION

This study explored the effects of a decision support system, Iliad, on the differential diagnosis of medical practitioners at three levels of training. It appears that medical students are heavily influenced by the diagnostic list that Iliad produces. In addition, these data show that the differential list produced by experienced practitioners is not significantly different with respect to the presence of the correct diagnosis than the lists produced by novices.

With the present data we cannot prove that Iliad caused the changes in the subject's differential diagnosis lists. To do so we would have to have two control groups: one that read the cases twice and gave a differential list, but did not use Iliad for consultation; and a second group that read the cases and entered information into Iliad but did not get any advice from Iliad. Both of these tasks would be tedious.

Our data do show that the students added a large number of diagnoses to their post-consultation lists that had not been on the their pre-consultation lists, but were on Iliad's lists. The number added by students was significantly higher than the number added by residents or attendings. On average, students added 2.40 diagnoses per case, residents added 0.85 diagnoses per case, and attendings added 0.54 diagnoses per case. These differences are probably due to both the difficulty of the cases, and to the inability on the part of students to assess the suitability of a diagnosis suggested by Iliad.

Residents had the most diagnoses from their pre-consultation list confirmed by Iliad's list, with attendings having the next most confirmations, and medical students having the least. We can speculate on the reasons for these differences. They may be due to the "intermediate effect" found in expert-novice studies, in which experts have greater knowledge but the intermediate level subjects (residents) may be more thorough in evaluating data which helps them perform better on measured tasks.[10] The residents may feel more of a challenge to get the Iliad to verify their initial list. The attendings may not feel the same challenge, or may view confirmation as an impossible task. The students may not have either the confidence in their original list nor the skill to get the system to verify their pre-consultation list. It is also possible that there was no deliberate strategy on the part of the subjects to confirm their original diagnoses.

Change in and of itself is not necessarily the desired outcome. What is important is whether the change results in a more correct diagnostic list. As one would expect, the students have fewer correct or near-correct diagnoses on both their pre- and post-consultation lists compared to the residents and attendings. These differences were statistically significant.

There was no significant difference between the groups on the quality of diagnostic advice from Iliad. This is an unexpected finding; we reasoned that the attending physicians and residents would be able to provide more insightful input and thereby get a more correct differential diagnosis list from the DSS. However, these were difficult cases and for some of them the case information did not match the disease profile very closely. As a result, entering only the significant information resulted in a differential diagnosis that was no better than the differential obtained when nearly every finding was entered.

Overall, Iliad had the correct diagnosis on its list in only 45% of the cases and in its top six diagnoses in only 20% of cases. This level of diagnostic accuracy is about the same as that reported by both Berner and Elstein.[5,11] Iliad appears to have had very little influence over whether the subject had the correct diagnosis on the post-consultation list. This may be partially due to a tendency of clinicians to adhere to their original list of diagnoses unless a diagnosis is clearly disconfirmed; a phenomenon called conservatism or anchoring, which has been studied by psychologists for decades.[12]

The cases used in this study were derived from actual patients and may be reasonably representative of patients who pose diagnostic dilemmas to practicing general internists. At the same time however, the cases were selected to be diagnostically challenging for attending physicians in general internal medicine. A clinical case is diagnostically challenging if it is an unusual presentation of a common disease, a typical presentation of an uncommon disease, or a patient presenting with multiple problems. Our cases came from all of these categories. As a result, these cases were extremely difficult for the fourth year medical students. In addition, the definitive diagnostic information was deliberately not included in the case write-up. While both Iliad and the subjects would have had greater accuracy if the definitive test was included, the case would no longer be challenging and the clinician would presumably no longer need a consultation if the conclusive evidence was included.

This study was designed to be as close to an actual clinical context as is possible in a controlled, experimental study. A pre- and post-test model was

used so that the change in the subject's diagnostic hypotheses in response to the use of a DSS could be assessed.

As the remainder of the data are collected we expect to be able to assess further the impact of the DSS on the subject's list of diagnostic hypotheses. Subjects from different geographical locations will add a dimension that will broaden the generalizability of our results.

CONCLUSIONS

As computer-based diagnostic consultation systems become commercially available to the medical community, their influence and usefulness needs to be evaluated by unbiased researchers using subjects who are similar to the intended user population. This study did that using researchers who have no financial stake in the outcome, and subjects who are generalists with varying levels of training (fourth-year medical students, second- and third-year residents in medicine, and attending physicians in general internal medicine). The case abstracts used were diagnostically difficult so that they would be similar to cases for which a clinician might use a diagnostic consultation. In this preliminary report, we found that medical students add significantly more of Iliad's diagnoses to their post-consultation lists than residents or attending physicians. Residents have a significantly higher number of diagnoses that are confirmed by Iliad compared to medical students, but there is no difference between the three groups in the number of diagnoses that are changed in rank from the pre-consultation to the post-consultation list.

As one would expect, the medical students had fewer correct diagnoses compared to the other two groups. However, the lists produced by Iliad were not significantly different in terms of the number of correct diagnoses when we compare the three groups. As reported elsewhere, the ability of the decision support system to find the correct diagnosis, given difficult cases and incomplete data, is limited.[5,11]

In this study, the system's advice is viewed as a product of an interaction between the user, who selectively enters positive and negative findings, and the knowledge/knowledge processing routines of the diagnostic support system.

Data are now being collected from a larger sample of subjects which will deepen and extend these data. An improved method of scoring diagnostic accuracy, which incorporates both quality and rank of the diagnoses, is being developed. Additional variables for next steps in the work-up and the perceived need for a diagnostic consultation will be included in that analysis.

Acknowledgments: The authors would like to thank the following people for their assistance in this research: Paul Heckerling, M.D., Paul Fine, M.D., Kevin Biolsi, Xiao Mei, Amy Capitano, Sema Barlas Macy Ng, Keith Cogdill, Kevin Biolsi, Tim Franz, and David Potts.

References

1. Miller RA. Medical diagnostic decision support systems--past, present, and future: A threaded bibliography and brief commentary. *JAMIA*, 1994:1:8-27.
2. Lincoln MJ, Turner CW, Haug PJ, et. al. Iliad training enhances medical students' diagnostic skills. *J Med Sys* 1991:15(1):93-110.
3. Sumner W II, Shultz EK. Expert systems and expert behavior. *J Med Sys* 1992:16(5):183-93.
4. Turner CW, Lincoln MJ, Haug P, et. al. Iliad training effects: A cognitive model and empirical findings. *AMIA* 1992: 68-72.
5. Berner ES, Webster GD, Sugarman AA, et. al. Performance of four computer-based diagnostic systems. *NEJM* 1994:330:1792-6.
6. Lundsgaarde HP. Evaluating medical expert systems. *Soc Sci Med* 1987:24(10):805-19.
7. Bankowitz RA, McNeil MA, Challinor SM, et. al. A computer-assisted medical diagnostic consultation service: Implementation and prospective evaluation of a prototype. *Ann Int Med* 1989:110:824-32.
8. Gozum, M.E. Emulating cognitive diagnostic skills without clinical experience. *Symposium on Computer Applications in Medical Care,* 18:991, 1994.
9. Applied Medical Informatics. Iliad 4.2 User Manual. Salt Lake City: Applied Medical Informatics, 1994.
10. Kaufman DR, Patel VL. Problem solving in the clinical interview: A cognitive analysis of the performance of physicians, residents, and students. *Teach Learn Med* 1991:3:6-14.
11. Elstein AS, Friedman CP, Wolf FM, et.al. Effects of a decision support system on the diagnostic accuracy of users: A preliminary report. Submitted JAMIA 1996.
12. Edwards W. Conservatism in human information processing. In B. Kleinmuntz (ed.), *Formal representation of human judgment.* N.Y.: Wiley 1968.