Journal of Experimental Psychology:
Learning, Memory, and Cognition
1993, Vol. 19, No. 5, 1151–1164

0278-7393/93/$3.00

Determinants of Diagnostic Hypothesis Generation: Effects of Information, Base Rates, and Experience

Elke U. Weber, Ulf Böckenholt, Denis J. Hilton, and Brian Wallace

Physicians generated diagnostic hypotheses for case histories for which 2 types of diagnoses were plausible, with one having a higher population base rate but less severe clinical consequences than the other. The number of clinical and background symptoms pointing towards the 2 diagnoses was factorially manipulated. The order and frequency with which physicians generated hypotheses varied with the amount of relevant clinical and background information and as a function of population incidence rates, with little evidence of base rate neglect. Availability of a hypothesis, made possible by diagnosis of a similar case before, also made doctors generate this diagnosis earlier and more frequently. Physicians' experience affected hypothesis generation solely by increasing the availability of similar cases. The results are consistent with the use of similarity-based hypothesis generation processes that operate on memory for prior cases.

The process of hypothesis generation in problem solving has received inadequate attention. Increasing concern with problem-solving skills in areas as diverse as medicine or manufacturing has resulted in descriptive and prescriptive accounts of human problem solving that typically divide the solution process into the subprocesses of hypothesis generation, information acquisition, and hypothesis testing. However, psychological researchers from Bruner, Goodnow, and Austin (1956) through Wason and Johnson-Laird (1972), Newell and Simon (1972) to Klayman and Ha (1987), have almost exclusively studied the processes underlying the testing of established hypotheses, with little attention to the complementary processes involved in the generation of hypotheses from available evidence. This focus on hypothesis testing or information evaluation may result from the fact that these tasks are better structured (i.e., have well-specified inputs and goals) and thus are more easily studied. Yet, in many domains (in medicine as well as elsewhere), diagnosis has the characteristics of a loosely structured problem (i.e., it requires the problem solver to impose structure on the problem domain by eliciting information, ordering investigations, and generating hypotheses; Simon, 1973). The task of a family practitioner, for example, is closer to that of a detective than to that of a judge (i.e., he or she is actively involved in the generation of hypotheses and the acquisition of information to evaluate these hypotheses).

Although some investigators (Abelson & Lalljee, 1988; Gettys & Fisher, 1979) have addressed the dynamic aspects of adding or deleting hypotheses over time, few have investigated the processes by which particular working hypotheses are initially generated. In our study we attempted to address this imbalance by providing empirical information about three sets of questions regarding hypothesis generation. First, what is the role played by case information, diagnosis base rates, and problem solvers' expertise in the generation of initial diagnostic hypotheses? Second, what determines the size and diversity of the initial hypothesis set? Third, what stopping criterion terminates the initial generation of hypotheses? In the remainder of the introduction we first document the importance of a better understanding of hypothesis generation and then elaborate on each of these three questions.

Medicine has been an important area for studying hypothesis generation in problem solving (Elstein, Shulman, & Sprafka, 1978). A substantial body of evidence demonstrates that physicians generate general hypotheses or working interpretations of patients' presenting complaints early in the consultation. In an analysis of actual videotaped clinical interviews, Gale and Marsden (1982) found that physicians

Elke U. Weber, Center for Decision Research, Graduate School of Business, University of Chicago; Ulf Böckenholt, Department of Psychology, University of Illinois at Urbana-Champaign; Denis J. Hilton, Marketing Department, Ecole Supérieure des Sciences Economiques et Commerciales; Brian Wallace, University of Wales College of Medicine, Heathpark, Cardiff, United Kingdom.

This research was supported by Biomedical Research Support Grant RR 7030 from the National Institutes of Health and by a grant from the University of Illinois Research Board to Elke U. Weber, Ulf Böckenholt, and Denis J. Hilton. Elke U. Weber and Ulf Böckenholt contributed equally to the study. Denis Hilton was supported by fellowships from the University of Illinois Cognitive Science Committee and the Alexander von Humboldt Foundation. This article was completed while Elke U. Weber was a Fellow at the Center for Advanced Study in the Behavioral Sciences, Stanford, California, with financial support from National Science Foundation Grant SES-9022192 and the Graduate School of Business, University of Chicago. We thank Soyong Kwon and Veronique d'Estaintot for assistance with data coding; Sema Barlas for assistance with the data analysis; Paul Heckerling and Barbara Shapiro for sharing their medical expertise; and Sema Barlas, Lee Beach, Jerry Busemeyer, Bill Goldstein, Chip Heath, Joshua Klayman, Greg Murphy, Ilan Yaniv, and an anonymous reviewer for helpful comments on this article.

Correspondence concerning this article should be addressed to either Elke U. Weber, Center for Decision Research, Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago, Illinois 60637, or, Ulf Böckenholt, Department of Psychology, University of Illinois at Urbana-Champaign, 603 East Daniel Street, Champaign, Illinois 61820.

developed working hypotheses during the first 50 s of the consultation. Evidence of early hypothesis generation on the basis of limited information has also been obtained in neurological diagnosis (Barrows & Bennett, 1972), surgical diagnosis (Dudley, 1970, 1971), and internal medicine (Barrows, Norman, Neufeld, & Feightner, 1982; Elstein, Kagan, Shulman, Jason, & Loupe, 1972; Elstein et al., 1978). Barrows et al. (1982) demonstrated the critical importance that early diagnostic hypotheses may play in guiding further information gathering as well as in ultimately determining diagnostic success. In a high-fidelity simulation, 96% of the physicians whose initial hypothesis set included the correct diagnosis eventually settled on that diagnosis, whereas only 14% of those doctors whose initial hypothesis set did not include the correct diagnosis eventually arrived at it.

The process of hypothesis generation has, however, received little attention even in the medical literature. When discussed at all, it is frequently described by the perceptual metaphor of "pattern recognition" (Pauker, Gorry, Kassirer, & Schwartz, 1976). This metaphor reduces hypothesis generation to the well-defined task of categorization, a characterization that is also frequently found explicitly in the literature (Brooks, Norman, & Allen, 1991; Medin & Edelson, 1988). Although this characterization may apply in many situations, it is not sufficient to describe all aspects of hypothesis generation. Doctors are most apt to use the metaphor of pattern recognition when describing the process of hypothesis generation in nonproblematic, routine cases. Thus, the subjective feeling of a diagnosis "popping" into the physician's mind (Barrows & Bennett, 1972) might accompany the event of an uncontested activation of a single diagnostic category by the set of presenting symptoms. On the other hand, if more than a single diagnostic category is being activated, the case will probably be perceived as nonroutine, and the physician may engage in more complex analytical reasoning processes to generate one or more diagnostic hypotheses (Eddy & Clanton, 1982; Evans & Gadd, 1989).

A second related issue that complicates the description of hypothesis generation as categorization concerns the potentially large number of possible categories as well as diagnostic feature cues. Within medicine, Gordon (1970) estimated the number of diseases (i.e., categories) to be approximately 6,000 and the number of symptoms or measurements (i.e., features) to be approximately 20,000. To characterize hypothesis generation as categorization thus requires further specification of mechanisms by which the category search space is reduced to manageable proportions (e.g., by incorporating information about the base rates of occurrence of different categories).[1]

Historically, initial hypotheses were thought to be generated by categorization or reasoning processes using semantic knowledge. Elstein and Bordage (1979) assumed that physicians compare case information with the content of lists retrieved from a network of associations or rules that relate symptoms to particular diseases or conditions. Bordage and Zacks (1984) proposed prototype models of diagnostic feature conditions to represent physicians' knowledge of diagnostic categories. Brooks et al. (1991) more recently argued that hybrid models of memory that include memory for particular instances as well as generalized knowledge are necessary to account for existing categorization effects. They showed, for example, that personal experience with similar cases by physicians facilitated the activation of a particular diagnostic class from a provided set of possible hypotheses. Reviewing the related distinction between the use of episodic versus semantic memory traces in frequency judgments about past events, Means and Loftus (1991) similarly concluded that semantic knowledge abstracted from episodic information does not replace such individual traces but simply coexists with it as a more accessible supplement and that both types of information are subsequently used. Individual differences in hypothesis generation may thus be related not only to differences in the general knowledge representation of problem solvers but also to differences in their memory for previous cases.

Many of the formal rules embodied in the semantic representation of a domain tend to be transmitted during the early stages of formal training (e.g., in medical school). It takes additional years of clinical practice to build up a representative knowledge base of previously diagnosed cases, from which more refined semantic rules or symptom–disease associations will be abstracted that can also serve as a store of problems with known solutions. Experience may thus offer expert problem solvers the advantage of being able to supplement more analytic rule-based generation processes with case-based recognition processes when diagnosing a case. The idea that problem solvers draw analogies to previously solved cases is, of course, not new (e.g., Gentner, 1983). Ross (1987) and Ross and Kennedy (1990) demonstrated the importance of "remindings" and the use of previously solved problems in people's choices of solution algorithms for newly presented problems. Similar processes may also guide people's generation of initial diagnostic hypotheses.

Much controversy in the categorization literature has centered on the twin issues of memory representation (e.g., exemplars vs. prototypes) and memory processes (e.g., induction of abstract principles or prototypes from specific exemplars at storage vs. at retrieval). Barsalou (1990) argued that it may be difficult to differentiate between different types of memory representation on the basis of behavioral data because any difference in representation can usually be compensated by appropriate retrieval processes (however, see

[1] Artificial intelligence approaches to problem solving have had to deal with this issue. Fox (1980), who simulated clinical diagnosis with nonprobabilistic inference processes, supplemented his model with a memory mechanism reminiscent of Morton's (1970) logogen model of word recognition, which provides concepts (e.g., diagnoses) that are activated more frequently with a higher level of standing activation. This mechanism allows the population base rate of diagnostic categories to have an effect on hypothesis generation that is absent from associative network models in which hypothesis generation is driven solely by the degree of co-occurrence of feature cues and diagnostic categories (i.e., by the degree of representativeness).

Nosofsky, 1992). Our study was designed to provide empirical information about the role of some basic informational and individual-differences variables on hypothesis generation without attempting to attribute such effects conclusively to particular memory representation or memory processes, even though our results may inform future experimental hypotheses along those lines.

Our first question concerned the effects of case information, diagnosis base rates, and clinical experience on the generation of hypotheses. To determine the effects of experience on hypothesis generation, we selected a large sample of physicians with a broader range of experience than commonly found in studies of this type. To determine the effects of available information, we systematically manipulated the amount of clinical and background information available to problem solvers. *Clinical information* refers to current and previous symptoms that describe the presenting condition of a patient (e.g., complaints, family history, or substance abuse). *Background information* refers to patient information that is nonmedical and not specific to the case at hand (i.e., age, sex, or occupation). Medical texts and reference books often provide information about the incidence rates of particular diseases as a function of background variables such as age or sex. Similar to the way demographic information is used by actuaries in the determination of the expected rates of particular accidents, background information about a patient can serve to modify the overall incidence rate of particular diagnostic hypotheses by delineating a more specific subset of the general population as the relevant reference class. In this context, it may also play a critical role in deciding how physicians interpret particular clinical symptoms (Feltovich & Barrows, 1984).

In an effort to extend the characterization of hypothesis generation as categorization, we designed cases that had some interpretative ambiguity (i.e., did not have only one obvious diagnosis). For each case, at least two types of hypotheses were plausible, with one diagnosis having a higher population base rate but less severe clinical consequences than the other in an attempt to determine the effect of the overall likelihood and clinical severity of diagnostic hypotheses on their generation. Diseases or disorders differ in the relative frequency with which they occur in a population (i.e., in their base rates). However, physicians reportedly often fail to appreciate the significance of base rates when presented with questions that require the integration of numerical base rate information (Casscells, Schoenberger, & Grayboys, 1978; Eddy, 1982; Wallsten, 1981). Textbooks on clinical diagnosis, in fact, propose maxims such as "statistics are for dead men" and "the patient is a case of one" (from Eddy & Clanton, 1982), presumably to discourage physicians from incorporating base-rate information when making their diagnoses. On the other hand, base-rate neglect has been found to be at least partially a function of the way base-rate information is provided (Christensen-Szalanski & Beach, 1982; Gigerenzer, Hell, & Blank, 1988). For example, base-rate knowledge about diseases acquired through direct experience has been found to affect diagnostic judgments (Medin & Edelson, 1988). Beyth-Marom and Arkes (1983) and Christensen-Szalanski and Beach (1983) suggested that people may give probability estimates in accordance with Bayes's theorem by means other than actually applying the theorem and doing so correctly. Instead, they argued that subjects may estimate relevant conditional probabilities directly from their memory as the relative frequencies of a characteristic in a subpopulation. Memory-based heuristics that use ease or strength of recall to make relative likelihood judgments work well when memory is a veridical reflection of actual frequencies, leading to problems only when biased reporting or differential vividness distorts those memory representations (Detmer, Fryback, & Gassner, 1978; Tversky & Kahneman, 1973). With experience, physicians' memory strengths for different diagnoses may provide fairly reliable estimates of the base rates of those diagnoses in the population to which the physicians have been exposed. It is an open question, addressed in this study, whether base-rate information will be used when it is manipulated or presented in ways that allow people to use memory-based judgment strategies.

Medical students receive a variety of often-conflicting recommendations in texts on clinical diagnosis on how to deal with the overall likelihood as opposed to the clinical severity of diagnostic hypotheses. Some rules of thumb advise future doctors to bias their diagnosis in the direction of the higher base-rate category: "If you hear hoofbeats, think of horses, not zebras" or "Rare manifestations of common diseases are often more likely than common manifestations of rare diseases." On the other hand, the same textbooks also contain advice such as "The first priority is to think about diseases you cannot afford to miss." Given that there is an empirical negative correlation between the frequency and clinical severity of diseases, following this last advice may often conflict with the implications of the previous admonitions. Given the Hippocratic oath of *primum non nocere*, physicians may try to avoid "harming" their patients by being more willing to entertain hypotheses about conditions with more severe consequences than warranted by base rates. Consequently, we investigated how much relative consideration physicians give to the base rate as opposed to the clinical severity of diagnostic categories when generating hypotheses.

The second question addressed in our study concerned the size and diversity of the set of generated hypotheses. The possible number of diagnostic hypotheses that one may theoretically consider for a given set of symptoms can be large. Early diagnostic hypotheses may have a considerable impact on further information-gathering processes because new hypotheses are often not introduced at later stages. Thus, the heterogeneity of the initial set of hypotheses can be crucial for a successful diagnosis. We therefore investigated what effects our independent variables (i.e., amount of clinical and background information, diagnosis base rates and severity, and physicians' experience) would have on the heterogeneity and size of the generated set of hypotheses.

Our third question concerned stopping rules. Why do physicians decide to stop generating further diagnoses? What criteria do problem solvers use in deciding when to stop generating initial hypotheses? One possibility is some form of *satisficing*. Studying the detection of abnormalities in radiographs, Berbaum et al. (1991) found that radiologists often halt their search after finding one lesion, leaving additional lesions undetected. By analogy, physicians might stop generating hypotheses after obtaining one satisfactory explanation of the presenting symptoms. Thus, physicians would be expected to entertain more hypotheses when the clinical information is less consistent or when the presenting symptoms are less specific (i.e., "past chest infection" vs. "past recurrent dyspepsia"). However, satisficing is only one example of the kind of cost–benefit considerations that might direct the termination of hypothesis generation. Thus, the final goal of this study was to investigate the nature of the stopping rules that guided physicians in their generation of initial hypotheses.

Method

Subjects

Participants were family practitioners in South Glamorgan, Wales, United Kingdom, who were working under subcontract to the British National Health Service to provide general medical services to a registered population of patients. These physicians work in group practices that vary in size from 2 to 8, with a mean doctor–patient ratio of 1:2,000. Some practices (approximately 30%) are recognized as training practices for the postgraduate education of trainee general practitioners (GPs). All doctors working in training practices in this district were asked to participate.

Contacted physicians were either trainers (fully registered GPs recognized as teachers of trainee GPs; $n = 53$), nontrainers (fully registered GPs who were partners of trainers in training practices; $n = 41$), or trainees (recently registered medical practitioners undergoing a 12-month postgraduate training program to become GPs; $n = 37$). The response rate was 64.1% (similar for all three categories of respondents), for a total of 84 returned questionnaires. Clinical experience among fully registered GPs varied from 1 year to 40 years. Trainees had 1–12 months of experience after medical school.

Case Histories

The three case histories used in the study were based on real patients whose names and personal details were changed to protect their identity. Two pilot studies established that physicians would accept the case histories as realistic and determined plausible diagnostic hypotheses. On the basis of the results of the pilot studies, the physician on our research team in consultation with another experienced physician identified two plausible diagnostic hypotheses (A and B) for each case history. In all three cases, medical reference books showed the A diagnosis as having a higher population base rate but less severe clinical consequences than the B diagnosis. Other (O) diagnoses were also conceivable but not plausible given the symptomatic and background information. In Case 1, for example, the A diagnosis was "upper gastrointestinal disease," the B diagnosis was "ischemic heart disease," and O diagnoses included "lung condition," "anxiety," or "gall bladder disease."[2]

Using the results of the pilot studies, we created 16 versions of each case history, wherein each version had a different combination of clinical symptoms and background information (as described shortly). To provide the reader with an impression of the case histories and the information manipulations, Table 1 lists the most complete and the most reduced version for each case.

Design and Procedure

The experimental manipulation of information followed a 2 × 2 × 2 × 2 factorial design that varied the amount of clinical and background information indicative of Diagnosis A and the amount of clinical and background information indicative of Diagnosis B, with two levels (full vs. reduced) for each of the four information factors. Table 2 illustrates the design by listing the clinical and background information indicative of Diagnoses A and B, respectively, that doctors received under the full and reduced conditions. By crossing the four clinical conditions shown in Table 2 (full vs. reduced clinical information indicative of A, B, or both) with the four background information conditions also shown there (full vs. reduced background information indicative of A, B, or both), 16 different versions of each case history were generated that were presented to physicians in a between-subjects design.

Doctors filled out a paper-and-pencil questionnaire that contained one randomly selected version of each case. For each case history, they answered the following set of questions: (a) What could be wrong with this patient? Please list as many possibilities as you would consider in real life. If there is more than one, please list them in order of likelihood. (b) Which of your ideas would you explore first? Why have you chosen this one? (c) Have you encountered a similar problem before in practice? If so, what was the diagnosis on that occasion? Physicians were then asked to consider the diagnosis that they had listed first and to answer the following additional questions about it: (d) Please identify which items of the information provided first made you think of this idea. (e) Please rate the likely significance of this diagnosis for the patient. They were then prompted to consider the diagnosis that they had listed in second place and to answer questions (d) and (e) again, now in reference to their second hypothesis. Significance ratings were made by placing a mark on a 4-in.-long graphic rating scale that ranged from *very significant* to *not at all significant*.

The questionnaire was mailed to the general practitioners in our sample. A cover letter asked physicians for their voluntary participation. Physicians were instructed to answer the questions for each case as if the patient described in the case history had just walked in for a consultation. Participants returned the completed questionnaire by mail. Although responses were anonymous (except for the identification of age, sex, and level of experience), the coding of provided return envelopes allowed for the identification of participants who had failed to return their questionnaire after some period of time. These physicians received a follow-up phone call asking them for their participation a second time. Doctors were assured that their responses would not be used to evaluate them in any way and were promised a summary of the results of the study.

Results

Manipulation Checks

Several measures provided manipulation checks for our construction of A-type diagnoses as having higher population base rates and lower clinical significance than B-type diagnoses. For each case, physicians were asked if they had encountered a similar problem before in their practice.

[2] A table providing the general A, B, and O hypotheses for each case, together with a listing of the specific diagnostic labels generated by our respondents, is available on request.

Table 1
Most Complete and Most Reduced Versions of Three Case Histories

Case 1

Most Complete Version

Mr. Brooks is 45 years old, married with two daughters, and works as a long-distance truck driver. He has had a burning lower retrosternal discomfort while driving for 3 months. In the past he has been troubled by recurrent dyspepsia and frequent chest infections. He smokes 40 cigarettes a day. His father died, aged 52, following a myocardial infarction.

Most Reduced Version

Mr. Brooks is 25 years old, married with two daughters, and works as a postman. He has had a lower retrosternal discomfort for 3 months. In the past he has been troubled by frequent chest infections. He smokes 40 cigarettes a day.

Case 2

Most Complete Version

Miss Maria Curtis is 21 years old and has a son aged 2. She lives with her parents and works as a part-time shop assistant. She has had intermittent abdominal pain and diarrhea over the past 6 weeks and has lost about one half stone (7 lb) in weight. Ten years ago she made a full recovery from infective hepatitis. Her father is dying at home with bronchial carcinoma. She does not smoke.

Most Reduced Version

Mrs. Maria Curtis is 32 years old and lives with her husband and child, aged 2. She works as a part-time shop assistant. She has had intermittent diarrhea over the past 6 weeks. Ten years ago she made a full recovery from infective hepatitis. She does not smoke.

Case 3

Most Complete Version

Mr. James is 52 years old and is a social worker. He has had a constant epigastric discomfort for 3 weeks, which is usually relieved by food. He has been awakened at night by severe bouts of pain. Five years ago he was found to have a hiatus hernia, and last year sebaceous cysts were removed from his scalp. He smokes 25 cigarettes a day. His father died from gastric carcinoma.

Most Reduced Version

Mrs. James is 32 years old and works as a civil servant. She has had a constant epigastric discomfort for 3 weeks, which is usually relieved by food. Last year sebaceous cysts were removed from her scalp. She is a nonsmoker.

This question was answered in the affirmative 82% of the time. Of those who had seen a similar case before, 74% reported that their diagnosis had been an A type and only 8% that it had been a B type, consistent with our construction of A diagnoses as having higher population base rates. Eighteen percent reported that their diagnosis had been an O type.

For the first two hypotheses generated, physicians rated the clinical significance of that diagnosis for the patient. Diagnoses were either A, B, or O. Physicians' clinical significance ratings varied significantly as a function of type of diagnosis, $F(2, 247) = 9.86$, $p < .0001$, and $F(2, 236) = 29.66$, $p < .0001$, for the first and second hypotheses, respectively. On a 10-point rating scale with larger values denoting greater clinical significance, A-type diagnoses had mean significance ratings of 6.5, whereas B-type diagnoses had mean significance ratings of 8.6, consistent with our construction of B-type diagnoses as having greater clinical significance than A-type diagnoses. O-type diagnoses had significance ratings not different from those for A-type diagnoses, with a mean of 6.0.

To validate our assumptions that particular items of clinical information are indicative of Diagnosis A (but not of B) and vice versa as indicated in Table 2, we analyzed the frequency with which physicians mentioned each symptom to support either an A, B, or O diagnosis in response to the question, "Which item of the information provided first made you think of this diagnosis?" For all three cases, symptoms designed to be indicative of a particular diagnosis were mentioned significantly more frequently for that diagnosis, whereas those symptoms designed to be common were mentioned with equal frequency in support of all three diagnoses. Thus, across cases and symptoms, clinical symptoms intended to be indicative of A were mentioned 56% of the time when doctors were supporting an A diagnosis, but only 15% and 19% of the time when doctors were supporting a B or O diagnosis, respectively. Clinical symptoms intended to be indicative of B were mentioned 81% of the time when supporting a B diagnosis but only 15% and 48% of the time when supporting an A or O diagnosis, respectively. Common clinical symptoms, on the other hand, were mentioned with about equal frequency for the three types of diagnoses (i.e., 34%, 33%, and 23% of the time when doctors were supporting A, B, and O diagnoses, respectively).

Table 2
Clinical and Background Information Indicative of Diagnoses A and B Provided in the Different Information Conditions, by Case

	Information condition			
Case	Reduced for A, reduced for B	Full for A, reduced for B	Reduced for A, full for B	Full for A, full for B
Case 1[a]				
Clinical information				
Common	Lower retrosternal discomfort; for 3 months; past chest infections; smokes 40 cigarettes/day	Lower retrosternal discomfort; for 3 months; past chest infections; smokes 40 cigarettes/day	Lower retrosternal discomfort; for 3 months; past chest infections; smokes 40 cigarettes/day	Lower retrosternal discomfort; for 3 months; past chest infections; smokes 40 cigarettes/day
Indicative of A		Burning; past recurrent dyspepsia		Burning; past recurrent dyspepsia
Indicative of B			While driving; father died at age 52 of myocardial infarction	While driving; father died at age 52 of myocardial infarction
Background information				
Common	Gender: Male; married, two daughters	Gender: Male; married, two daughters	Gender: Male; married, two daughters	Gender: Male; married, two daughters
Indicative of A	Occupation: postman	Occupation: long-distance truck driver	Occupation: postman	Occupation: long-distance truck driver
Indicative of B	Age: 25	Age: 25	Age: 45	Age: 45
Case 2[b]				
Clinical information				
Common	Intermittent diarrhea; for 6 weeks; past infective hepatitis; father with bronchial carcinoma; nonsmoker	Intermittent diarrhea; for 6 weeks; past infective hepatitis; father with bronchial carcinoma; nonsmoker	Intermittent diarrhea; for 6 weeks; past infective hepatitis; father with bronchial carcinoma; nonsmoker	Intermittent diarrhea; for 6 weeks; past infective hepatitis; father with bronchial carcinoma; nonsmoker
Indicative of A		Abdominal pain		Abdominal pain
Indicative of B			Lost one half a stone (7 lb) in weight	Lost one half a stone (7 lb) in weight
Background information				
Common	Gender: female; occupation: part-time shop assistant	Gender: female; occupation: part-time shop assistant	Gender: female; occupation: part-time shop assistant	Gender: female; occupation: part-time shop assistant
Indicative of A	Married; lives with husband	Single; lives with parents; father dying at home	Married; lives with husband	Single; lives with parents; father dying at home
Indicative of B	Age: 32	Age: 32	Age: 21	Age: 21
Case 3[c]				
Clinical information				
Common	Epigastric discomfort; constant; for 3 weeks; relieved by food; past sebaceous cyst	Epigastric discomfort; constant; for 3 weeks; relieved by food; past sebaceous cyst	Epigastric discomfort; constant; for 3 weeks; relieved by food; past sebaceous cyst	Epigastric discomfort; constant; for 3 weeks; relieved by food; past sebaceous cyst
Indicative of A		Woken at night; past hiatus hernia; smokes 25 cigarettes/day		Woken at night; past hiatus hernia; smokes 25 cigarettes/day
Indicative of B			Father died from gastric carcinoma	Father died from gastric carcinoma
Background information				
Indicative of A	Gender: female; Age: 32	Gender: female; Age: 32	Gender: male; Age 52	Gender: male; Age 52
Indicative of B	Occupation: civil servant	Occupation: social worker	Occupation: civil servant	Occupation: social worker

[a] Diagnosis A: upper gastrointestinal disease. Diagnosis B: ischemic heart disease. [b] Diagnosis A: irritable bowel syndrome. Diagnosis B: inflammatory bowel disease. [c] Diagnosis A: benign upper gastrointestinal disease. Diagnosis B: upper gastrointestinal cancer.

Strength of Hypothesis Generation

Two measures served as an indication of the strength with which a particular hypothesis was generated. The first measure was the rank of the first mention of the hypothesis, with smaller ranks indicating an earlier listing and thus greater strength. In addition, physicians often listed multiple diagnostic labels for A- or B-type hypotheses, usually at different levels of specificity (e.g., dyspepsia as well as gastroesophageal reflux). We interpreted the frequency with which a physician listed different versions of a particular type of hypothesis as a second indicator of the generation strength of that hypothesis. The mean ranks and frequencies of A-, B-, and O-type diagnoses generated for each case are shown in Table 3. Because O-type diagnoses are a collection of miscellaneous other hypotheses, their frequency of generation is similar to that of B-type diagnoses. However, O-type diagnoses were generated significantly later (i.e., had higher ranks) than were B-type diagnoses.

Table 4 shows the ranks at which A- and B-type diagnoses were first mentioned, aggregated across physicians and cases. (The pattern is the same when the data are broken down by case.) The table shows that A-type diagnoses were included in the hypothesis set (i.e., had ranks of 1, 2, or 3+) 96% of the time. B-type diagnoses were included in the hypothesis set only 70% of the time. A-type diagnoses were mentioned first on the list of possible diagnoses 74% of the time. B-type diagnoses were mentioned first only 8% of the time. The most frequent rank of the first generation of a B-type diagnosis was in the third position.

Doctors frequently generate another variant of the A-type disorder as their second diagnosis. For multiple mentions of the same diagnostic hypothesis, subsequent diagnoses could become either more specific, less specific, or stay at the same level of specificity. Theories about causal inference such as Mackie's (1974) progressive localization of cause as well as common sense (i.e., the fact that more general diagnoses are more likely to be true than more specific ones) would predict that physicians who list multiple instances of the same diagnostic category should do so in a general-to-specific order. In our data, physicians acted accordingly the majority of the time (64%, 95%, and 86% of the time for Cases 1–3, respectively). In addition, deviations from this predicted pattern were less frequent for more experienced physicians, as evidenced by a positive association between experience (i.e., years of clinical practice) and the ordering of multiple diagnoses in the logical general-to-specific direction, $r(82) = .32$, .22, and .26, for Cases 1, 2, and 3, respectively (all $ps < .05$).

Table 3
Means of Rank and Frequency Measures of Strength of A, B, and O Diagnoses, by Case

	Rank			Frequency		
Case	A	B	O	A	B	O
1	1.2	3.1	3.8	2.0	0.7	1.3
2	1.6	2.4	3.5	1.2	1.3	1.5
3	1.1	3.0	3.3	2.5	0.5	0.5

Effect of Information Manipulation on Hypothesis Generation

We analyzed the effect of our four experimental information factors (clinical and background information indicative of diagnoses A and B, respectively, effect coded for full vs. reduced levels) on both the frequency with which A, B, and O diagnoses were generated as well as on the rank order of generation. The frequencies of A, B, and O diagnoses were analyzed jointly using multivariate regression. There was no evidence of an interaction between the effects of clinical and background information for any of the diagnostic frequencies in any of the three cases, indicating that physicians did not use the background information to reinterpret the significance of clinical symptoms. However, there were significant main effects, indicating that physicians used both types of information. Clinical as well as demographic background information affected the frequency with which physicians generated hypotheses, as shown in Table 5, which summarizes the mean frequencies as a function of the amount of information provided. Doctors generated more A diagnoses when they received full (rather than reduced) clinical and background information indicative of A. Although the effect was smaller, the same pattern was observed for the number of generated B diagnoses, which were more frequent when full (rather than reduced) clinical and background information indicative of B was provided. In addition, full clinical information indicative of a high-base-rate, low-consequence A diagnosis decreased the number of other low-consequence O diagnoses but did not affect the number of high-consequence B diagnoses. Full clinical information indicative of a low-base-rate, high-consequence B diagnosis, on the other hand, decreased the number of high-base-rate, low-consequence A diagnoses.

To remove interdependencies between the ranks of hypotheses (i.e., negative correlations due to the fact that a higher rank for one hypothesis resulted in lower ranks for the other two hypotheses), we modeled the rank data by a log-linear formulation of Pendergrass and Bradley's (1960) ranking model. In analogy to Luce's (1959) pairwise choice model, which assumes that choice between two alternatives is a probabilistic function of the alternatives' selection strength, the ranking model assumes that the rank order in which diagnoses are generated (e.g., listing an A diagnosis first, followed by a B diagnosis, and then by an O diagnosis) is a probabilistic function of the underlying generation strength of each diagnosis. The probability of observing the three hypotheses A, B, and O being generated in the A–B–O order, for example, can be represented as the product of the three implicit pairwise comparisons: generating A before B, A before O, and B before O. More formally, the probability of observing the A–B–O rank order is equal to $\Pr(A, B, O) = C\,\{\pi_A/(\pi_A + \pi_B)\}\{\pi_A/(\pi_A + \pi_O)\}\{\pi_B/(\pi_B + \pi_O)\}$, where π_A, π_B, and π_O represent generation strength parameters for the A, B, and O diagnoses, respectively, under the constraint that $\pi_A + \pi_B + \pi_O = 1$, and C is a normalizing constant. There are six distinct orders in which the three A, B, or O diagnoses can occur (ABO, AOB, BAO, BOA, OAB, and OBA). The relative frequencies with which these six rank

Table 4
Frequencies (Across Physicians and Case Histories) of First Mention of Diagnoses A and B, respectively, in the First (1), Second (2), Third or Later (3+) Positions or Not at All (0)

Rank of Diagnosis A	Rank of Diagnosis B				
	1	2	3+	0	Sum
1	X	44	85	69	*198* (79%)
2	11	X	15	5	*31* (12%)
3+	7	5	2	0	*14* (5%)
0	3	4	1	1	*9* (4%)
Sum	*21* (8%)	*53* (21%)	*103* (41%)	*75* (30%)	*252*

Note. X = combination is impossible. Numbers in parentheses show the percentage of A and B diagnoses generated at each rank.

orders occurred were computed for each of the 16 experimental information conditions outlined in Table 2. This 16 × 6 (information condition by rank order) frequency table was analyzed with a multinomial logit regression model, using standard maximum-likelihood methods (for details, see Critchlow, Fligner, & Verducci, 1991, and Fienberg & Larntz, 1976). The effect-coded experimental information conditions served as independent variables. Model fits were assessed, using likelihood-ratio chi-square statistics, by comparing the observed and expected frequencies of rank orders.

For all three cases, a main effects model of the four information conditions provided a satisfactory representation of the ranking data. Estimates of the generation strength parameters for A, B, and O diagnoses (i.e., π_A, π_B, π_O) are shown in Table 5 as a function of information conditions and averaged across cases. The results were highly similar to those for the frequency data.[3] The generation strength of a particular diagnosis (A or B) increased when more (full) diagnosis-relevant clinical and background information was provided. Moreover, the increment in generation strength of A diagnoses when full clinical or background information indicative of A was provided came at the expense of both B and O generation strength, with a greater decrement in O than in B generation strength. On the other hand, the increment in generation strength of B diagnoses when full clinical or background information indicative of B was provided came at the exclusive expense of A generation strength. This asymmetric effect of diagnosis-inconsistent information on generation strength paralleled the asymmetry in the effect on generation frequency.

Effect of Availability of Diagnosis on Hypothesis Generation

For each case, physicians were asked if they had encountered a similar problem before in their practice. This question was answered in the affirmative 82% of the time. Physicians then gave their final diagnosis of that case. Similarity of a case to previous cases will, of course, be influenced by the amount of clinical and background information provided about the case. Across the 16 experimental information conditions, doctors reported diagnosing the similar case as an A, B, or O type 74%, 8%, and 18% of the time, respectively. Most of the O-type similar cases were reported for those information conditions in which clinical and background information indicative of A and B were reduced. More information indicative of an A or B diagnosis increased the likelihood that this diagnosis would be reported as the similar case, an effect that was statistically significant for the amount of clinical information indicative of A, $\chi^2(2, N = 252) = 27.07, p < .0001$. Thus, we tested for the effect of availability of a similar diagnosis on hypothesis generation after statistically controlling for the effect of the four information conditions (by including them as prior variables in the analyses). We again analyzed both the frequency and rank measures of hypothesis generation strength. For the frequency data, all three cases showed a significant increase in explained variance when availability dummy variables indicating reported prior experience with an A, B, or O diagnosis, respectively, were added to the multivariate regression of diagnosis frequency on the information factors, Fs(3, 67) = 3.27, 3.18, and 3.86 for Cases 1–3 respectively (all ps < .05). Availability worked as one might expect, with prior experience of an A case leading to a stronger (i.e., more frequent) generation of A-type diagnoses. In the same way, prior B or O cases led to a stronger generation of B- and O-type diagnoses, respectively.

The rank data, modeled as described earlier, were analyzed in a similar way. The decrease in the chi-square statistics obtained after adding dummy variables corresponding to a prior A-, B-, or O-case diagnosis to the experimental information design variables served as the test statistic. The additional effect of availability was significant for all three cases, χ^2s$(6, N = 252) = 13.1$, 18.4, and 14.3, for Cases 1–3, respectively, in each case creating an increase in the generation strength of a hypothesis when a similar case with the same hypothesis had been seen before.

[3] This similarity in findings is not a result of the fact that the ranking and frequency data were derived from the same list of generated hypotheses. The two derived variables are conceptually distinct, and differences in results would have been theoretically possible.

Table 5
Average Frequencies and Strength Parameters of Generation of A-, B-, and O-Type Diagnoses as a Function of Experimental Information Condition, Across Cases

	Generation frequency			Generation strength		
	Diagnosis			Diagnosis		
Information condition	A	B	O	A	B	O
Clinical for A						
Full	**2.19**	0.80	0.84	**.883**	.065	.051
Reduced	1.73	0.77	**1.40**	.623	**.146**	**.224**
Clinical for B						
Full	1.81	**0.87**	1.10	.755	**.126**	.119
Reduced	**2.05**	0.73	1.06	**.867**	.051	.082
Background for A						
Full	**2.04**	0.77	1.01	**.861**	.063	.075
Reduced	1.81	0.85	1.15	.768	**.101**	**.130**
Background for B						
Full	1.91	**0.93**	1.06	.798	**.100**	.102
Reduced	1.89	0.69	1.11	**.837**	.069	.104

Note. Entries that are significantly larger for each full–reduced pair appear in boldface. Summaries of the individual case analyses with significance levels for the overall multivariate *F* tests and chi-square tests are available on request.

Effect of Experience on Hypothesis Generation

In addition to the positive association between clinical experience and doctors' ordering of multiple hypotheses in the logical general-to-specific direction as discussed earlier, our data also revealed a significant positive correlation between years of clinical practice and the availability of a similar case, rs(82) = .24, .22, and .26, for Cases 1–3, respectively (all ps $< .05$). Physicians with more experience were more likely to report that they had seen a similar case before, resulting in an indirect effect of experience on hypothesis generation via increased availability of previous diagnoses as outlined in the last section. After statistically controlling for this effect of availability, however, there was no further effect of experience on either the frequencies or the ranks of diagnostic hypotheses.[4]

Size of Hypothesis Set and Stopping Rule

As discussed earlier, our manipulations of clinical and background information affected both the order and the frequencies with which A, B, and O diagnoses were generated. However, these information variables had no effect on the overall size of the diagnostic set (i.e., on the total number of generated hypotheses). Physicians generated a similar number of diagnoses regardless of the amount of clinical and background information provided in the different experimental conditions.

The absence of any apparent effect of the information manipulation on number of generated hypotheses could have been the result of two opposing effects canceling each other out. More information may provide more cues that trigger associated hypotheses otherwise not generated but may also constrain the set of hypotheses that are plausible given the presented symptoms.[5] The latter effect would predict that the heterogeneity of the hypothesis set will decline as more diagnostic information is provided. Heterogeneity of the hypothesis set was operationalized as the standard deviation of diagnoses in the set, wherein different types of diagnoses were assigned different numerical codes. The degree of heterogeneity was affected by the information variables, but not in the way predicted by the aforementioned hypothesis. Doctors generated a more heterogeneous set of hypotheses when they received less clinical information indicative of the high-base-rate A diagnosis, $F(1, 244) = 16.67, p < .0001$, but also when they received more clinical and background information indicative of the lower base-rate B diagnosis, $F(1, 244) = 11.53$, $p < .001$. Thus, the two opposing effects explanation did not account for the absence of an effect of the information variables on hypothesis set size.

Variation in the size of the hypothesis set occurred between cases, which was smaller for Case 3 than for Case 1, $F(2, 247) = 4.42$, $p < .02$, and between physicians, $F(83, 166) = 2.94$, $p < .0001$. These individual differences in the size of the hypothesis set were not related to differences in physicians' experience but were highly stable across cases. Intercorrelations among the total number of hypotheses each physician generated for Cases 1–3 were highly significant, rs(82) = .47, .42, and .39 for intercorrelations 1–2, 1–3, and 2–3, respectively (all ps $< .0001$), even though the particular levels of clinical and background information seen by a given doctor differed across cases.[6] These results suggest that the

[4] The effect of availability of a previous case on generation of the corresponding hypotheses, on the other hand, remained significant after statistically controlling for years of experience, in addition to the amount of clinical and background information for several frequency-of-diagnoses measures, Fs(1, 66) = 3.83 and 4.17 for A and O diagnoses for Case 1 (ps $< .05$); Fs(1, 66) = 5.97 and 2.87 for A and B diagnoses for Case 3 (ps $< .10$).

[5] We thank Lee Ross for suggesting this hypothesis.

[6] Differences in the number of hypotheses generated by doctors

size of the initial hypothesis set is more strongly influenced by individual differences between problem solvers than by the information available about the problem. The only variable that correlated with the size of the hypothesis set for all three cases was the rank of the first mention of a B diagnosis, $rs(82) = .44$, .40, and .66 for Cases 1, 2, and 3, respectively. This positive relationship persisted after statistically controlling for the experimental information factors. The longer it took doctors to generate their first alternative B diagnosis, the more diagnoses they produced. Across cases, physicians generated on average 2.8 diagnoses prior to their first B diagnosis but produced only 0.1 additional diagnoses after their first B diagnosis, a number that is not significantly different from zero. There was variation in both the total number of hypotheses ($SD = 1.32$) and in the rank of the first B diagnosis ($SD = 1.77$). That is, the positive relationship between the size of the hypothesis set and the rank of the first B diagnosis did not arise because physicians routinely generated only four hypotheses (e.g., the size of their short-term memory; Crowder, 1976), with the fourth hypothesis always being of a B type. O diagnoses were generated on average at higher ranks than were B diagnoses (see Table 3), but there was no corresponding correlation between the rank of the first O diagnosis and the size of the hypothesis set. The average ranks of diagnoses in Table 3 are confounded with the size of the hypothesis set. Thus, differences in the percentile ranks of the first generation of B versus other diagnoses speak more clearly to the question of whether physicians were more likely to stop their generation of hypotheses after generating a B diagnosis than other diagnoses. The mean percentile rank of the first B diagnoses was significantly greater than that of O diagnoses, $t(158) = 5.02, p < .0001$.

Discussion

Given the importance of early diagnostic hypotheses for ultimate diagnostic success, what would one wish physicians' hypothesis generation to look like? First, strength of generation ought to be responsive to available information, clinical as well as background symptoms. Second, physicians ought to consider both the likelihood and severity of diagnoses when generating diagnoses, in a way that reflects their loss functions for hits and misses in both categories. Finally, because additional diagnoses are often not introduced at later stages (Barrows et al., 1982), we would want physicians to initially generate all diagnoses that have a reasonable chance of being accurate (in our study, both A and B types). In light of the bad reputation that problem solvers have acquired for their performance in later stages of the process (e.g., the confirmation bias in hypothesis testing; Bruner et al., 1956; Nisbett & Ross, 1980), it may come as a surprise that the general practitioners of our study largely satisfied these characteristics of good hypothesis generation.

Information Usage

Both clinical and background information indicative of a diagnosis had significant effects on the strength with which that hypothesis was generated. Diagnosis-consistent clinical information increased the strength of both A and B diagnoses. Our manipulation of background information (i.e., information that affected the case-specific likelihood of particular diagnoses) also had significant effects on the strength with which these diagnoses were generated. Thus, likelihood information that is provided not as a numerical probability but by defining a relevant reference class in memory for which differential estimates of the dependent variable exist or for which such relative frequency estimates can be generated appears to be used. For such a segmentation process to operate, it is necessary that people initially encode the value of the dimension along which the segmentation is supposed to occur. A causal connection between the background (segmentation) variable and the judged (dependent) variable (e.g., diagnosis incidence rates being affected by occupation) will probably ensure such encoding, but knowledge that the variable under judgment occurs with differential frequency for different categories of an incidental background variable also seems to be sufficient (e.g., diagnosis incidence rates differing as a function of age and sex). Either direct experience or instruction (e.g., by medical reference books) can provide such information. In the absence of either (e.g., in Kahneman & Tversky's, 1973, cab problem), people may ignore the segmentation variable, meaning that they will not encode or use it as a memory-segmentation variable and that they will not use numerical likelihood information about the dependent variable that is provided as a function of it. Encountering causal rather than incidental segmentation (base-rate) variables may encourage both more likely encoding and greater usage of the variable (e.g., Bar-Hillel, 1980).

In addition to effects of the manipulated information factors, we found consistent and strong effects attributable to physicians' personal prior experience. Doctors generated A-, B-, or O-type diagnoses, respectively, earlier and more frequently in instances in which they had seen a corresponding similar case before. These results are consistent with those of medical studies about the diagnosis of skin disorders in which the physician having seen a similar case activates a diagnostic class (Brooks et al., 1991; Norman, Rosenthal, Brooks, Allen, & Muzzin, 1989). Our study extends these findings by demonstrating that having seen a similar case not only facilitates the choice of that diagnostic category from a provided set of possible diagnoses but that it also increases the spontaneous generation of that hypothesis.

Weber, Goldstein, and Busemeyer (1991) outlined some benefits of incorporating considerations of memory representation into models of judgment, decision making, and problem solving. We hope that the results of this study will also encourage further considerations of the memory representations of physicians and other experts, in particular considerations of the use of memory for prior cases. Wason (1983) suggested that many inferential processes may not be instances of formal reasoning but instead applications of relevant past experience. Similar mechanisms may underlie the

and consistency in numbers across cases were not an artifact of the size of physicians' handwriting. Sometimes, doctors filled only 25% of the available space on the questionnaire; at other times, they continued to write hypotheses in the margin.

generation of initial working hypotheses. Memory integration of past experience, for example the differential accumulation of prior experiences of high-frequency A-type and low-frequency B-type diagnoses, may give rise to the base-rate sensitivity observed in this study. Beyth-Marom and Arkes (1983) and Christensen-Szalanski and Beach (1983) made a similar memory-based argument to explain the base-rate sensitivity found by Christensen-Szalanski and Beach (1982). Questioning the assumption that direct experience of a relationship between base rate and diagnostic information turns people into true Bayesians, they argued that, instead, certain information conditions may allow people to take advantage of simple memory strategies with the result that their judgments look as if they had applied Bayes's theorem. This distinction has important implications for attempts to improve judgment or problem-solving performance. Giving physicians training in formal statistical reasoning may not be the only or perhaps the best way of encouraging them to incorporate base-rate information. Instead, it may be better to structure their information environment in a way that allows them to take best advantage of their memory base and memory-based judgment processes, which may enable them to make more normative judgments and decisions. Assuming that base-rate effects are mediated by processes operating on memory for prior instances would explain why manipulations of base rates that allow people to capitalize on their prior knowledge when making judgments (e.g., the manipulation of background symptoms in our study, or Study 2 of Gigerenzer et al., 1988) find people sensitive to base-rate information, whereas studies that provide base-rate information in numerical form tend to find base-rate neglect.

Attributing base-rate sensitivity to memory storage and retrieval processes involving prior instances is also consistent with Wallsten (1981), who found base-rate sensitivity in experienced physicians but not in medical students. In our study we found a positive relationship between experience and availability of diagnostic hypotheses (i.e., increased likelihoods of having seen a similar case before, resulting in an indirect effect on hypothesis generation) but no additional effects of experience on generation strength after controlling for this availability effect, reinforcing the hypothesis that increased base-rate sensitivity with experience is being mediated by experts' more representative memory base for previously diagnosed cases.

Likelihood–Severity Considerations

Medical diagnosis, especially in a family practice setting, can be seen as a signal-detection problem. The great majority of cases seen by physicians are either high-likelihood, low-consequence routine diagnoses of the A type or involve the management of chronic (known) diseases. Only a small fraction of cases are medical problems with serious consequences if they go undetected (i.e., B types). The main challenge for the family practitioner is to detect those cases (the signals) among the noise of routine cases. Given the empirical negative correlation between the likelihood and severity of diseases, a fact of which both physicians (Schiffmann, Cohen, Nowik, & Selinger, 1978) and the general public (Weber & Hilton, 1990) seem to be well aware, hypothesis generation will thus involve an implicit decision about the relative importance or priority of likelihood as opposed to severity considerations. The signal-detection view suggests that this decision may depend on the costs and benefits associated with hits and misses in the two categories. As discussed earlier, clinical textbooks provide aspiring physicians with often-conflicting advice regarding the relative importance of these two attributes. In addition, people's statements of perceived relative importance do not always agree with the tradeoffs implicit in their choices or judgments. Medin and Edelson (1988), for example, found that some subjects reported using the severity of symptoms when choosing a diagnosis, whereas their responses showed that they were not.

The general practitioners in our study seemed to be sensitive first and foremost to the likelihood or base rate of hypotheses when generating diagnoses.[7] A-type diagnoses were listed both earlier and more numerously than were B-type diagnoses. However, there was an asymmetry in the effect of diagnosis-inconsistent clinical information on A- and B-type diagnoses. Looking at the frequency-of-diagnoses data in Table 5, which were not constrained to sum to any constant, more clinical information indicative of a B diagnosis reduced the frequency of A diagnoses, but not vice versa. More clinical information indicative of an A diagnosis reduced only the frequency of O diagnoses but not B diagnoses. This might have been partly the result of a floor effect for the generation of B diagnoses. Alternatively, it might also have been the greater clinical significance of B-type diagnoses that diminished the relative generation of more common but less severe diagnoses when symptoms indicative of B were present. The fact that tradeoffs in the frequency of generated hypotheses occurred at all indicates some cognitive limitations on the part of the physicians. Doctors could have generated more B diagnoses without reducing the frequency with which they generated A diagnoses.

One result of the asymmetric tradeoffs in the frequency of different diagnoses was an asymmetric effect of A versus B information on the heterogeneity of the hypothesis set. Heterogeneity of the diagnostic set was greatest when more information indicative of a high-severity B diagnosis and less information indicative of a high-likelihood A diagnosis was present. This suggests that high-base-rate diagnoses (but not high-severity diagnoses) seem to be generated by default.

Role of Experience

Experienced physicians were more likely than novices to list multiple instances of the same general hypothesis in a general-to-specific order, consistent with expert–novice differences in hypothesis generation skills postulated by Camerer and Johnson (1991) and with expert–novice differences in knowledge representation reported by Joseph and Patel (1990) and Pauker et al. (1976).

[7] This and related observations need be interpreted with some caution because the likelihood and severity of hypotheses in our study were correlated rather than independently manipulated.

Joseph and Patel (1990) observed that expert physicians generated accurate diagnostic hypotheses significantly faster than did novices, a difference in the time course of hypothesis generation similar to the speed differences found in the generation of diagnoses by masters versus beginning chess players (Chase & Simon, 1973) as well as experts versus novices solving physics problems (Larkin, McDermott, & Simon, 1980). Norman et al. (1989) argued for a two-process model of diagnosis, with a fast associative pattern-recognition process in which the set of presenting symptoms is considered as a whole, and a slower analytical feature-by-feature analysis, activated after the failure of the pattern-recognition process. Using response time measures of diagnoses made by expert and novice dermatologists, Norman et al. (1989) found that improvements in diagnostic accuracy with clinical experience were not the result of improvements in the second slower analytic process but of improvements in the operation of the more rapid episodic pattern-recognition process. The results of our study are consistent with this explanation. Experience (or rather, its imperfect proxy: years of clinical practice) affected hypothesis generation by making it more likely that doctors would recall having seen a similar case before, presumably because clinical experience had enlarged their memory for prior cases. After controlling for this effect, experience held no further advantage.

Stopping Rule

The total number of working hypotheses doctors generated in our study was around four (ranging from one to eight), which was highly similar to the number of hypotheses generated by endocrinologists in the course of unconstrained "think-aloud" protocols (Joseph & Patel, 1990) and to the number of hypotheses generated by internists in encounters with live patients (Barrows et al., 1982). This number of generated hypotheses was also highly similar to the number of hypotheses found to be considered by physicians at any given time during hypothesis testing (Elstein et al., 1978) regardless of the complexity of the problem, a result usually attributed to doctors' cognitive processing limitations (e.g., short-term memory capacity). Thus, it seems that problem solvers may limit the size of their initial hypothesis set to the number of hypotheses they can use in subsequent problem-solving stages. If so, then individual differences in cognitive processing capacity or processing strategies may account for the observed stable individual differences between doctors in the size of their hypothesis sets across cases.

Caveats and Future Research Questions

In a theoretical analysis, Elstein and Bordage (1979) suggested that physicians may arrive at initial hypotheses by two routes, first generating one or more hypotheses by association from the symptom cues and then thinking about competitors to these hypotheses, presumably by using associations between diagnoses. Our study provides some empirical evidence consistent with this view. Although the presence of symptom cues indicative of A had a strong effect on the rank and frequency with which A diagnoses were generated, the corresponding effect of B symptoms on the generation of B diagnoses was much weaker. Nonetheless, there was some evidence to suggest that physicians frequently continued to generate hypotheses until their set included a low-probability, high-significance diagnosis. Doctors might have been influenced by the greater severity of B diagnoses to include one in their set or by the fact that it explained symptoms otherwise unaccounted for. These alternative explanations will need to be distinguished. The generality of this or other stopping criteria (Busemeyer & Rapoport, 1988) across different types of problems and response modes will also have to be established. A more general theory should, for example, examine to what extent the stopping criterion is influenced by the costs of generating additional hypotheses versus the costs of omission of hypotheses (Böckenholt & Kroeger, in press). The dynamic nature of a (medical) diagnosis in which hypotheses can be added or deleted over time also necessitates further examination of the function of any stopping criterion for the initial generation of working hypotheses.

The order in which information is presented has been found to affect performance in a variety of cognitive tasks from free recall to belief updating. The order in which information presented to doctors was kept constant in our study, partly to constrain the size of the experimental design and partly as a reflection of some "natural" order in which doctors may obtain information in an actual consultation. Effects of the order of information on hypothesis generation thus await further study.

Methodology

The orientation of this study was descriptive rather than prescriptive or normative. Our method constituted an attempt to find some middle ground between field studies and protocol analysis methods on the one hand and controlled laboratory experimentation on the other hand. We used stimulus material that was based on actual patients and respondents who were practicing physicians in order to achieve realism and external validity. However, by constructing and manipulating the stimulus material in systematic ways and constraining doctors' responses with a structured set of questions, we tried to keep our dependent measures manageable and interpretable, thus allowing us to test a large sample of physicians with a broad range of experience. Overall, our results suggest that such a middle ground paradigm holds promise in providing interpretable yet generalizable data. Most results were consistent across all three cases, suggesting underlying processes of hypothesis generation that transcend case specificity.

Summary

The processes of hypothesis generation observed in our sample of medical problem solvers seemed to have many desirable characteristics. Hypothesis generation was sensitive to both clinical information and background factors.

Physicians responded more to the likelihood of diagnoses than to their severity, but they frequently continued to generate hypotheses until their set included at least one higher severity hypothesis. When mentioning a particular hypothesis at multiple levels of specificity, doctors generated them in a logically consistent general-to-specific order most of the time and increasingly so with more clinical experience.

Deviations from this order, particularly in physicians with less clinical experience, as well as other instances of "suboptimal" hypothesis generation (e.g., generating O-type diagnoses early on the list) could be explained as by-products of the operation of similarity-based generation processes that are based on memory for prior cases. In contrast to more analytic semantic processing of information, similarity-based processing offers no guarantees of logical consistency, and its accuracy depends on the extent to which the physician's episodic knowledge base is representative of population frequencies. On the positive side, as demonstrated in our study, similarity-based processing provides a simple mechanism for the generation of initial diagnostic hypotheses that, by capitalizing on prior experience, allows for natural usage of clinical as well as base-rate information.

References

Abelson, R. P., & Lalljee, M. (1988). Knowledge structure and causal explanation. In D. J. Hilton (Ed.), *Contemporary science and natural explanation* (pp. 23–41). New York: New York University Press.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44,* 211–233.

Barrows, H. S., & Bennett, K. (1972). The diagnostic (problem solving) skill of the neurologist: Experimental studies and their implications for neurological training. *Archives of Neurology, 26,* 273–277.

Barrows, H. S., Norman, G. R., Neufeld, V. R., & Feightner, J. W. (1982). The clinical reasoning of randomly selected physicians in general medicine practice. *Clinical and Investigative Medicine, 5,* 49–55.

Barsalou, L. W. (1990). On the indistinguishability of exemplar memory and abstraction in category representation. In T. K. Srull & R. S. Wyer (Eds.), *Advances in social cognition* (Vol. 3, pp. 61–88). Hillsdale, NJ: Erlbaum.

Berbaum, K. S., et al. (1991). Time course of satisfaction of search. *Investigative Radiology, 26,* 640–648.

Beyth-Marom, R., & Arkes, H. R. (1983). Being accurate but not necessarily Bayesian: Comments on Christensen-Szalanski and Beach. *Organizational Behavior and Human Performance, 31,* 255–257.

Böckenholt, U., & Kroeger, K. (in press). The effect of time pressure in multiattribute binary choice tasks. In O. Svenson & A. J. Maule (Eds.), *Time pressure and stress in human judgment and decision making.* Hillsdale, NJ: Erlbaum.

Bordage, G., & Zacks, R. (1984). The structure of medical knowledge in the memories of medical students and general practitioners: Categories and prototypes. *Medical Education, 18,* 406–416.

Brooks, L. R., Norman, G. R., & Allen, S. W. (1991). Role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General, 120,* 278–287.

Bruner, J. S., Goodnow, J., & Austin, G. A. (1956). *A study in thinking.* New York: Wiley.

Busemeyer, J. R., & Rapoport, A. (1988). Psychological models of deferred decision making. *Journal of Mathematical Psychology, 32,* 91–134.

Camerer, C., & Johnson, E. J. (1991). The process–performance paradox in expert judgment: How can experts know so much and predict so badly? In A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limitations* (pp. 101–129). Cambridge, England: Cambridge University Press.

Casscells, W., Schoenberger, A., & Grayboys, T. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine, 299,* 999–1000.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 1,* 55–81.

Christensen-Szalanski, J. J. J., & Beach, L. R. (1982). Experience and the base-rate fallacy. *Organizational Behavior and Human Performance, 29,* 270–278.

Christensen-Szalanski, J. J. J., & Beach, L. R. (1983). Believing is not the same as testing: A reply to Beyth-Marom and Arkes. *Organizational Behavior and Human Performance, 31,* 258–261.

Critchlow, D. E., Fligner, M. A., & Verducci, J. S. (1991). Probability models on ranking. *Journal of Mathematical Psychology, 35,* 294–318.

Crowder, R. B. (1976). *Principles of learning and memory.* Hillsdale, NJ: Erlbaum.

Detmer, D. E., Fryback, D. G., & Gassner, K. (1978). Heuristics and biases in medical decision-making. *Journal of Medical Education, 53,* 682–683.

Dudley, H. A. F. (1970). The clinical task. *Lancet, 2,* 1352–1354.

Dudley, H. A. F. (1971). Clinical method. *Lancet, 3,* 35–37.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* Cambridge, England: Cambridge University Press.

Eddy, D. M., & Clanton, C. H. (1982). The art of diagnosis: Solving the clinicopathological exercise. *New England Journal of Medicine, 306,* 1263–1268.

Elstein, A. S., & Bordage, G. (1979). Psychology of clinical reasoning. In G. Stone, F. Cohen, & N. Adler (Eds.), *Health psychology: A handbook* (pp. 333–367). San Francisco: Jossey-Bass.

Elstein, A. S., Kagan, N., Shulman, L. S., Jason, H., & Loupe, M. J. (1972). Methods and theory in the study of medical inquiry. *Journal of Medical Education, 47,* 85–92.

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

Evans, D. A., & Gadd, C. S. (1989). Managing coherence and context in medical problem-solving discourse. In D. A. Evans & V. L. Patel (Eds.), *Cognitive science in medicine* (pp. 211–255). Cambridge, MA: MIT Press.

Feltovich, P. J., & Barrows, H. S. (1984). Issues of generality in medical problem solving. In H. G. Schmidt & M. L. De Volder (Eds.), *Tutorials in problem-based learning: A new direction in teaching the health professions* (pp. 128–142). Assen, Netherlands: Van Gorcum.

Fienberg, S. E., & Larntz, K. (1976). Log-linear representations for paired and multiple comparison models. *Biometrika, 63,* 245–254.

Fox, J. (1980). Making decisions under the influence of memory. *Psychological Review, 87,* 190–211.

Gale, J., & Marsden, P. (1982). Clinical problem-solving: The beginning of the process. *Medical Education, 16,* 22–26.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155–170.

Gettys, C. F., & Fisher, S. D. (1979). Hypothesis plausibility and hypothesis generation. *Organizational Behavior and Human Performance, 24,* 93–110.

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 513–525.

Gordon, B. L. (1970). Terminology and content of the medical record. *Computational Biomedical Research, 3,* 436–444.

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

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94,* 211–228.

Larkin, J., McDermott, J., & Simon, D. P. (1980). Expert and novice performance in solving physics problems. *Science, 208,* 1335–1342.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* New York: Wiley.

Mackie, J. L. (1974). *The cement of the universe.* London: Oxford University Press.

Means, B., & Loftus, E. F. (1991). When personal history repeats itself: Decomposing memory for recurring events. *Applied Cognitive Psychology, 5,* 297–318.

Medin, D. L., & Edelson, S. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General, 117,* 68–85.

Morton, J. (1970). A functional model for memory. In D. A. Norman (Ed.), *Models of human memory* (pp. 203–260). San Diego, CA: Academic Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice Hall.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice Hall.

Norman, G. R., Rosenthal, D., Brooks, L. R., Allen, S. W., & Muzzin, L. J. (1989). The development of expertise in dermatology. *Archives of Dermatology, 125,* 1063–1068.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honor of William K. Estes* (Vol. 1, pp. 149–167). Hillsdale, NJ: Erlbaum.

Pauker, S. G., Gorry, G. A., Kassirer, J. P., & Schwartz, W. B. (1976). Towards the simulation of clinical cognition. *American Journal of Medicine, 60,* 981–996.

Pendergrass, R. N., & Bradley, R. A. (1960). Ranking in triple comparisons. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp. 331–351). Stanford, CA: Stanford University Press.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 629–639.

Ross, B. H., & Kennedy, P. T. (1990). Generalizing from the use of earlier examples in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 42–55.

Schiffmann, A., Cohen, S., Nowik, R., & Selinger, D. (1978). Initial diagnostic hypotheses: Factors which may distort physicians' judgment. *Organizational Behavior and Human Performance, 21,* 305–315.

Simon, H. A. (1973). The structure of ill structured problems. *Artificial Intelligence, 4,* 181–201.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5,* 207–232.

Wallsten, T. S. (1981). Physician and medical student bias in evaluating diagnostic information. *Medical Decision Making, 1,* 145–164.

Wason, P. C. (1983). Realism and rationality in the selection task. In J. Evans (Ed.), *Thinking and reasoning: Psychological approaches* (pp. 54–81). London: Routledge & Kegan Paul.

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content.* London: Batsford.

Weber, E. U., Goldstein, W. M., & Busemeyer, J. R. (1991). Beyond strategies: Implications of memory representation memory processes for models of judgment and decision making. In W. E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays in honor of Bennet B. Murdock* (pp. 75–100). Hillsdale, NJ: Erlbaum.

Weber, E. U., & Hilton, D. J. (1990). Contextual effects in the interpretations of probability words: Perceived base rate and severity of events. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 781–789.

Received June 26, 1992
Revision received November 30, 1992
Accepted December 8, 1992 ■