Premature Conclusions in Diagnostic Reasoning

Anthony E. Voytovich, M.D., Robert M. Rippey, Ph.D., and Anthony Suffredini, M.D.

Abstract—The purpose of the study reported here was to explore the characteristics of premature diagnostic conclusions in a group of physicians, medical students, and residents. When the subjects were asked to construct complete, precise problem lists from three case abstracts, premature closure occurred frequently, it could be recognized with good interrater reliability, and it seemed to appear with equal frequency regardless of the level of training.

The initial assembly of multiple clinical clues from a patient's history, physical examination, laboratory tests, and X-rays into working diagnoses or conclusions is a critical phase in the medical care process. Specialized diagnostic procedures and therapies are frequently begun from this starting point. The process may be compared with a navigator's initial fix, for regardless of how precise and sophisticated subsequent maneuvers are, a start from the wrong point will be costly.

Traditional medical teaching exercises, including patient management problems and clinical pathological conferences, focus on arriving at a single diagnostic entity. However, most medical patients (particularly the elderly) have multiple problems, some with overlapping features. Some conditions are clearly defined, and some are ambiguous. Often patients have multiple diagnostic entities converging in a single manifestation (such as dizziness or fatigue). The authors have suggested that a complete, precise problem list might serve as an inventory on which a physician might show each condition in relation to others. Choice of terminology allows ambiguity to be expressed and strategies in the clustering of data to be represented.

The first author described common error patterns in the construction of such precise lists at a meeting of the Association of American Medical Colleges Council of Deans in 1975. This work led to the conclusion that error patterns in diagnostic reasoning could be classified into four categories:

Omission—An important clinical clue is simply ignored. For example, anemia or proteinuria is ignored and does not appear in the formulation.

Premature closure—The diagnosis of the patient's condition is less than justified by existing data. For example, anemia and rectal bleeding are presumed to confirm the presence of iron deficiency.

Wrong synthesis—Available data con-

Dr. Voytovich is associate professor, Department of Medicine, and Dr. Rippey is professor, Department of Research in Health Education, University of Connecticut School of Medicine, Farmington. Dr. Suffredini is a medical staff fellow in the Critical Care Medicine Department, National Institutes of Health Clinical Center, Bethesda, Maryland.

tradict the conclusions. For example, trace ketouria, pH of 7.42, and bicarbonate of 30 are called ketoacidosis.

Inadequate synthesis—Conclusions that could be supported by data are not drawn. For example, history of angina with 80 percent occlusion of the left main coronary is called organic heart disease.

Since totally wrong synthesis is rare, investigation has centered on the other three categories of error. Omission and inadequate synthesis have been studied extensively in both real and simulated teaching situations (1, 2). These errors can be detected with striking interrater reliability. Those making the diagnoses display similar tendencies on similar case material. The number and the type of errors discriminate for level of experience; the independent occurrence of error types on individual problem lists suggests that these behaviors reflect separate diagnostic reasoning skills. Rater agreement has been shown to be strong on simulated cases. Also, the first author presented data at the Conference on Research in Medical Education sponsored by the Association of American Medical Colleges in 1979 suggesting that similar rater agreement can be attained when auditing actual problem lists on hospital wards. In addition, the second author presented work at a meeting of the American Educational Research Association in 1983 describing a computer program capable of recognizing the four errors in problem lists typed in free-form at a keyboard.

The evolution of diagnostic skill as revealed by studying the pattern of errors in reasoning is not always as expected. More experienced clinicians may recognize complex patterns and perform better than the less experienced clinicians with regard to inadequate synthesis; however, they tend to focus on a central concern and ignore seemingly peripheral (but frequently significant) issues. Inexperienced students, on the other hand, often note each abnormality and produce a more thorough but less well integrated list (1, 2).

Methods

Three actual cases that presented likely opportunities for premature closure were abstracted for the study reported here.

Case 1—A 46-year-old woman with polydipsia, polyuria, and weight loss who was told in the past that her blood pressure was high. The abstract contained insufficient data to substantiate diabetes or hypertension.

Case 2—A 65-year-old male with Laennec's cirrhosis who presents with rectal bleeding. He has a history of varices and a coagulopathy, but there is insufficient data to pinpoint the cause of the bleeding. He also has a low grade anemia and data to suggest emphysema, again both inviting etiologic conclusions without substantive data.

Case 3—A 76-year-old woman with acute alteration in mental status and congestive heart failure with atrial fibrillation. The mental status and failure cannot be explained etiologically with the data given, nor is there evidence of causal relationship between them.

A list of major anticipated premature conclusions was constructed and agreed upon by the authors in advance. This list served to provide examples of premature closure. It was not meant to be all inclusive, as it would be impossible to anticipate every response. Some judgment by the raters was required.

Sixty volunteers at the University of Connecticut Health Center were asked in 1980–81 to construct a complete precise problem list for each of the three cases and were cautioned to avoid conclusions not substantiated by data given. (A diagnosis was said to be substantiated if treatment could be begun without further di-

agnostic maneuvers to establish the conclusion as stated at the outset.) Fifty-eight useful sets of responses covering a total of 164 cases were returned from 11 second-year medical students, 15 third-year students, 10 fourth-year students, seven first postgraduate year (PGY-1) residents, eight PGY-2 and PGY-3 residents, and seven faculty members in the School of Medicine's Department of Medicine.

The responses were coded to blind the two raters to the subjects' identities and levels of training. Instances of premature closure were noted and counted. Totaling the number of omissions and inadequate syntheses was done in order to replicate prior work, to provide some estimate of the general ability of these cases to discriminate for level of training using known measures, and to provide a reference against which to compare performance on premature closure.

In order to determine interrater reliability, an administrative secretary at the medical school scored 20 randomly selected sets of three cases (60 in all), noting what would appear to be premature conclusions to her. Rater agreement was determined by simple correlation of the number of instances of premature closure noted by the third author and by the secretary on each case.

In order to determine parallel form validity (consistency of individual performance across cases), all three possible pairs of the three cases were correlated on an individual-by-individual basis.

Multivariate analysis of variance was employed to determine the relationship between performance and level of training.

Results

Fifty-three (out of the 58) subjects made errors of premature closure. Although the cases were of a nature that would lead to unwarranted conclusions, errors of omission and inadequate synthesis were more frequent than premature closure.

Twenty-one of the 58 subjects made premature closures on all three cases, 15 on two cases, and 17 on one case. Only five subjects made none. Table 1 depicts the mean numbers of errors committed in the three cases by the level of training of the respondents.

The two raters agreed closely on the number of instances of premature closure on each problem list despite the fact that the lists represented the respondents' own choice of words and that one rater had no medical training or background (though she understood the purpose and structure of a problem list). The correlations were: case 1 = .80; case 2 = .68; case 3 = .78 ($p < .0005$).

Correlations for the number of instances of premature closure on the three possible case pairs were: cases 1 and 2 = .29; cases 1 and 3 = .29; cases 2 and 3 = .49. Application of the Spearman-Brown prophecy formula to the pooled correlations (using Fisher's Z transformation) suggests a reliability of .62 on a hypothetical 10-case test.

The reliability of .62 for premature closure was modest compared with the .85 and .78 previously described for inadequate synthesis and omission (2) yet certainly worthy of note given the general difficulty in demonstrating consistent subject performance across case material in most studies.

Improvement in diagnosis with more training can be seen in the data in Table 1 in omissions (second-year students with a mean number of 23.9 and faculty members with a mean number of 13.4) and in inadequate synthesis (second-year students, 3.5; second- and third-year residents, 0.5; faculty members, 0). Multivariate analysis of variance indicated significant differences with respect to training level of $<.001$ and $<.003$ for these two

TABLE 1

Mean Number of Errors in Diagnoses on Three Case Abstracts by 58 Medical Students, Residents, and Faculty Members, University of Connecticut Health Center, 1980–81

Respondent	Omissions		Inadequate Synthesis		Premature Closure	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Students						
Second-year	23.9	9.5	3.5	4.3	2.6	2.4
Third-year	16.5	8.0	0.7	1.1	3.5	2.6
Fourth-year	17.8	7.6	1.1	2.0	2.6	1.3
Residents						
First-year	15.1	5.1	0.3	0.5	4.2	3.0
Second- and third-year	12.1	5.4	0.5	0.8	3.9	3.1
Faculty members	13.4	3.5	0.0	0.0	1.9	2.4

errors. In contrast, premature closure appeared randomly among the respondents and had no tendency to be characteristic of the least or most experienced group.

Discussion

In the study described here, the frequency of premature conclusions by the respondents making diagnoses seemed, in contrast to the other errors, independent of training and level of ability.

It must be acknowledged that the likelihood of failing to detect differences between training levels is high. Even if the entire group of respondents were split into an upper and lower half (29 per group) and the means compared by simple independent *t*-test, power is only .46, with alpha set at .05 and .5 standard deviations regarded as a real difference (3). This suggests a 54 percent chance of missing an association between premature closure and training level. Although a larger sample would increase the likelihood of detecting such an association, the data in Table 1 suggest that the effect would be small for premature closure when compared with omission and inadequate synthesis.

Prior studies have suggested that premature closure might be independent of knowledge. In a study (4) in which 83 clinical students were given two problem lists to construct and were tested on related and unrelated subject matter using confidence testing, it was possible to obtain simultaneous data on erroneous synthesis, content, and the subjects' opinions regarding the certainty of correctness of their answers. It was shown that errors of inadequate synthesis correlated strongly with knowledge, while premature closure occurred equally in well informed and poorly prepared students. Premature closure, however, correlated significantly with confidence in case-relevant (but not irrelevant) subject matter. That is, students who had difficulty perceiving whether their multiple-choice test answers were correct were more likely to make premature closures, regardless of actual test score.

In many teaching settings, there seems to be an inordinate value placed on making a diagnosis with minimum information, even if guessing is involved. Although answers are important, premature closure is potentially more dangerous than no answer at all for the following reason: diagnostic terminology carries an implicit "flag" signaling that conclusions in the case are complete. Multiple elements from the patient history, the phys-

ical examination, and laboratory and X-ray results have been consolidated into a compact package, pathophysiologic abnormalities have been recognized, and finally a diagnosis with an etiology has been specified. The etiologic diagnosis frequently indicates that no further specification is needed. It is surprising how often an even tentative holding of such a conclusion stops further diagnostic consideration. Thus, premature closure in effect represents closure. Inadequate synthesis, in contrast, by nature of the nonspecific terminology, demands further thought. Although inadequate synthesis may cause a delay in treatment, premature closure, in addition to the delay, may lead to the wrong treatment and a false sense of confidence. It may be more difficult to disregard or disassemble an ongoing diagnosis than it is to arrive at a fresh one. Therefore, the consequences of premature closure may be more costly than other types of errors in diagnosis. Furthermore, its occurrence cannot be reliably predicted by the stature or experience of the person making the diagnosis, and it may result in inappropriate therapy.

Given a tendency toward case specificity and no appreciable correlation between premature closure and level of experience and training in the present study, whether premature closure is a viable variable at all can be questioned. The authors believe it is for several reasons. First, the cases in the study here were not parallel cases. A higher correlation may have been achieved if they involved similar conditions. Second, the criterion of training and experience may not have been appropriate, since premature closure had not been taught to the same extent as inadequate synthesis had. There was less emphasis on recognizing and avoiding it. Developers of medical training curricula in this country have not shown much interest in developing courses in risk and uncertainty, in spite of excellent models abroad (5, 6). Until such development takes place, it seems probable that a proper criterion group will not be available. The validity of measures of premature closure can be disputed because of this lack of a criterion group. Hopefully, proposals for the training of such a group will not be deferred on this basis. The authors feel that premature closure is important because it can be identified dependably in real and simulated cases and because it is related to an independent measure of confidence (4).

Another diagnostic behavior reported as independent of training and experience is "anchoring," first described by Tversky and Kahneman (7, 8). Anchoring is the tendency of persons to retain early hypotheses in spite of subsequent information. Anchoring differs from premature closure in the following respect. Once closure is made, either premature or otherwise, anchoring leads to retention of the closure in the face of new evidence. Premature closure, on the other hand, takes place before the "anchor" is set. Premature closure hinders the search for new evidence. Anchoring retards a response to new evidence once it is available.

The research reported here may have important implications for clinical teachers. Medical educators should encourage a skeptical attitude in their students and teach them to recognize fully justified conclusions. They should discourage hasty decisions, even if they are correct. Even the best trained specialist may close prematurely on what might be considered a random basis. Thus, physicians should encourage independent review of their conclusions and realize that knowledge provides no shield against premature closure.

References

1. VOYTOVICH, A. E., and RIPPEY, R. M. Audit of the Structured Problem List as a Measure of Clinical Judgment. *J. Irish Med. Assoc.*, **71**:346–349, 1978.
2. VOYTOVICH, A. E., RIPPEY, R. M., and COPERTINO, L. Scoreable Problem Lists as Measures of Clinical Judgment. *Eval. Health Prof.*, **3**:159–170, 1980.
3. COHEN, J. *Statistical Power Analysis.* New York, New York: McGraw-Hill, 1977.
4. VOYTOVICH, A. E., and RIPPEY, R. M. Knowledge, Realism, and Diagnostic Reasoning in a Physical Diagnosis Course. *J. Med. Educ.*, **57**:461–467, 1982.
5. HOGARTH, R. Cognitive Process and the Assessment of Subjective Probability Distributions. *J. Am. Statis. Assoc.*, **70**:271–289, 1975.
6. WOOLRIDGE, C., DRAPER, N., and HUNTER, E. (Eds.). *Risk: A Second Level Course.* Milton Keynes, England: Open University Press, 1980.
7. FRIEDLANDER, M., and PHILLIPS, S. Preventing Anchoring Errors in Clinical Judgment. *J. Consult. Clin. Psychol.*, **52**:366–371, 1984.
8. TVERSKY, A., and KAHNEMAN, D. Judgment Under Uncertainty: Heuristics and Biases. *Science*, **185**:1124–1131, 1974.