DECISION MAKING AND LABORATORY TEST UTILIZATION: EXPERT-NOVICE DIFFERENCES

Stewart A. Haight, M.S., Donald P. Connelly, M.D., Ph.D., and Lael C. Gatewood, Ph.D.,* and M. Desmond Burke, M.D.**

*Division of Health Computer Sciences, University of Minnesota
**Department of Pathology, SUNY at Stonybrook

ABSTRACT

This study investigates the medical decision making process in both expert and novice physicians in an attempt to identify specific weaknesses in this decision making process which might be targeted for medical decision support. Two groups of physicians, practicing gastroenterologists and third and fourth year medical students, were given simulated patient management problems in a paper-and-pencil format. The two problems used were both from the same medical domain, liver disease, and consisted of an initial clinical scenario followed by two opportunities to order and obtain laboratory test results. The results were analyzed using a hypothetico-deductive model of decision making as a basis for evaluation. It is assumed that weaknesses of the novices relative to the experts in one of the four primary decision making phases described by this model may indicate a possible need for decision support.

INTRODUCTION

In spite of the large amount of effort that has been put into the development of medical decision support systems (1, 2, 3), few investigators have addressed the question of what form of decision support is best in a given clinical problem solving situation. All too often the selection of a decision support technique has been based on conjecture and intuition regarding some perceived weakness in the physician's decision-making process. The significance of this is that while a decision support technique may prove useful in a given clinical situation, the underlying mechanism of that success is not known. Consequently there can be no systematic means of determining whether or not this is the best possible mechanism for that particular situation or if the mechanism can be effectively applied to other types of problems. The underlying difficulty has been a lack of a valid theoretical framework for identifying the weaknesses the support mechanisms are intended to ameliorate.

Recent studies of the medical decision-making process may provide a framework for classifying basic weakness in this process. These studies indicate that a hypothetico-deductive model of decision making provides a reasonably accurate description of the medical decision-making process (4, 5, 6). According to this hypothetico-deductive decision-making (HDDM) model, decision making consists of interaction among several well-defined cognitive processes. These processes can be labelled data acquisition, hypothesis generation, data interpretation, and hypotheses evaluation (5). In the initial phase of the patient encounter, certain clues trigger the formation of hypotheses. The first hypotheses are formed after the collection of a relatively small amount of data. Some hypotheses may trigger the consideration of other competing or related hypotheses. Once initial hypotheses are formed, knowledge of what is expected if the hypotheses were true is used to direct further data collection. Cue interpretation, which links the hypotheses to the known clinical information, and hypothesis evaluation, in which the aggregate weight of cues for competing hypotheses are compared, lead to the selection of the hypothesis or set of hypotheses which best explains the patient's condition.

A technique for evaluating the decision-making performance of physicians in terms of the HDDM model has been developed by Burke and Connelly at the University of Minnesota (7). The basis of this technique is a diagnostic problem simulation patterned after the patient management problem (PMP) of Rimoldi (8) and further developed by McGuire and her associates (9). By giving a simulation of this type to both experienced and novice physicians and comparing the results, it may be possible to identify specific areas within the framework of the hypothetico-deductive model in which novices are deficient relative to experts. Decision support might be appropriately directed toward such areas.

METHODOLOGY

Two clinical problem-solving tasks were developed in one medical domain (liver disease). The tasks were matched for difficulty. From an analysis of the relationship between PMP structure and the HDDM model measures useful in the evaluation of subject performance were defined. The PMP design includes several scoring indices. Two of these, hypothesis appropriateness (HA) and cue interpretation appropriateness (CIA) are useful for...
the evaluation of hypothesis generation. The HA index assesses the validity of a hypothesis given all of the information in the initial clinical scenario, while the CIA index is a measure of the validity of associating a particular hypothesis with a single clinical cue or test result. Using these two indices it was possible to determine the proportion of hypotheses which were appropriately generated. The number of hypotheses generated was also used as an indicator of hypothesis generation performance. The CIA index was also used as a measure of the cue interpretation phase.

Cue acquisition was directly assessed by the test selection appropriateness (TSA) index. This index assesses the validity of selecting a particular hypothesis. Useful derivations include the proportion of test selections which are appropriate and number of appropriate test selections which were missed.

Hypothesis evaluation is very difficult to assess in a paper-and-pencil PMP format. The phase of hypothesis evaluation uses multivariate information and it is difficult to know exactly what information is being considered in the evaluation of each hypothesis. Measures which attempt to indirectly assess this phase were devised. These measures used the ranked order of a group of hypotheses together with HA indices to assess the appropriateness of the hypothesis ordering.

Two PMP’s, one related to chronic active hepatitis and the other to alcoholic hepatitis, were sent to thirteen experienced, practicing gastroenterologists who formed the criteria or expert group. From these we obtained eight usable data sets for the first problem and eleven for the second. Novice data was taken from the same PMP’s used by Burke and Connolly (22) as pretests for the evaluation of their course. These yield 41 and 39 first and second problem data sets respectively. All of the novices were in their third or fourth year of medical school.

One rater (MDB) designed the PMP’s and also generated the scoring matrices for the HA, CIA, and TSA indices. Each problem was designed to be solvable using only the data given in the initial clinical scenario along with the available laboratory test results. The PMP’s were first coded by hand with numeric codes being assigned to all hypotheses and cues. These codes were then keypunched and the data analyzed using a Pascal program to generate the various scoring indices. The output of this program was used as input to various SPSS analysis programs. T-tests were used to test for differences in proportions and counts.

**RESULTS**

The most striking result was some fairly strong evidence that the experts did better than novices in the hypothesis generation phase. Not only did the experts generate significantly more new hypotheses after getting back the first set of test results in Problem 1 (1.6 versus 0.4; p = 0.056), but a greater proportion of their hypotheses were appropriate (1.00 versus .84; p = .047). In Problem 2 the experts generated more hypotheses in the initial stage after reading the clinical scenario but before tests results were obtained (4.8 versus 3.9) but this was not statistically significant. When both problems are taken together the experts generated an average of 6.2 hypotheses (n = 19) while the novices generated 5.0 (n = 80; p = .016). In addition, a higher proportion of the expert’s hypotheses were appropriate (.93 versus .84; p = .019).

Measures associated with the test selection and cue interpretation phases were little different between the two groups. None of the differences were significant for Problem 1. For Problem 2, the only significant differences were for the overall proportion of possible test selections which were missed (0.48 versus 0.57; p = .079) and the proportion of possible cue interpretations which were missed at Stage 2, for the first set of test result interpretations (0.46 versus 0.26; p = 0.037).

In terms of overall success in solving these problems, the physicians clearly did better than the students. Of the physicians, 63 percent had the correct hypothesis at the end of Problem 1, while none of the students did. On the second problem, 18 percent of the physicians were correct versus 3 percent of the students. Though none of the students got the correct answer of chronic active hepatitis on Problem 1, 95 percent of them were relatively close with some variant of viral hepatitis.

**DISCUSSION**

Patient Management Problems. The use of a paper-and-pencil patient management problem (PMP) simulation presents difficulties but at the same time offers several advantages over other approaches. What is lost in low fidelity and cueing is gained in greater experimental control, allowing observation of many decision makers’ behaviour on the same problem.

The difficulty in coding the rating criteria is a problem which is not easily solvable. The sheer mass of rating scores which must be generated, more than three thousand for each problem, made it necessary to rely on the judgment of one rater (MDB) in the creation of the rating index matrices for these problems. While this provides some consistency in the rating index matrices, it also introduces the individual biases of the rater. Future studies could consider combining the rating scales of two or more raters. Another approach might be to use a different rater on each problem. This would require a broader interpretation of the results over all problems, but in cases where potential raters cannot take the time to rate several problems it may be the only practical solution.

268
In spite of these difficulties there are several reasons why these PMPs were used. The PMPs had already been designated and used at the inception of this project, a major factor in making this study feasible. An important advantage of this moderate fidelity simulation, as opposed to a high fidelity one, is the ability to examine decision-making behaviour across many individuals. This is a necessity if the results of such studies are to be used in the design of decision support which is intended to provide the greatest aid to a large group of decision makers. The higher fidelity approaches would be very expensive to apply to many decision makers working on many different problems and would also not provide the degree of experimental control which the moderate fidelity approach, such as the one used here, can provide.

Test Administration Problems. Several problems occurred in the course of administering the tests to the subjects. A few subjects did not provide linkages from hypotheses to cues or hypotheses to test orders. The most common example of this was observed where students ordered some laboratory tests which they did not link back to specific hypotheses. In some cases subjects stopped in the middle of a test because a test or procedure which they wanted to use was not available.

The Hypothetico-Deductive Decision Making Model. A fundamental assumption underlying all of this work is that the HDDM model is an accurate one, or at least one able to serve as a valid predictor of problem solving. While much evidence has been provided by others in support of this model (5), it is unlikely that there is one common approach to medical inquiry. For the purposes of this study, however, this model offers two important and related advantages: 1) the model is simple, and 2) it provides a useful structure for the organization of decision support. Clearly, since this study was organized around this model, there are some biases in favor of it. For this reason it cannot be argued that the success of a decision support technique based on this model would help to verify the model.

Expertise. A different question entirely is how well medical students represent the expertise of that group of physicians to whom decision support might be directed: practicing clinicians in general practice or internal medicine with no more than the usual training in gastroenterology. A number of studies have demonstrated that the decision-making process appears to be about the same in both experts and novices (5, 10). The primary differences in performance are thought to be related to the extent of the knowledge base which the problem solver brings to the problem. Thus, if we accept that diagnostic performance improves along a continuous gradient in the course of medical education, then we may assume that the diagnostic errors made by practicing physicians are due to a lack of knowledge in the same areas in which medical students are lacking, but not to as great a degree. In this case, the errors made by medical students will be of the same type as those made by practicing physicians, but to a greater degree. The implication of this assumption for this study is that it will be easier to detect such errors using medical students than practicing physicians. Future studies are needed to verify this assumption. Assuming this assumption is valid, such studies will probably require larger numbers of both "expert" and "novice" practicing physicians in order to identify the smaller differences which will exist.

Hypothesis Evaluation Assessment. Although expert-novice differences regarding hypothesis evaluation measures were small, the difference indicated better performance by the novice group. This result tends to undermine the validity of the indirect approach taken to assess hypothesis evaluation. This was not totally unexpected. Such an assessment really needs to consider what information the problem solver is using in the evaluation. Hypothesis evaluation is clearly a multi-variate operation which cannot be directly assessed by the simple simulation technique used in this experiment. The best that could be done here was to assess outcome of the evaluation process in the form of hypothesis ranking. However, this ranking not only depends on what information goes into the hypothesis evaluation process, but also on the validity of that information which varies considerably among phases in the decision-making process. Another problem with the rank score is its rather arbitrary definition. Unfortunately a better one could not be found to compare rank orders with differing numbers of elements.

Expert-Novice Differences. The fact that the physicians did better than the students on both problems tends to confirm the validity of this approach. This expert-novice difference, however, varied considerably between the two problems. In Problem 2 only a small proportion of either the expert physicians or the novice students ended with "correct" answer of alcoholic hepatitis as the first choice. The greater difference in success on Problem 1 between these two groups should provide a more fertile area for identifying functional differences between them. Indeed, the strongest evidence for a functional difference on either problem appears for the hypothesis generation phase of Problem 1. In this problem the experts generated the same number of hypotheses as the novices after reading the clinical scenario but generated four times as many following return of the first set of laboratory test results. In addition, the overall proportion of hypotheses generated which were appropriate was higher for the experts than for novices. Also, the experts' hypotheses were closer to the goal of a chronic active hepatitis diagnosis. From these observations we can conclude that the experts are more successful at generating appropriate hypotheses than students. For Problem 1, then, it appears that the students' diagnostic expertise might be improved by providing some sort of hypothesis generation aid.
The physicians performed rather poorly on Problem 2, though they did do better than the students. Overall the experts missed fewer of the important reasons (hypotheses) for selecting a test as well as fewer of the test-to-hypothesis links. This suggests that the experts have a richer, more detailed knowledge of the linkages between laboratory tests and hypotheses and that this richer knowledge base may account for their slightly better performance. Problem 2, the alcoholic hepatitis problem, was expected to be somewhat more difficult than Problem 1, the chronic active hepatitis problem, but the poor performance of the physicians was still surprising. Successful completion of Problem 2 required the problem solver to look at the ratio of the serum aspartate aminotransferase (AST or SGOT) level to that of serum alanine aminotransferase (ALT or SGPT). Such a strategy may be too complex for problem solvers who appear to simply weigh various pieces of evidence against each other as the HDDM model of the problem solving process suggests.

Thus for the chronic active hepatitis problem, it appears that just thinking of the appropriate hypothesis was critical to success, while appropriate test selection and interpretation was not. This can be attributed to the relatively direct laboratory test-to-hypothesis linkages, as well as the existence of the nearly pathognomonic Hepatitis B surface antigen test. For the Alcoholic Hepatitis problem, though, these linkages were far less straightforward, and successful problem resolution required a stronger knowledge of the hypothesis-test linkages.

CONCLUSIONS
For patient management problems problem solving performance appears to be markedly problem-dependent. Any investigator of clinical problem solving who expects to reach broadly applicable conclusions will find it necessary to study a broad range of problem scenarios. The identification of distinct expert-novice differences suggest that this technique could be used to identify decision-making weaknesses toward which decision support could be directed. Confirmation of this theory will require a controlled experiment in which various forms of decision support targeted at each of the specific HDDM model phases and their effects are assessed.

REFERENCES