Assessing the Behavioral Impact of a Diagnostic Decision Support System

Yu-Chuan Li, M.D., Ph.D., Taipei Medical College, Peter J. Haug, M.D., LDS Hospital/University of Utah, Michael J. Lincoln, M.D., University of Utah, Charles W. Turner, Ph.D., University of Utah, T. Allan Pryor, Ph.D., IHC/University of Utah, Homer H. Warner, M.D., Ph.D., University of Utah

This paper describes a prototype for research to evaluate the impact of diagnostic decision support systems on the behavior of physicians. Several indices that can be used to quantify the magnitude of impact are proposed. A large medical diagnostic knowledge base in internal medicine (the Iliad knowledge base) was used in this evaluation. The impact on behavior when different inference models are run against this knowledge base is evaluated for two different case domains and physician's specialties.

INTRODUCTION

The potential benefit of applying diagnostic decision support systems in medical care has been much discussed [1-3] in the past. But before any such system can be applied in a clinical setting, it must be tested and evaluated thoroughly. Evaluation of a diagnostic decision support system is a complex issue [4-6]. Not only must the accuracy of the computer-generated diagnoses be ensured, the more delicate issue of impact on physician's diagnostic decision should also be explored.

This study proposes several indices that can be used to measure the degree of change in diagnostic decision making by the physician under the guidance of a medical diagnostic support tool. We applied these indices to different implementations of a diagnostic support system called Iliad, which has a large knowledge base (KB) comprised of 2300 diseases and intermediate diagnoses and 9000 relevant findings in internal medicine [7]. Using this KB, we devised and prototyped a scheme to evaluate system impact in the context of three variables: the inference model used, the medical domain from which the test cases are taken, and the specialty of the subject physicians. Details of the variables and study design are described in the following sections.

METHODS

The Inference Models

Two inference models were used in this study. One is the standard algorithm used in most versions of Iliad. We refer to this as Iliad-Knowledge Representation (Iliad-KR) in this paper [8]. The Iliad-KR model runs under the normal consultation mode of Iliad system [7]. This model is referred to as ILD in the following text. The second model uses a relatively new knowledge representation called Bayesian networks, which, in brief, are based on a graphical representation of probabilistic dependencies [8,9]. The Bayesian network model is referred to as BYN in the following text. By using a set of mathematical algorithms, we were able to implement a computer program that reads directly the Iliad KB and transforms it into a Bayesian network.

Patient Selection and Case Domains

The Iliad KB encompasses most of the subspecialties in the internal medicine domain. We chose to concentrate on cardiology and gastroenterology since the Iliad KB is quite mature in these areas. The patient selection process was begun by downloading a population of 7855 patient-profiles from the database that is a part of the HELP medical information system [10, 11]. Out of this population, 1251 cases involved cardiovascular (CV) diagnoses and 609 cases involved gastroenterology (GI) diagnoses. We applied a set of criteria based on age and length of stay to exclude newborns and pediatric patients as well as admission for scheduled procedures and chronically ill patients. The remaining cases were then rated independently by two physicians on their suitability. Cases with enough evidence to propose a reasonably circumscribed differential diagnostic list but without adequate evidence to determine a definite diagnoses were deemed most suitable. The ratings of the two reviewers were averaged. The ten cases with highest average ratings were selected from each domain as the evaluation cases.

Based on emergency room reports and admission summaries, patient information was abstracted into patient vignettes. These were presented to four different physicians who were required to perform case-evaluation. All the information contained in the vignettes were translated into the terminology and coding scheme used in Iliad and were submitted to the two test versions of Iliad described above.

The Evaluation Processes

After identifying the final 10 CV and 10 GI cases for evaluation, two cardiologists (labeled as CV DOC-1

0195-4210/95/$5.00

and CV DOC-2) and two gastroenterologists (labeled as GI DOC-1 and GI DOC-2) from among the attending physicians at the University of Utah Medical Center were recruited to participate in this study. They were provided with copies of the vignettes, forms for recording diagnostic evaluations, and the output from one of the two diagnostic models. Each subspecialist evaluated all 20 cases (10 CV and 10 GI) based on the following instructions:

(1) First, read the patient vignette.
(2) Second, write down a differential diagnostic list and record a percentage representing the likelihood for each diagnosis.
(3) Third, read the computer's suggested differential diagnoses generated by one of the inference models (BYN or ILD). Only one differential list was included in each packet.
(4) Fourth, in light of the computer-generated suggestions, revise the differential diagnostic list and the set of likelihoods proposed in step two. The physicians were free to add/drop diagnoses or change the estimated likelihoods.
(5) Finally, rate the usefulness of the differential diagnostic list provided by the computer on a scale from one to five. In this scale, one implies the list is very misleading, three implies neutral and five implies very useful.

In step 3, a physician only received one computer-generated diagnosis list for each case. This list came either from the BYN or the ILD model. Two physicians from each domain were involved, so that we were able to measure the impact of both inference models within each domain. In addition, the participating physicians were exposed to cases both within and outside their own specialty.

Defining the Behavioral Impact

Two kinds of impact on physicians' diagnostic behavior were assessed in this study. We called these "sheer impact" and the "positive impact" respectively. The *sheer impact* represents the magnitude of change in diagnostic decision making induced by the computer-generated advice. This impact will be represented by scores derived from the change in the physicians' differential diagnoses before and after exposure to the computer's advice. The *positive impact* identifies whether the impact is in the right direction, that is, does the differential diagnostic list move closer to the gold standard.

We used the differential diagnosis lists produced by the subspecialist when they evaluated cases within their subspecialty as the source of the gold standard. The diagnosis list generated by the CV DOCs for CV cases were aggregated by averaging the likelihood proposed for each diagnosis. The resultant combined list was used as the gold standard for CV cases. Similar procedures were applied to each GI case to obtain its gold standard.

A score we call DxNum was used to represent the sheer impact of the computer's diagnostic advice. DxNum was obtained by observing the difference between the physicians' differential diagnosis lists before and after they saw the computer's advice. The number of diagnoses changed (between the two lists) consistent with the computer's advice is defined as DxNum. A diagnosis was considered changed if it appeared in only one of the two lists or the probability assigned to it differed in the two lists. However it was only counted in DxNum if the change was consistent with the computer-generated diagnosis list and the change was greater than 5%. DxNum is a discrete number that ranges from 0 to the number of diagnoses in the union of the Before-list and the After-list.

A score named DxGold was calculated to represent the positive impact. To create this score we defined two intermediate values, DxCon and DxIncon. DxCon is identical to DxNum except that it counts the number of changed diagnoses consistent with the gold standard; whereas the DxIncon counts the number of changed diagnoses inconsistent with the gold standard. DxGold is obtained by subtracting DxIncon from DxCon, i.e., DxGold = DxCon - DxIncon. DxGold is a positive integer when the computer's advice generates more consistent changes in the physicians second diagnostic list than inconsistent ones. If fewer changes in the second list are consistent with the gold standard, DxGold is negative.

Perceived Usefulness

In addition to the objective impact measured by these ad hoc scores, physicians' subjective ratings of the usefulness of the computer's advice were also analyzed in this study. This rating is referred to as "perceived usefulness" in following sections.

Statistical Analyses and Study Design

This study was conducted under a mixed design ANOVA. Three independent variables were involved in the analysis, namely, inference model (INF-MOD), case domains (CASE-DOM) and physicians' specialty (PHY-SPEC). Each independent variable has two levels (see Table 1).

Table 1. Levels of the independent variables.

Independent Variables	Levels
INF-MOD	BYN, ILD
PHY-SPEC	Cardiologist, Gastroenterologist
CASE-DOM	CV cases, GI cases

The DxNum and the DxGold scores were used as dependent variables in this analysis to measure the sheer impact and the positive impact respectively.

Correlation coefficient between the indices of impact and the perceived usefulness was also analyzed to investigate the relation between subjective rating and objective scores.

RESULTS

In response to the patient vignettes, the four participating physicians gave an average of 4.4 differential diagnoses to each case. No significant difference was found in the number of diagnoses made by the physicians before and after they saw the computer-generated diagnostic suggestions.

Effect of the Sheer Impact

Using the DxNum score as the dependent variable, physicians' specialty (PHY-SPEC) was the only significant main effect ($P < 0.0005$) where the mean DxNum score was 1.7 for the cardiologists and 0.475 for the gastroenterologists. No significant difference was found on either CASE-DOM or INF-MOD in terms of main effect. The only significant interaction between independent variables was PHY-SPEC × CASE-DOM ($P < 0.005$).

Effect of the Positive Impact

Using the DxGold score as the dependent variable, the difference between inference models (INF-MOD) was highly significant ($P < 0.005$) where the mean DxGold score was 0.750 for the BYN model and -0.450 for the ILD model. No significant difference was found on CASE-DOM in terms of main effect. The interaction INF-MOD × CASE-DOM was significant in this analysis ($P < 0.05$). The interaction bar chart is shown in Figure 1. This interaction suggests that the pattern of positive impact imposed by BYN and ILD is significantly different in different case domains. The interaction bar chart shows an apparent negative influence by the ILD model when physicians were evaluating cases in the GI domain.

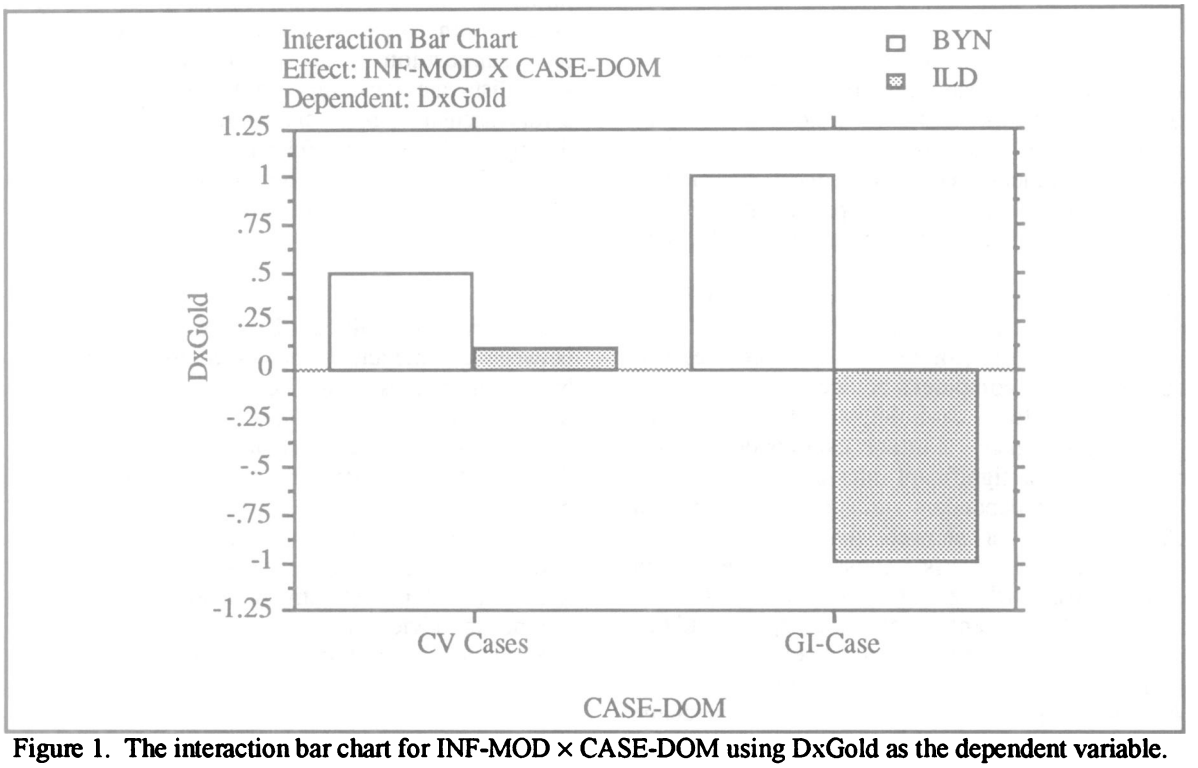

Figure 1. The interaction bar chart for INF-MOD × CASE-DOM using DxGold as the dependent variable.

Correlation between the Impact Scores and Perceived Usefulness

Although no significant correlation was found between DxNum and the perceived usefulness rated by each physician, the correlation between DxGold and the perceived usefulness produced quite intriguing results. The R-square of DxGold and the perceived usefulness was 0.174 ($P < 0.001$). However, the slope of the correlation was negative. Further investigation of the data revealed a strong negative correlation between the index of positive impact and the perceived usefulness in the GI cases, whereas those in the CV cases were not significant.

DISCUSSION

The study described here must be viewed as a prototype. Analyzing four physicians allows us to test the usability and potential sensitivity of the metrics chosen to evaluate the different inference models and sub-specialties, but does not allow us to generalize broadly from the results. Within these constraints, the results reported here increase our confidence that the impact of computerized diagnostic systems on physicians decision making can be measured.

Two aspects of the impact on physicians' diagnostic decision making by a decision support system were assessed in this study: the sheer impact, which indicates the pure change of diagnostic opinion before and after the physicians saw the computer's advice; and the positive impact, which goes one step further to discriminate "good" from "bad" influence associated with the inference models by using the sub-specialists' differential list as the gold standard.

In the assessment of the sheer impact, no significant difference was found between the Bayesian network and the Iliad-KR inference model. This suggests that the BYN model promoted as much change of diagnostic opinion as the ILD model, that is, the physicians did not perceive one as more convincing than the other in

giving diagnostic advice. However, a significant difference between the cardiologists and gastroenterologists (the PHY-SPEC variable) was observed. The computer's advice stimulated more change in the cardiologists than in the gastroenterologists. Further examination of the data revealed a significant interaction ($P < 0.005$) between the independent variable PHY-SPEC and CASE-DOM that showed that the computer's diagnostic advice is most influential on cardiologists working on GI cases. It seems likely that cardiologists tend to rely on computer's advice more than gastroenterologists when doing cases outside their specialty. This result may be attributed to the observation that internal medicine sub-specialists frequently see patients with cardiac disease independent of their subspecialty due to the high frequency of cardiovascular patients [12-14]. This experience makes them more confident in the cardiovascular domain.

In the assessment of the positive impact, a significant difference between inference models (INF-MOD) was observed. This would suggest that, in this limited study, the BYN model imposed more positive diagnostic impact than the original ILD model. A significant interaction was also detected between inference models and the case domains (see Figure 1). This interaction indicates a difference between the patterns of impact from the alternate inference models in different case domains. The difference is more prominent when evaluating GI cases. Here the Bayesian network model is associated with a greater than one unit increase in the DxGold score while the Iliad-KR model appears to decrease the DxGold.

These results suggested that the BYN model may have had a positive influence on these physicians; the DxGold score suggested that the BYN model contributed to about one additional correct diagnosis in GI cases. Considering that only an average of 4.4 differential diagnoses were listed on each case, one more appropriate diagnosis may be a valuable addition. On the other hand, the ILD model tended to have minimal to negative influence on physicians receiving its advice especially in GI cases. Again, the data supported the supposition that gastroenterology sub-specialists are generally more knowledgeable in cardiovascular cases, and thus were able to distinguish between useful and misleading advice. In the GI domain, cardiologists may be confused by the advice from the ILD model and made questionable decisions.

We observed a significant but negative correlation between DxGold and the perceived usefulness rated by the physicians. This negative correlation was not observed when correlating the index of sheer impact to the perceived usefulness, however. Since all the cases done by the physicians in the positive impact study were outside their specialty (the cases within each subspecialty were used to define the gold standard), this finding indicates that they did not effectively distinguish helpful diagnostic advice from misleading advice when dealing with unfamiliar case domains.

It is certainly not trivial to evaluate a diagnostic decision support system. We emphasize that the behavioral impact should be measured in addition to theoretical diagnostic accuracy. The study presents a prototypic framework that can be used to assess the impact of medical diagnostic support systems on physicians; it also demonstrates potential indices that can be used to measure the magnitude and direction of an expert system's impact.

*This publication was supported in part by grant number 5 R01 LM05323 from the National Library of Medicine.

References

1. Miller RA. Pople HEJ, Myers JD. Internist-I: An experimental computer-based diagnostic consultation for general internal medicine. N Engl J Med 1982;307:468-76.
2. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplan: An evolving diagnostic decision-support system. JAMA 1987;258:67-74.
3. Warner HR Jr. Iliad - Moving medical decision-making into new frontiers. In: Proceedings of International Symposium of Medical Informatics and Education. Salamon R, Protti D, Moehr J. eds. University of Victoria, B.C., Canada, 1989:267-70.
4. Habbema JDF, Hilden J, Bjerregaard B. The measurement of performance in probabilistic diagnosis (parts 1, 2, and 3). Meth Inf Med 1978; 17:217-46.
5. Bankowiz RA, McNeil MA, Challinor SM, Miller RA. Effect of a computer-assisted general medicine diagnostic consultation service on housestaff diagnostic strategy. Meth Inf Med 1989; 28:352-6.
6. Li YC, Haug PJ. Evaluating the quality of a probabilistic diagnostic system using different inferencing strategies. In: Proceedings of the 17th Symposium on Computer Applications in Medical Care (SCAMC), Washington DC, 1993:471-77.
7. Warner HR, Haug PJ, Bouhaddou O, Lincoln MJ, Warner HRJ, Sorenson D, Williamson JW, Fan C. Iliad as an expert consultant to teach differential diagnosis. In: Proceedings of the 12th Symposium on Computer Applications in Medical Care (SCAMC), IEEE Computer Society Press 1988:371-376.
8. Li YC, Haug PJ, Warner HR. Automated transformation of probabilistic knowledge for a medical diagnosis system. In: Proceedings of the 18th Symposium on Computer Applications in Medical Care (SCAMC), Washington DC, 1994:765-9.
9. Pearl 1988. Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufman, San Mateo, CA, 1988.
10. Warner HR, Olmsted CM, Rutherford BD. HELP - A program for medical decision making. Comp Biomed Res 1972; 5:65-74.
11. Kuperman GJ, Gardner RM, Pryor TA. The HELP System. Springer-Verlag New York, Inc. 1991.
12. Schmidt HG, Norman GR, Boshuizen HP. A cognitive perspective on medical expertise: theory and implication. Acad Med. 1990;65:611-21.
13. Voytavich AE, Rippey RM, Suffredini A. Premature conclusion in diagnostic reasoning. J Med Educ 1985;60:302-7.
14. Elstein AS, Shulman LS, Sprafka SA. Medical problem solving: a ten year retrospective. Eval Health Prof 1990;112:221-226.