Diagnostic Reasoning of High- and Low-Domain Knowledge Clinicians: A Re-analysis

Arthur S. Elstein, Benjamin Kleinmuntz, Mitchell Rabinowitz, Robert McAuley, James Murakami, Paul S. Heckerling, James M. Dod,
Department of Medical Education (M/C 591)
University of Illinois at Chicago
PO Box 6998, Chicago, IL 60680

Abstract

Thinking aloud protocols previously obtained by Joseph and Patel were re-analyzed to determine the extent to which their conclusions could be replicated by independently developed coding schemes. The data set consisted of protocols from 4 cardiologists (low domain knowledge = LDK) and 4 endocrinologists (high domain knowledge = HDK), individually working on a diagnostic problem in endocrinology. Both analyses found that HDK physicians related data to potential diagnoses more than the LDK group, and that there were trends for HDK physicians to be more focused on the correct diagnostic components and to employ more single-cue inference and less multiple-cue inference. However, the re-analysis found no meaningful differences between groups in diagnostic accuracy, speed of diagnosis, or in the breadth of the search space used to seek a solution. The generalizability of results of protocol-analysis studies can be assessed by using several complementary coding schemes.

Analyzing the verbalization ("thinking aloud") of physicians as they respond to a sequentially presented clinical problem has been a popular strategy for studying the study of medical cognitive processes. These studies typically present a diagnostic or treatment planning problem in a series of discrete steps, thereby assuring that each subject is exposed to the same information in the same order, and ask subjects to think aloud as the clinical situations unfolds.

Comparing, contrasting and then generalizing across studies has been difficult. Investigators have used different problems, different approaches to or frameworks for protocol analysis, and different groups of subjects (medical students, residents, academic physicians, practitioners in various practice settings). Since there is ample evidence that medical problem solving is partly contingent on the content of the problem, the phenomenon variously called domain- or content- or case-specificity (Elstein et al., 1978, 1990; Grupen, Wolf, Van Voorhees, 1988; Norman, 1988; Swanson, 1990), generalizing about cognitive processes from a limited sample of cases is clearly a problem. Substantial within-group variability has also been observed in many studies (e.g., Grant & Marsden, 1987; Johnson, et al., 1982). Yet protocol analysis is so labor-intensive that these studies invariably employ small samples of both cases and subjects.

In addition to variation across studies due to cases and subjects, investigators have employed protocol analysis schemes that were not subsequently used by other research teams. These analytic methods usually contain many qualitative or subjective elements, and it is often unclear that the conclusions asserted would have been reached by a different team of investigators, even had they employed the same approach to protocol analysis. Consequently, it has been difficult to compare and contrast conclusions of different studies, and to determine to what extent the conclusions of any investigation are independent of the method of protocol analysis, or what aspects of the data are highlighted by each method of analysis.

One approach to these difficulties is to re-analyze protocols obtained by one set of investigators using another framework. This paper reports one such analysis, presenting a re-analysis of data originally collected and analyzed by Joseph and Patel (1990). Its aim is to determine the reproducibility of their conclusions by re-analyzing the raw data with independently constructed schemes.

Method

Subjects

Joseph and Patel's study employed a sample of 9 senior physicians, all associated with the Faculty of Medicine at McGill University, 4 endocrinologists and 5 cardiologists. The clinical problem concerned the

This research was supported in part by a grant from the National Library of Medicine, RO1-LM-4583. We are grateful to Vimla Patel for providing materials from her files for this re-analysis.

diagnosis of an endocrine disorder, so the endocrinologists were considered as High Domain Knowledge (HDK) subjects and the cardiologists as Low Domain Knowledge (LDK) subjects. All of the subjects were board-certified, practicing physicians with five to ten years of experience in their fields. Our re-analysis was performed on protocols of 8 subjects, all that were made available.

Clinical Problem

The complete text is presented in Joseph & Patel (1990). The vignette was organized into three parts, corresponding to the traditional partition of a case presentation into history, findings from the physical examination, and x-ray and laboratory test results. These parts were further subdivided into 25 segments presented sequentially. The first part consisted of 4 segments of medical history, the next 14 presented physical exam information, and the final 7 segments included x-ray and laboratory test information. According to Joseph and Patel, the accurate diagnosis of this case is Hashimoto's hypothyroidism with myxedema pre-coma. The diagnosis can be further divided into three components of increasing specificity: hypothyroidism, myxedema pre-coma, and an autoimmune condition called Hashimoto's thyroiditis. The cognitive task is diagnosis, not treatment or management. The time elapsed during the events in this case is not specified in the vignette, but is probably not more than 3 hours.

Original Procedure

All subjects were individually tested. The stimulus material was presented on a computer monitor one segment at a time. Subjects were asked to think aloud after each segment, verbalizing about the role and importance of the information in each segment in reaching the correct diagnosis. They controlled the rate of presentation of each segment and there was no time limit for their responses. Transcripts of tape recorded verbalizations provided the raw data.

Protocol analysis procedures

In prior studies, we had developed two coding schemes as part of an investigation of expert-novice differences in clinical reasoning about cases in a medical intensive care unit. One focused upon problem solving processes (Kleinmuntz & Elstein, 1990), the other upon categories of knowledge used in problem solving (Rabinowitz, Cottrell & Elstein, 1990).

The analysis of problem solving processes utilized 5 constructs: 1) formulates new diagnostic hypothesis; 2) confirms or strengthens diagnostic hypothesis using new facts or findings; 3) rejects or weakens diagnostic hypothesis using new facts or findings; 4) bases inference on multiple cues; and 5) bases inference on single cue. Each process was coded every time it occurred in the 25 segments, except that diagnostic formulations (construct #1) were noted only the first time they were mentioned, as only new formulations were coded. To facilitate coding, specific inclusionary and exclusionary criteria were developed and representative examples were used to define further the characteristics of each construct. Scores were summed for each subject within and across the three sections. For the present study, the coding scheme was revised by adding a distinction between diagnoses (more specific and precise formulations) and problems (more vague and global). The diagnostic hypotheses were reviewed by one of the authors (PSH) and classified into diagnoses and problems. The coding was done by a single rater (JM).

The knowledge utilization analysis focused on the categories of knowledge used in diagnostic reasoning and on the question of whether differences in category utilization can distinguish groups of clinicians. The protocols were divided into units consisting of either a single sentence or a sentential phrase with subject and object. No unit was longer than a sentence; some were even briefer. These units were identified and then coded by a single rater (RM) utilizing a scheme designed to identify the categories of knowledge and information in each unit. Four primary categories were employed: context (personal and social history, risk factors, circumstances of acute onset), findings (relevant medical history and physical findings), tests, and diagnosis. Additional "relational categories were used to identify units in which two primary categories were related to each other.

Analysis of Data

While quantitative analytic methods were employed, statistical tests of significance are not very useful. We have chosen to describe the quantitative findings and to suggest trends in the data.

Our analysis focussed upon 4 central questions. For each, we present initially the findings of Joseph and Patel, and then our own conclusions, noting agreements and divergences, as appropriate. The questions are:

1. Diagnostic accuracy: Did HDK and LDK physicians differ in diagnostic accuracy?

2. Timing and course of diagnostic hypotheses: When in the sequence of data presentation were the correct diagnostic components generated? Were HDK physicians correct earlier?

3. Diagnostic process: What are the differences between groups in diagnostic process, particularly in generating and testing alternative hypotheses? Do they differ in the number of alternative hypotheses considered? Do HDKs and LDKs differ in the breadth of the search space used in seeking a solution?

4. Knowledge utilization: Are there differences between groups in the categories of knowledge used in problem solving, particularly in relating data to diagnostic hypotheses?

Results

Diagnostic Accuracy

Joseph and Patel did not provide any data regarding the number of clinicians in each group who generated each component of the correct diagnosis, but they reported that "those LDK subjects who produced accurate diagnostic components produced them later than did the HDK subjects." Our analysis shows that the first diagnostic component, hypothyroidism, and the second diagnostic component, myxedema, were generated by all clinicians. The third diagnostic component, Hashimoto's thyroiditis was generated by two HDK and two LDK clinicians. Diagnostic accuracy is not a function of the number of hypotheses proposed. We conclude that, with respect to diagnostic accuracy, there is no meaningful difference between groups.

Timing and Course of Hypotheses

The common wisdom about diagnostic skill is that experts reach conclusions more rapidly and efficiently than those less experienced in a particular domain. Joseph and Patel's findings accord with this general view. They reported that HDK clinicians generated each diagnostic component earlier than LDK clinicians, and illustrated this conclusion by providing data from one subject from each group rather than by providing aggregate data from all.

Our analysis identified the point in the case at which each subject generated each diagnostic component. The conclusions of Joseph and Patel are affirmed in broad outline with respect to the first and third diagnostic component, hypothyroidism and Hashimoto's thyroiditis, although there are minor differences in our analyses. We found that HDKs generated these components somewhat earlier than LDKs (although not so much earlier as to be clinically significant). However, our interpretation of the data regarding the second diagnostic component, myxedema, is quite different. Two HDK clinicians developed this component very early (by segment 4), while two did not mention it (even with very liberal criteria) until segment 8 and 17, respectively. The LDK clinicians all generated this aspect of the diagnosis in segment 5 or 6. Thus, HDK physicians were more variable in the timing of this diagnostic component and are not necessarily earlier, although the lag is clinically unimportant. Both groups tend to move from general diagnoses to more specific diagnoses.

Diagnostic Process

The cumulative sum of the mean number of new diagnoses produced was the dependent variable Joseph and Patel used to examine diagnostic process. They reported that after HDK physicians had produced the accurate diagnostic components, they generated few new diagnostic hypotheses and spent the rest of their time confirming hypotheses. They sporadically ruled out some of the hypotheses generated earlier, and most often used the new findings from the physical examination to confirm the diagnosis and determine secondary problems. In contrast, LDK subjects continued to generate new diagnostic hypotheses throughout the case. They focused primarily on associating clinical findings and test results with new possibilities. They generated very few secondary problems, and did not rule out the diagnostic hypotheses that they had generated earlier, even when the new hypotheses were contradictory to some of the earlier ones. In general, the HDK subjects narrowed uncertainty while LDK subjects increased it. Their results imply that the HDK subjects "coned down" on a diagnosis, working in a progressively narrower problem space.

Our reanalysis hypothesis and problem generation found that the number of new problems identified levels off for both groups after segment 15. The mean number of new diagnostic hypotheses continued to climb slowly for both groups. Both HDK and LDK clinicians considered similar numbers of new diagnostic hypotheses across all segments. In contrast to the view of Joseph and Patel, this analysis argues that the two groups do not differ quantitatively in the breadth of the search space used to seek a solution.

The diagnostic reasoning process may be further examined by referring to measures of hypothesis confirmation, disconfirmation, and types of inference. The mean number of attempts to confirm and to rule out diagnoses was the same for both groups. We tabulated the numbers and percentages of efforts at hypothesis confirmation that involved the correct diagnostic components vs. all other formulations. HDKs produced slightly more instances of confirma-

tion of the correct components.

Single-Cue and Multiple-Cue Inferences.

A single-cue inference relates a single item of information to a diagnostic hypothesis, in an effort either to confirm or rule out; a multiple-cue inference relates two or more cues to a diagnostic hypothesis in a semantic unit such as a sentence. The data show that LDK physicians produced slightly more multiple-cue inferences than did HDKs, and a lower percentage of this output was concerned with the correct components. This pattern is reversed for single-cue inferences: HDKs produced more such processes and a higher percentage was focused on the correct diagnostic components. Thus, there is a consistent trend for HDK clinicians but this trend is not statistically significant. The mean ratio of single-cue to multiple-cue inferences was 2.65:1 for HDK physicians and 1.86:1 for LDKs. These ratios show a trend for HDK clinicians to rely more on single-cue inference and less on multiple-cue inference. These results might suggest that as the case moves along, HDK subjects rely more on direct associations from data to cognitive schemata and less on step-by-step aggregation of findings to build a clinical picture than do the LDKS.

Knowledge Utilization

Joseph and Patel found that HDK subjects used more relations to connect important information and ignored irrelevant information. The evidence presented in support of this claim was largely anecdotal.

The quantitative analysis of knowledge utilization focused on the ratios between the primary categories of "Findings" (History in part 1 of the case, physical findings in part 2) and "tests" (Laboratory test data and x-rays, in part 3) and the relational categories Findings-to-Diagnosis,Diagnosis-to-Findings,Test-to-Diagnosis, and Diagnosis-to-Test.

Both groups handle Part 1 in a fairly similar fashion, with statements about findings alone equalling the number of relational statements. In parts 2 and 3, however, the HDKs move toward a much more relational mode of discourse, consistently relating findings or test results to diagnoses much more than do the LDKs. The LDKs' statements typically stayed closer to the data, and related them less often to the diagnoses under consideration. These measures provide a concise quantitative picture of information processing by HDK and LDK clinicians and in general support the results of Joseph and Patel. At the level of a sentence-by-sentence analysis of the discourse of these physicians, HDK clinicians do relate the data to diagnoses more than do LDKs. Our analysis supports Joseph & Patel's conclusions that HDK physicians on the average used more relations to connect important information, but we also found substantial within-group variation.

Joseph & Patel reported that HDKs concentrated more on the correct diagnostic components while LDKs were more concerned with incorrect hypotheses. We analyzed the content of the relational diagnostic statements made by the members of each group to determine if they referred to one of the three correct diagnostic components of the case. 82% of the HDKs' diagnostic relational statements and 76% of the LDKs' concerned one or more of the three diagnostic components. Our interpretation is that there is probably no meaningful difference between groups.

Discussion

In this paper, data previously reported by another group of investigators have been independently reanalyzed by a set of methods for protocol analysis that are quite different than the techniques they employed. Joseph and Patel had studied two small groups of experienced clinicians on one endocrinology case. One group was quite expert in the domain (High Domain Knowledge, HDK) and the other was not (Low Domain Knowledge). They identified several differences in the reasoning of the two groups. According to their account, HDK clinicians produced the components of an accurate diagnosis earlier than the LDK group and were also more efficient and focused. HDK clinicians generated fewer new diagnostic alternatives, spent more time confirming prior hunches, used more relations to connect important information, and ignored irrelevant information.

We found some trends in quantitative measures of diagnostic reasoning that are partial support for their overall conclusions. Both analyses agree that HDK physicians related data to potential diagnoses more than the LDK group, and that there was a trend for HDK physicians to be more focused on the correct diagnostic components. It can be argued that HDKs know the links between findings and diagnoses, while LDKs recognize clinical abnormalities both in the physical examination and laboratory data, but do not have the linkages between the data and diagnoses needed for understanding and solving difficult diagnostic problems.

On the other hand, some of their conclusions are not supported by our reanalysis. Specifically, we found no important or clinically meaningful differences between the groups in overall accuracy of diagno-

sis, speed of accurate diagnosis (as measured by the segment number at which each component is first mentioned), or the quantitative breadth of the search space (a measure of diagnostic focus). We found that both groups considered comparable numbers of new diagnostic hypotheses and more general problem formulations, continued to generate a small but steadily increasing number of diagnostic alternatives as the problem unfolded, and were very similar in mean numbers of efforts at hypothesis confirmation or disconfirmation.

References

Barrows, H. S., & Feltovich P. 1987. The clinical reasoning process. *Medical Education*, 21, 86-91.

Elstein, A. S. 1988. Cognitive processes in clinical inference and decision making. In D. C. Turk and P. Salovey (Eds.), *Reasoning, inference and judgment in clinical psychology* (pp. 17-50). New York: Free Press/Macmillan.

Elstein, A. S., Shulman, L. S., & Sprafka, S.A. (1978). *Medical problem solving: An analysis of clinical reasoning*. Cambridge, MA: Harvard University Press.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1990). Medical problem solving: A ten-year retrospective. *Evaluation and the Health Professions*, 13, 3-36.

Grant, J., & Marsden, P. (1987). The structure of memorized knowledge in students and clinicians: An explanation for diagnostic expertise. *Medical Education*, 21, 92-98.

Gruppen LD, Wolf FM, Van Voorhees C, Stross JK. The influence of general and case-related experience on primary care treatment decision making. Arch Intern Med 1988;148:2657-2663.

Johnson, P.E., Duran, A.S., Hassebrock, F., Mollar, J., Prietula, M., Feltovich, P., & Swanson, D.B. (1981). Expertise and error in diagnostic reasoning. *Cognitive Science*, 5, 235-285.

Joseph, G. M., & Patel, V. L. (1990). Domain knowledge and hypothesis generation in diagnostic reasoning. *Medical Decision Making*, 10, 31-46.

Kassirer, J. P., (1989). Diagnostic reasoning. *Annals of Internal Medicine*, 110, 893-900.

Kassirer, J. P., & Kopelman, R. I. (1989). Cognitive errors in diagnosis: Instantiation, classification and consequences. *American Journal of Medicine*, 86, 433-441.

Kassirer, J. P., Kuipers, B. J., & Gorry, G. A. (1982). Toward a theory of clinical expertise. *American Journal of Medicine*, 73, 251-259.

Kleinmuntz, B., & Elstein, A.S. (1990). Toward medical decision making expertise. Paper presented at the Annual Meeting of The Judgement and Decision Making Society, New Orleans, November 18, 1990.

Norman, G. R. (1988). Problem-solving skills, solving problems and problem-based learning. *Medical Education*, 22, 279-286.

Swanson, D. B. (1990). Issues in assessment of practical skills in medicine. *Professions Education Researcher Quarterly*, 12, 3-6.