9

Categorical and Probabilistic Reasoning in Medical Diagnosis

Peter Szolovits and Stephen G. Pauker

In the mid-1970s, when Gorry left M.I.T. to go to Baylor College of Medicine, Peter Szolovits took over as head of the Clinical Decision-Making Group at Project MAC (now known as the Laboratory for Computer Science). He renewed ties with the collaborators at Tufts University with whom Gorry had previously worked (Pauker, Schwartz, and Kassirer). The following chapter is an early result of those developing ties. It was written for a special issue of Artificial Intelligence *that dealt solely with applications of AI in biomedicine (Sridharan, 1978). In the article Szolovits and Pauker review the lessons of the major four AIM programs of the early 1970s.*

The review begins by noting that medical decision making can be viewed along a spectrum, with categorical (or deterministic) reasoning at one extreme and probabilistic (or evidential) reasoning at the other. The authors discuss classical flow charts as the prototype of categorical reasoning and decision analysis as the prototype of probabilistic reasoning. Within that context they compare MYCIN, PIP, CASNET, and INTERNIST—the four systems described in Chapters 5 through 8. They note that, although all four systems can exhibit impressive expertlike behavior, none of them is capable of truly expert reasoning. They argue that a program that can demonstrate expertise in the area of medical consultation will have to use a judicious combination of categorical and probabilistic reasoning—the former to establish a sufficiently narrow context and the latter to make comparisons among hypotheses and eventually to recommend therapy. We include the paper here because it nicely summarizes and integrates the

From *Artificial Intelligence,* 11: 115–144 (1978). Copyright © 1978 by North-Holland Publishing Company. All rights reserved. Used with permission.

discussions of the systems in the four preceding chapters. By citing the limitations of the early systems, this article helped define and clarify some of the research issues that evolved later in the decade and are discussed in subsequent chapters.

9.1 Introduction

How do practicing physicians make clinical decisions? What techniques can we use in the computer to produce programs that exhibit medical expertise? Our interest in these questions is motivated by our desire:

1. to provide (by computer) expert medical consultation to general practitioners or paramedical personnel in communities where such consultation is normally unavailable;
2. to come to understand the reasoning processes of expert doctors so that we may improve the teaching of their skills to medical students; and
3. to advance the techniques of artificial intelligence, especially as applied to medicine (AIM), to support our other goals.

In other publications, we have described research by our group on programs to take the history of the present illness of a patient with renal disease (Pauker and Gorry, 1976; Szolovits and Pauker, 1976) and to advise the physician in the administration of the drug digitalis to patients with heart disease (Gorry et al., 1978; Silverman, 1975; Swartout, 1977). Here, we would like to review the reasoning mechanisms[1] used by our own programs, by other AI programs with medical applications, and, by inference, by physicians.

9.2 Categorical and Probabilistic Decisions

Most decisions made in medical practice are straightforward. Whether the physician is taking a history of a patient's illness, performing a routine physical examination, or ordering a standard battery of laboratory tests, he or she makes few real decisions. To a large extent his or her expertise

[1] In this discussion, we take *reasoning* to be synonymous with *decision making*. Although the former is a broader term, we are specifically concerned with that aspect of reasoning that yields medical decisions. An earlier review of work in this area was made by Pople et al. (1975).

consists of mastery of the appropriate set of *routines* with which he or she responds to typical clinical situations.

This view is corroborated, in part, by the observed differences between the diagnostic approach of a medical student or newly minted doctor and that of a practicing expert. The novice struggles "from first principles" initially to propose plausible theories and then to rule out unlikely ones, whereas the expert simply recognizes the situation and knows the appropriate response. We might say that the expert's knowledge is *compiled* (Rubin, 1975; Sussman, 1973). Similar differences have even been noted among expert consultants in different specialties when they are presented the same case, and even between the performance of the same consultant on cases within compared to cases outside his or her specialty. The expert doctor dealing with a case within his or her own specialty approaches the case parsimoniously; the expert less familiar with the case resorts to the more general diagnostic style associated with the nonexpert (Miller, 1975).

An important characteristic of expert decision making, then, is the use of an appropriate set of routines or rules that apply to the great majority of clinical situations. We shall identify this as *categorical reasoning.*[2] A categorical medical judgment is one made without significant reservations: if the patient's serum sodium is less than 110 mEq./l., administer sodium supplements; if the patient complains of pain on urination, obtain a urine culture and consider the possibility of a urinary tract infection. These rules, as applied by the physician, are not absolutely deterministic. Although their selection and use do not involve deep reasoning, the doctor may withhold his or her full commitment from conclusions reached by even such categorical rules. The doctor thereby establishes the flexibility to modify his or her conclusions and rethink the problem if later difficulties arise.

A categorical decision typically depends on a relatively few facts; its appropriateness is easy to judge, and its result is unambiguous. A categorical decision is simple to make, and the rule that forms its basis is usually simple to describe (although its validity may be complicated to justify). Physicians most often work with categorical decisions, and, to whatever extent possible, computer experts should do the same.

Unfortunately, not every decision can be categorical. No simple rule exists for deciding whether to perform a bone marrow biopsy or when to discharge a patient from the cardiac intensive care unit. Those decisions must be made by carefully weighing all the evidence. Although we know that doctors do so, we do not understand just how they weigh the evidence that favors and that opposes various hypotheses or courses of action; this is an important unsolved problem for both AI and cognitive psychology (Newell and Simon, 1972; Tversky and Kahneman, 1974).

[2] Webster's defines *categorical* as "unqualified; unconditional; absolute; positive; direct; explicit; . . . "

A number of formal schemes for the weighing of evidence are used, and we shall concentrate on one of them, the *probabilistic,* to contrast with the categorical mode of reasoning[3] discussed above. We do not believe or suggest that formal probabilistic schemes are naturally used in decision making by physicians untrained in the use of such schemes. Indeed, there is convincing evidence that people are very poor at probabilistic reasoning (Tversky and Kahneman, 1974). Yet we believe that, with appropriate limitations as discussed below, probabilistic reasoning can be an appropriate component of a computerized medical decision-making system, especially for the difficult decisions for which categorical reasoning is inappropriate.[4]

In this paper we examine prototypical categorical and probabilistic reasoning systems, their limitations, and their successful applications, and then describe and analyze the reasoning mechanisms of some current AIM programs in terms of these schemes. We conclude with some comments and speculations on the requirements for reasoning mechanisms in future AIM programs.

9.2.1 Purely Categorical Decision Making—The Flow Chart

Categorical reasoning is exemplified by the simplest *flow chart* programs for guiding frequent decisions based on a well-accepted rationale. The flow chart is a finite state acceptor in which every nonterminal node asks a question whose possible answers are the labels of the arcs leaving that node. The machine has a unique initial state corresponding to initial contact with the user and a number of possible terminal states, each labeled by an outcome—a diagnosis, patient referral, selected therapy—relevant in its domain of application.[5] Every answer to every question is decisive; the formalism is simple and attractive.

[3] Other potentially appropriate schemes include the theory of *belief functions* (Shafer, 1976) and the application of *fuzzy set theory* (Gaines, 1976; Zadeh, 1965). All share the characteristic that arithmetic computations are performed to combine separate beliefs or implications to determine their joint effect. We are not convinced of the uniform superiority of any of these formalisms. Because we are most familiar with the probabilistic scheme, we have chosen to examine it in detail.

[4] Although our approach to the construction of expert medical systems has been, in general, to follow the way we think expert physicians reason, the known deficiencies in people's abilities to make correct probabilistic inferences suggest that this is one area in which the computer consultant could provide a truly new service to medicine. However, it is not universally accepted in medicine that probabilistic techniques are a valid way to make clinical decisions (Feinstein, 1977b).

[5] In some flow chart schemes, the structure of the acceptor is a tree. In that case, every terminal node can be reached only by a unique path. In other flow charts, the acceptor is augmented to retain information collected during questioning (e.g., in history-taking systems). Even in those systems, it is uncommon for a piece of information to be used to select a branch in the flow chart in any place except where it is determined. Thus that augmentation does not provide the program with any additional state information.

Perhaps the most successful use of categorical decision-making programs is in patient-referral triage.[6] Nurse-practitioners using standardized information-gathering and decision-making protocols can effectively handle routine orders for noninvasive laboratory tests and the scheduling of emergency or routine visits with a doctor. Such a system is now used in the walk-in clinic at the Beth Israel Hospital in Boston (Perlman et al., 1974), actually employing pen and printed forms rather than computer-generated displays and keyboard input.

Although every decision in a flow chart is categorical, the development of that flow chart may have been based on extensive probabilistic computations. Optimal test selection studies (Peters, 1976) and treat versus no-treat decision models (Pauker and Kassirer, 1975) are examples of probabilistic means of generating categorical decision models.

Whereas patient referral deals with a broad problem domain that may require only shallow knowledge, the problem of providing the physician with advice about the administration of digitalis requires a great deal of knowledge about a narrow medical domain. That domain is, in fact, sufficiently well understood at the clinical (although not the physiological) level that a reasonably straightforward program has been implemented (Silverman, 1975) that gathers relevant clinical parameters about the patient, projects digitalis absorption and excretion rates, adjusts for patient sensitivities, and monitors the patient's clinical condition for signs of therapeutic benefit or toxic effect. Although the numerical models used by the program are complex, its data-gathering strategy and its heuristic techniques for adjusting dosages are simple enough that most parts of the program can be explained to the user by simply translating the computer's routines into English (Swartout, 1977). This program relies largely on categorical reasoning.

Why are categorical decisions not sufficient for all of medicine? Because the world is too complex! Although many decisions may be made straightforwardly, many others are too difficult to be prescribed in any simple manner. When many factors may enter into a decision, when those factors may themselves be uncertain, when some factors may become unimportant depending on other factors, and when there is a significant cost associated with gathering information that may not actually be required for the decision, then the rigidity of the flow chart makes it an inappropriate decision-making instrument.[7]

[6] Triage is "the medical screening of patients to determine their priority for treatment; the separation of a large number of casualties, in military or civilian disaster medical care, into three groups: those who cannot be expected to survive even with treatment; those who will recover without treatment; and the priority group of those who need treatment in order to survive" (Stedman, 1961).

[7] Of course, one could, in principle, anticipate every complication and degree of uncertainty to every answer in the flow chart. If medical diagnosis is a finite process, then a gigantic flow chart could capture it all. This is, however, the equivalent of playing chess by having precomputed every possible game; it is probably equally untenable. It suffers similarly from losing all of the parsimony of the underlying model that the physician must have, from which the giant flow chart would be produced.

9.2.2 Purely Probabilistic Decision Making—Bayes' Rule and Decision Analysis

In a typical probabilistic decision problem,[8] we are to find the true state of the world, H_T, which is one of a fixed, finite set of exhaustive and mutually exclusive hypotheses, $H_1, H_2, \ldots, H_n$. We start with an initial estimate of the probability that each H_i is the true state. We then perform a series of tests on the world and use the results to revise the probability of each hypothesis. Formally, we have a probability distribution, P, that assigns to each H_i a prior probability, P_{H_i}. The available tests are $T_1, T_2, \ldots, T_m$, and for each test, T_i, we may obtain one of the results, $R_{i,1}, R_{i,2}, \ldots, R_{i,r_i}$.

Consider the case where we perform a series of the tests. We define the *test history* of the patient after the ith test to be the list of <test, result> pairs performed so far:

$$Q_i = (<T_{\mathrm{sel}(1)}, R_{\mathrm{sel}(1),k_{\mathrm{sel}(1)}}>, \ldots, <T_{\mathrm{sel}(i)}, R_{\mathrm{sel}(i),k_{\mathrm{sel}(i)}}>) \tag{1}$$

where sel is the test selection function.

If for every H_j and for every possible testing sequence, Q_i, we can assess how likely we would be to observe Q_i in the situation where H_j were known to be the true state, then we may apply Bayes' Rule to estimate, after any possible test history, the likelihood that H_j is H_T. In other words, if we know the conditional probability of any test history given any hypothesis, $P_{Q_i|H_j}$, for each j and Q_i, then we can apply Bayes' Rule to compute the posterior probability distribution over H:

$$P_{H_j|Q_i} = \frac{P_{Q_i|H_j} \cdot P_{H_j}}{\sum_{k=1}^{n} P_{Q_i|H_k} \cdot P_{H_k}} \tag{2}$$

A straightforward application of the above methodology would be to perform every test for every patient in a fixed order, obtaining Q_n, and then to use formula (2) to compute the posterior probabilities. Less naive applications of the methodology involve sequential diagnosis, in which the order of tests selected depends on previous results and in which diagnosis may terminate before all tests are performed. In sequential diagnosis, the next test to be performed may be selected by an expected information-maximizing function (Gorry et al., 1973) or a classical decision analysis that maximizes expected utility. The diagnostic process may terminate when the likelihood of the leading hypothesis exceeds some threshold[9] or when

[8] Here we follow Gorry (1967). This is the Bayesian approach to probabilistic decision problems.

[9] Sometimes, it is the ratio of the likelihood of the leading hypothesis to that of the next hypothesis that must exceed a threshold.

the expected cost of obtaining further information exceeds the expected cost of misdiagnosis due to missing those further data. Each of these techniques has been applied in diagnosis.

The failure of the pure probabilistic decision-making schemes lies in their voracious demand for data. Consider the size of the data base that would be needed for a direct implementation of the Bayesian methodology described above. In performing i of m possible tests, we can choose ${}_mP_i$ $(=m!/(m - i)!)$ possible test sequences. If every test has r possible results, then there will be $r^i\,{}_mP_i$ possible test histories after i tests. If we want to know the probability distribution over the H_i after each test (to help to select the next one), then we need to sum over test histories of every length and to multiply by the number of hypotheses, n, to get a total of

$$n \cdot \sum_{i=1}^{m} r^i \cdot {}_mP_i \tag{3}$$

conditional probabilities. For even a relatively small problem—e.g., $n = 10$ hypotheses, $m = 5$ binary tests ($r = 2$)—the analysis requires 63,300 conditional probabilities.[10]

Although the methodology described above is a complete view of medical diagnosis, it is certainly not an efficient one. To improve the scheme's efficiency, researchers typically make a series of assumptions about the problem domain that permit the use of a more parsimonious version of this decision method. First, it is usually assumed that two tests will yield the same results if we interchange the order in which they are performed.[11] That assumption reduces the number of conditional probabilities needed to

$$n \cdot \sum_{i=1}^{m} r^i \cdot {}_mC_i = n \cdot ((1 + r)^m - 1) \tag{4}$$

(2,420 in our example), which is still unwieldy.

A second assumption often made is that test results are conditionally independent—i.e., given that some hypothesis is the true state of the world, the probability of observing result $R_{i,k}$ for test T_i does not depend on what results have been obtained for any other test. This assumption allows all

[10] We are actually underestimating the amount of data required for such an analysis. In addition to the conditional probabilities, we also need other values to construct an optimal test-selection function. For example, we might use the costs of performing each test (possibly different after each different test history) and the costs and benefits of each possible treatment.

[11] Although this seems very reasonable, it is not strictly true. The effect of one test may be to interfere with a later one. For example, the upper GI series can interfere with interpretation of a subsequent intravenous pyelogram (IVP). The situation is even more complex since the effect of the former test on the latter often depends on the time that elapses between them. Even so, the assumption is so useful that it is worth making.

information from previous tests to be summarized in the revised probability distribution after the *i*th test, and the data requirements are reduced to approximately $m \cdot n \cdot r$ conditional probabilities (100 in our example), which is reasonable for some applications (Flehinger and Engle, 1975).

Unfortunately, three serious problems arise with the above scheme and its simplifications. The assumption of conditional independence is usually false, and the basic premises of the applicability of Bayes' Rule, that the set of hypotheses is exhaustive and mutually exclusive, are often violated. These may all lead to diagnostic conclusions that are wrong.

In a small study of the diagnosis of left-sided valvular heart disease, we have found that assuming conditional independence between observations of systolic and diastolic heart murmurs leads (not surprisingly) to erroneously reversed conclusions from those obtained by a proper analysis. To the extent that anatomical and physiological mechanisms tie together many of the observations that we can make of the patient's condition and to the extent that our probabilistic models are incapable of capturing those ties, simplifications in the computational model will lead to errors of diagnosis.

A similar error is introduced when conditional probabilities involving the negation of hypotheses are used. $P_{R|\sim H}$, being the probability of a test result R given that hypothesis H is *not* the true state of the world, cannot be assessed without knowing the actual probability distribution over the other hypotheses (unless, of course, there is only one other hypothesis). In fact, in our formalism,

$$P_{R|\sim H_i} = \sum_{j \neq i} P_{H_j} \cdot P_{R|H_j} /(1 - P_{H_i}) \tag{5}$$

which obviously depends on the probability distribution over the hypotheses. Even if we make the usual assumption of conditional independence, the practice of considering $P_{R|\sim H_i}$ to be a constant is unjustified and leads to further errors. Formalisms that employ a constant likelihood ratio implicitly commit this error, often without recognizing it (Duda et al., 1976; Flehinger and Engle, 1975). The likelihood ratio is defined as $P_{R|H_i} / P_{R|\sim H_i}$. Assuming conditional independence of the test results guarantees only that the numerator is constant, while, in general, the denominator will vary according to formula (5) as new results alter the probability distribution over the hypotheses. Using a constant likelihood ratio evaluates the current result in the context of the *a priori* probabilities, wrongly ignoring the impact of all of the evidence gathered up to that point.

A far more serious objection to the use of pure probabilistic decision making is that in most clinical situations the hypotheses under consideration are neither exhaustive nor mutually exclusive. If we perform a Bayesian calculation in the absence of exhaustiveness within the set of hypotheses, we will arrive at improperly normalized posterior probabilities. Their use in assessing the relative likelihoods of our possible hypotheses is appropriate, but we may not rest absolute prognostic judgments or compute expected values on the basis of such calculations.

The absence of mutual exclusivity is a more serious flaw in this methodology. Doctors find it useful to describe the clinical situation of a patient in terms of abstractions of disorders. When a patient is described as having acute poststreptococcal glomerulonephritis (AGN), for example, no one means that this patient exhibits every symptom of the disease as described in a textbook or that every component of the disease and its typical accompaniments is present. Having accepted such a description of the patient with AGN, diagnosis may then turn to consideration of whether such common (but not necessary) complications as acute renal failure and hypertension are present as well. Mapping this process into the view imposed by classical probabilistic methods requires the creation of independent hypotheses for every possible combination of diseases. That technique leads to a combinatorial explosion in the data collection requirements of the system and at the same time destroys the underlying view the practicing physician takes toward the patient.

Because of the distortions that the pure probabilistic scheme imposes on the problem and because of the enormous data requirements it implies, it tends to be used successfully only in small, well-constrained problem domains.

9.3 Reasoning in Current AIM Programs

Medical judgment, by the physician and by computer programs, must be based on both categorical and probabilistic reasoning. The focus of research in applying artificial intelligence techniques to medicine is to find appropriate ways to combine these forms of reasoning to create competent programs that exhibit medical expertise. In this section, we will outline in brief the central reasoning strategy of four major AIM programs and compare their methods to the two "pure cases" presented above.

9.3.1 The Present Illness Program

Perhaps the best way to explain the reasoning of our program is to describe the data that are available to it. The Present Illness Program (PIP) (Szolovits and Pauker, 1976) (also see Chapter 6) can deal with a large set of possible *findings* and a separate set of *hypotheses.* Findings are facts about the patient that are reported to the program by its user. Hypotheses represent the program's conjecture that the patient is suffering from a disease or manifesting a clinical or physiological state. Associated with hypotheses are sets of prototypical findings that can either support or refute the hypothesis.

Relation to Findings

TRIGGERS	<findings>
FINDINGS	<findings>

Logical Decision Criteria

IS-SUFFICIENT	<findings>
MUST-HAVE	<findings>
MUST-NOT-HAVE	<findings>

Complementary Relation to Other Hypotheses

CAUSED-BY	<hypotheses>
CAUSE-OF	<hypotheses>
COMPLICATED-BY	<hypotheses>
COMPLICATION-OF	<hypotheses>
ASSOCIATED-WITH	<hypotheses>

Competing Relation to Other Hypotheses

DIFFERENTIAL-DIAGNOSIS
(<condition 1> <hypotheses>) . . . (<condition *k*> <hypotheses>)

Numerical Likelihood Estimator

SCORE
((<condition 1,1><score 1,1>) . . . (<condition 1,n_1 > <score 1,n_1 >))
. . .
((<condition *m*,1> <score *m*,1>) . . . <condition *m*,n_m > <score *m*,n_m >))

FIGURE 9-1 Structure of a hypothesis frame in PIP.

Findings reported by the user are matched against these prototypical findings and, if a match occurs,[12] PIP's belief in the hypothesis is reevaluated. Figure 9-1 shows the structure of a hypothesis in PIP.

Presentation

Both TRIGGERS and FINDINGS are often associated with the hypothetical disorder. If a reported finding matches one of the triggers of a hypothesis, that hypothesis is immediately *activated.* If it matches a nontrigger

[12] The details of this matching process are not relevant to the questions addressed here and will not be discussed. The prototype finding can express either the presence or absence of a sign, symptom, laboratory test, or historical finding. For example, it is possible to use the *absence* of increased heart muscle mass (which takes months to develop) to argue in favor of acute rather than chronic hypertension. In general, many possible findings may match a prototype finding pattern. Thus, within each frame, only those aspects of a finding that are important to the hypothesis at hand need be mentioned, and any of the category of possible findings thus defined will match successfully.

finding, its relevance to that hypothesis is only noticed if the hypothesis is already under consideration. The logical decision criteria are used by the program to make categorical decisions about the likelihood of the patient's suffering from the currently considered hypothesis. IS-SUFFICIENT covers the case of pathognomonic findings, in which the presence of a single finding is in itself sufficient to confirm the presence of the hypothesized disorder; logical combinations (by NOT, AND, and OR) may also be used to specify more complex criteria. MUST-HAVE and MUST-NOT-HAVE specify necessary conditions, in the absence of which the hypothesis will not be accepted as confirmed.[13]

The *complementary* hypotheses identify other disorders that may be necessary in addition to the hypothesis under consideration to account for the condition of the patient.[14] The relationship may be known as *causal* if the physiology of the disorders is well understood, may be *complicational* if one disorder is a typical complication of the other, or may be *associational* if the two may be related by some known but incompletely understood association. Although all noncomplementary hypotheses are competitors, medical practice specifically identifies those that may often be confused—that is the role of the DIFFERENTIAL-DIAGNOSIS relationships in the frame.

The complementary and competing relations to other hypotheses are used in controlling the activation of hypotheses. In an anthropomorphic analogy, we think of an *active* hypothesis as corresponding to one about which the physician is consciously thinking. Active hypotheses offer the possible explanations for the patient's reported condition and are the basis from which the program reasons to select its next question. *Inactive* hypotheses are all those possible disorders that play no role in the program's current computations; they may be inactive either because no findings have ever suggested their possibility or because they have been considered and rejected by evaluation in light of the available evidence. *Semiactive* hypotheses bridge the gap between active and inactive ones and allow us to represent hypotheses that are not actively under consideration but that may be "in the back of the physician's mind." As mentioned above, if a trigger of any hypothesis is reported, that hypothesis is made active. When a hypothesis is activated, all of its closely related complementary hypotheses are semiactivated. Whereas nontrigger findings of inactive hypotheses do not lead to consideration of those hypotheses, any reported finding of a semiactive hypothesis causes it to be activated (i.e., each of its findings is treated as a trigger). This models the observation that physicians are more likely to pay attention to the minor symptoms of a disease related to the diagnosis that they are already considering than to the minor symptoms

[13] For logical completeness, we could have an IS-SUFFICIENT-NOT-TO-HAVE criterion, which would confirm a hypothesis in the absence of some finding, but this is just not useful.

[14] Note that we use the word *complement* in the sense of completion, not as implying negation or something missing. This is the sense of the word used in Pople (1975).

of an unrelated disorder. Each of the complementary hypotheses identifies another disorder that may be present along with the one under consideration and that is therefore to be semiactivated. The DIFFERENTIAL-DIAGNOSIS relation identifies a set of competing hypotheses that are to be semiactivated if the appropriate condition holds.

We need to assign to every hypothesis some estimate of its likelihood. In PIP, that estimate forms one basis for deciding whether the hypothesis ought to be *confirmed,* if the estimate is sufficiently high, or *inactivated,* if it is sufficiently low. Further, PIP bases its questioning strategy in part on the likelihood of its leading hypothesis. That likelihood is estimated by combining a function that measures the fit of the observed findings to the expectations of the hypothesis with a function that is the ratio of the number of findings that are accounted for by the hypothesis to the total number of reported findings. These two components of the likelihood estimate are called the *matching score* and the *binding score.*

PIP allows us to define clinical and physiological states (not only diseases) as hypotheses. Thus it is not necessary to list every symptom of a disease with that disease hypothesis; commonly co-occurring symptoms can be made symptoms of a clinical state hypothesis, and their relation to the disease derives from the causal relation of the disease to the clinical state. This is an appropriate structure that is consistent with medical practice. It does, however, raise a problem in computing the matching and binding scores for a hypothesis. If a finding is accounted for by a clinical state that is related to a disease, then the binding score of the disease hypothesis should reflect that relation, and its matching score should also reflect that the finding has improved the fit of the facts of the case to the hypothesis. To effect this behavior, PIP uses a *score propagation* scheme, described below. A similar argument can be made to extend score propagation to disease hypotheses as well: if a disease is made more likely by the observation of one of its symptoms, causally related diseases should also be seen as more likely.

The numerical likelihood estimator (see Figure 9-1) is used to compute the *local score* part of the matching score. The local score reflects the degree to which the facts found support the hypothesis directly. It consists of a series of clauses, each of which is evaluated as a LISP COND.[15] The local score of a hypothesis is the sum of the values of the clauses, normalized by the maximum possible total score. Thus it ranges from a maximum of 1 (complete agreement) downward to arbitrarily large negative numbers (complete disagreement).

[15] That is, for clause *i*, first <condition *i*,1> is evaluated, and if it is true, the value of clause *i* is <score *i*,1>. If that first condition is false, then each other condition in the clause is evaluated in turn, and the value of the clause is the score for the first true condition. Prototypical finding patterns in the condition that have not yet been asked about—thus, whose truth is not yet known—are treated as false, unless the pattern requests a negative or unknown finding. If none of the conditions is true, the value of the clause is zero.

PIP now computes the matching score by revising the local score to include the effects of propagated information deriving from related hypotheses. Consider the case when PIP is trying to compute the score for the hypothesis, H_i. First we identify all those other hypotheses, H_j, that are possibly complementary to H_i.[16] PIP then computes the MATCHING-SCORE by adding up the contributions of every scoring clause of H_i and each H_j and normalizing by the maximum possible total for this virtual scoring function. The effect here is to mechanically undo the organization imposed by the use of clinical and physiological states, since we could achieve a similar effect by merely listing with each hypothesis the exhaustive set of symptoms to which it might lead. Figure 9-2 shows, as an example, the PIP frame for acute glomerulonephritis.

Discussion

PIP uses both categorical and probabilistic[17] reasoning mechanisms. We shall identify the various forms of reasoning that it undertakes and whether they are accomplished by categorical or probabilistic means. When a finding is reported to PIP, whether as a fact volunteered by the user or in response to the program's questions, it tries to characterize fully the finding in terms of all the descriptors known to apply to that finding. For example, if edema is reported, PIP will try to establish its location, severity, temporal pattern, and whether or not it is symmetrical, painful, and erythematous. Rather specific rules capture some of the physician's common sense: if the question of past proteinuria is raised, PIP can conclude its absence if the patient passed a military physical examination at that time. These inferences are purely categorical.

The main control over PIP's diagnostic behavior resides in the list of active and semiactive hypotheses. Recall that only these hypotheses are "under consideration"—only they are evaluated or used to select the pro-

[16] H_j may be directly linked as a complementary relation to H_i, or it may be linked by a causal path going through some other hypotheses. In the latter case, we insist that the flow of causality along such a linking path be unidirectional, for we do not want, for example, two independent causes of some disease to reinforce each other's likelihood merely by being possible causes of the same disorder. We also compute a LINK-STRENGTH between the hypotheses, which is the product of each LINK-STRENGTH along the component links. Those component link strengths are identified in the data base and reflect the strength of association represented by the links.

[17] As should be clear from the above discussion, we do not think of the score computations as representing a true probability (either objective or subjective). We have sometimes tried to think of our scores as log-transformed probabilities, but the analogy is weak. Rather, we must think of them as an arbitrary numeric mechanism for combining information, somewhat analogous to the static evaluation of a board position in a chess-playing program. It is useful, however, to contrast the scoring computations with a correct probabilistic formulation, because that analogy suggests an explanation for various deficiencies of the scoring scheme (Szolovits, 1976).

```
TRIGGERS      (EDEMA with LOCATION = FACIAL or PERI-ORBITAL,
                  PAINFULNESS = not PAINFUL,
                  SYMMETRY = not ASYMMETRICAL,
                  ERYTHEMA = not ERYTHEMATOUS)
FINDINGS      (COMPLEMENT with RANGE = LOW), (MALAISE), (WEAKNESS),
              (ANOREXIA), (EDEMA with SEVERITY = not MASSIVE),
              (PATIENT with AGE = CHILD or YOUNG, SEX = MALE)
CAUSED-BY       (STREPTOCOCCAL-INFECTION in RECENT-PAST)
CAUSE-OF      SODIUM-RETENTION, ACUTE-HYPERTENSION, NEPHROTIC-SYNDROME,
              GLOMERULITIS
COMPLICATED-BY    ACUTE-RENAL-FAILURE
COMPLICATION-OF     CELLULITIS
DIFFERENTIAL-DIAGNOSIS

   (CHRONIC-HYPERTENSION implies CHRONIC-GLOMERULITIS)
   (EDEMA with RECURRENCE = not FIRST-TIME
       implies NEPHROTIC-SYNDROME, CHRONIC-GLOMERULONEPHRITIS,
       FOCAL-GLOMERULONEPHRITIS)
   (ABDOMINAL-PAIN implies HENOCH-SCHOENLEIN-PURPURA)
   (RASH with PURPURA = PURPURIC implies HENOCH-SCHOENLEIN-PURPURA)
   (RASH with (either LOCATION = MALAR or PHOTOSENSITIVITY = PHOTOSENSITIVE)
       implies SYSTEMIC-LUPUS)
   (JOINT-PAIN implies HENOCH-SCHOENLEIN-PURPURA, SYSTEMIC-LUPUS)
SCORE

(((PATIENT with AGE = CHILD or YOUNG) → 0.8)
 ((PATIENT with AGE = MIDDLE-AGED) → -0.5)
 ((PATIENT with AGE = OLD) → -1.0))
(((COMPLEMENT with RANGE = LOW) → 1.0)
 ((COMPLEMENT with RANGE = NORMAL or MODERATELY-ELEVATED) → -0.7)
 ((COMPLEMENT with RANGE = VERY-HIGH) → -1.0))
(((EDEMA with LOCATION = FACIAL or PERI-ORBITAL, SYMMETRY = not ASYMMETRICAL,
     DAILY-TEMPORAL-PATTERN=WORSE-IN-MORNING, PAINFULNESS = not PAINFUL,
     ERYTHEMA = not ERYTHEMATOUS) → 1.0)
 ((EDEMA with LOCATION = FACIAL or PERI-ORBITAL, SYMMETRY = not ASYMMETRICAL,
     PAINFULNESS = not PAINFUL, ERYTHEMA = not ERYTHEMATOUS) → .5)
 ((EDEMA with SEVERITY = not MASSIVE) → 0.1)
 ((EDEMA with SEVERITY = MASSIVE) → -0.1)
(((PATIENT with SEX = MALE) → 0.3)((PATIENT with SEX = FEMALE) → -0.3))
(((ANOREXIA) → 0.3) ((ANOREXIA absent) → -0.3))
(((WEAKNESS) → 0.3) ((WEAKNESS absent) → -0.3))
```

FIGURE 9-2 The PIP hypothesis frame for acute glomerulonephritis.

gram's further questions. The activation (but not the evaluation) of all hypotheses is purely categorical. A hypothesis can come up for consideration only if one of its prototype findings is matched by a reported finding, if a complementary hypothesis is activated, or if a competing hypothesis is active and a finding matches a condition among its differential diagnosis clauses.

Once a hypothesis is under consideration, both categorical and probabilistic mechanisms exist to decide its merit. In 18 of the 38 fully developed hypothesis frames in the current PIP, we find categorical IS-SUF-

FICIENT rules to establish the presence of the hypothesized disorder.[18] By contrast, all frames have a scoring function by which a pseudoprobabilistic threshold test may confirm hypotheses. Similarly, 9 of the frames have necessary conditions that may be used categorically to rule out a hypothesis, whereas all may be inactivated if their scores fall below another threshold. In our experience, the program performs best when presented with cases decided on categorical grounds. Too often, small variations in a borderline clinical case can push a score just above or just below a threshold and affect the program's conclusions significantly. Of course, in a textbook case, even the probabilistic mechanism will reach the right conclusion because the evidence all points in a consistent direction. Perhaps it should not disappoint us when the program flounders on tough, indeterminate cases where we have neither certain logical criteria nor a consensus from the evidence.

Once the reevaluation of all hypotheses affected by the last finding introduced is done, PIP selects an appropriate question to ask the user. That selection depends on the probabilistic evaluation of each active hypothesis. PIP identifies the highest-scoring active hypothesis, and if one of its expected findings has not yet been investigated, that finding is asked about. If all its expected findings have already been investigated, then PIP pursues expected findings of hypotheses complementary to the leading one.

To its user, PIP's reasoning is discernible from the conclusions it reaches and the focus of its questioning. PIP appears unnatural when its focus frequently shifts, as the probabilistic evaluator brings first one and then another competing hypothesis to the fore. This major deficiency relates to the lack of categorical reasoning. Such reasoning might impose a longer-term discipline or diagnostic style (Miller, 1975) on the diagnostic process.

In summary, PIP proposes categorically and disposes largely probabilistically.

9.3.2 INTERNIST—The Diagnostic System of Pople and Myers

INTERNIST (Oleson, 1977; Pople, 1975; Pople et al., 1975) is a computerized diagnostic program that emphasizes a very broad coverage of clinical diagnostic situations. The INTERNIST data base currently covers approximately 80% of the diagnoses of internal medicine (Pople, 1976), and thus is the largest of these AIM programs. Although INTERNIST is close to its goal of covering most of internal medicine, other problems lie down-

[18] Currently, PIP contains a total of 69 hypothesis frames, but 31 of them are so skeletal that they can never be confirmed. They are there to maintain the appropriate complementary relationships, and they anticipate a future extension of our data base.

Portal-vein-occlusion

Manifestation	*L*	*F*
Hepatic-vein-wedge-pressure-normal	0	4
Splenomegaly	1	4
Gastro-intestinal-hemorrhage	1	4
Varices-esophageal	2	4
Portal-vein-obstruction-by-radiography	5	3
Anemia	1	3
Appendicitis-history	1	2
Ascites	1	2

FIGURE 9-3 A diagnosis and its manifestations in INTERNIST. *L* indicates evoking strength; *F* indicates frequency.

stream for these researchers, including human-engineering issues centered on usability of the program's interface, possibly significant costs of running the program and maintaining the data base, introducing some model of disease evolution in time, and dealing with treatment, as diagnosis is hard to divorce from therapy in any practical sense.

Presentation

The INTERNIST data base associates with every possible diagnosis, D_i, a set of manifestations, $\{M_j\}$. A manifestation is a finding, symptom, sign, laboratory datum, or another diagnosis that may be associated with the diagnosis. For every M_j listed under D_i, two likelihoods are entered. $L_{D_i|M_j}$, the *evoking strength*, is the likelihood that if manifestation M_j is seen in a patient, its cause is D_i. It is assessed on a scale of 0 to 5, where 5 means that the manifestation is pathognomonic for the diagnosis and 0 means that it lends virtually no support. $F_{M_j|D_i}$, the *frequency*, is the likelihood that a patient with a confirmed diagnosis, D_i, would exhibit M_j.

Although INTERNIST's developers resist identifying these numbers as probabilities, $F_{M_j|D_i}$ is clearly analogous to the conditional probability $P_{M_j|D_i}$. The evoking strength is like a posterior probability, $P_{D_i|M_j}$, that includes a population-dependent prior, P_{D_i}, that is not explicit in the data base. If we were to take such a probabilistic interpretation, all the usual complaints about the failure of Bayesian assumptions would be appropriate. The INTERNIST scoring function that computes with these numbers is, however, in no sense probabilistic, and the rough granularity of the data is undoubtedly equally significant. It is reported that small random perturbations of the frequencies and evoking strengths in the data base do not significantly alter the program's behavior. A small example of a diagnosis, its associated manifestations, and the evoking strengths and frequencies connecting them are shown in Figure 9-3 (Pople, 1976).

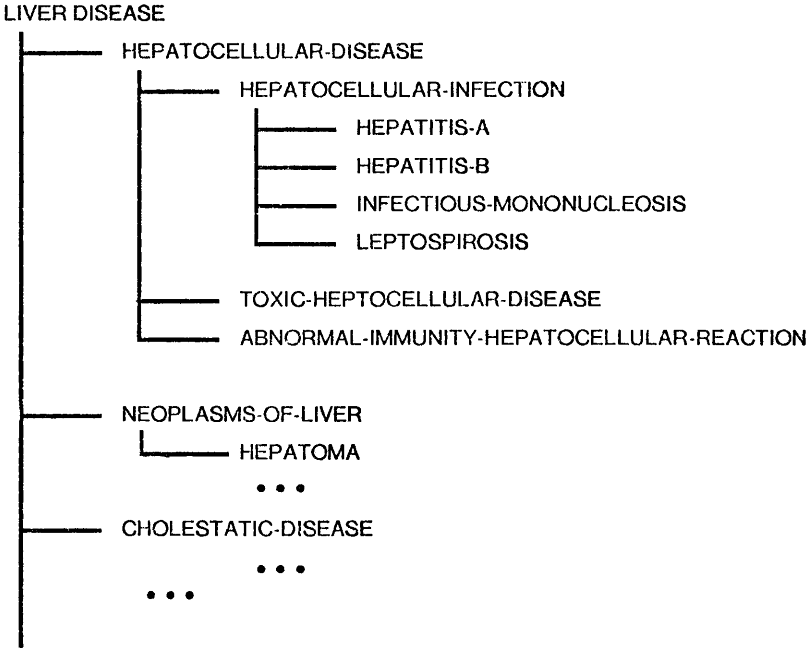

FIGURE 9-4 A small portion of INTERNIST's diagnosis hierarchy.

INTERNIST also classifies all its diagnoses into a disease hierarchy, a small part of which is shown in Figure 9-4 (Oleson, 1977). The use of hierarchy is an important mechanism for controlling the proliferation of active hypotheses during the diagnostic process because it allows a single general diagnosis to stand for all its possible specializations when no discriminating information is yet available to choose among them. This occurs, however, only when *all* specializations of the chosen general diagnosis have in common the same set of observed manifestations. Because INTERNIST wants to evaluate general as well as specific diagnoses, it *computes* for each general diagnosis a list of manifestations and their corresponding evoking strengths and frequencies. The manifestations for the general diagnosis are those common to each of its specializations, and the evoking strength and frequency of each are, respectively, the maximum evoking strength and minimum frequency of that manifestation among the specializations.

Borrowing the term from PIP, we will call a diagnosis *active* if at least one of its manifestations with a nonzero evoking strength has been observed, unless the diagnosis is a general one and must be replaced by its specializations (for example, because a manifestation occurring in one but not another of the more specific diagnoses has been reported). For each active hypothesis, a score is computed by summing the scaled evoking strengths of all its manifestations that have been observed, adding "bonus" points for confirmed causally consequent diagnoses, subtracting the sum of frequencies of those of its manifestations that are known to be absent, and also subtracting a weight of *importance* for each significant finding that is reported to be present but that is not explained by either the diagnosis or some other confirmed diagnosis. Thus evocative findings and confirmed consequences of a diagnosis count in its favor, while expected findings that are known to be absent and reported findings that are unexplained count against it.

Discussion

Drawing an analogy with PIP, INTERNIST's diagnoses are PIP's hypotheses, the manifestations are the findings and causally related hypotheses, and the evoking strengths are like the triggers—they and the frequencies play the role of the scoring function. INTERNIST's use of the importance measure for unexplained findings is superior to PIP's simple fractional binding score. Because the scoring function in PIP is explicit in each hypothesis frame, it requires more effort to create but provides a more general means of evaluating the significance of present and absent findings. Also, because PIP provides some logical criteria for confirming or denying a hypothesis, it provides a data base with the option of categorical hypothesis evaluation.

The lumping together of findings with causally consequent diagnoses, both as manifestations, leads INTERNIST to some difficulties. For it, any manifestation is either present, absent, or unobserved. This may be appropriate for findings, but when imposed on the evaluation of diagnoses, ignores the arguably real support of a strongly suspected though not confirmed causally consequent diagnosis for its antecedent. As Pople has pointed out, this effect may prevent INTERNIST from diagnosing a syndrome of connected hypotheses if no one of them is definitely provable even though the circumstantial evidence of their combined high likelihood is convincing to a physician. A similar deficiency arises because reported findings are explained only by confirmed diagnoses. Again, a strongly suspected but not confirmed complementary hypothesis will not be able to explain its significant findings, and so the correct diagnosis may have its score strongly penalized. As discussed above, PIP addresses these problems by dealing more explicitly with complementary disorders and accepting that a hypothesis accounts for a finding if one of its *active* complementary hypotheses accounts for it. We will argue below, however, that both of these solutions are weakened by not having a sufficiently explicit model of the hypothesis they are pursuing.

The most interesting part of INTERNIST is its focusing mechanism. After scoring all its active diagnoses, INTERNIST chooses to concentrate on the highest-ranking diagnosis. It partitions the others into two lists: the *competing* and the *complementary* diagnoses. A diagnosis is complementary to the chosen one if the two, together, account for more findings than either alone; otherwise the diagnosis is competing. The complementary list is then temporarily set aside, and a *questioning strategy* (one of RULE-OUT, NARROW, DISCRIMINATE, or PURSUE) is selected, depending on the number of high-scoring competitors and whether the information to be requested is low or high in cost. The complete scoring, partitioning, and strategy-selection processes are repeated after each new fact is reported. Confirmation is by numerical threshold. The partitioning heuristic is credited by Pople with having a very significant effect on the performance of the program, focusing its questioning on appropriate alternative diagnoses.

Because its intended coverage of disorders and findings is universal, INTERNIST relies on a uniform processing strategy and a simply structured data base. Much of its decision making falls under our probabilistic designation. The use of a hierarchic tree of diagnoses and of the rule for moving from a general to more specific diagnoses is categorical and captures an important part of a clinician's diagnostic behavior. The selection of questioning strategy is also categorical, although, interestingly, it depends on a probabilistic computation of the likelihood of each diagnosis.

9.3.3 CASNET—A Model of Causal Connectives

In a domain where normal and diseased states are well understood in physiological detail, it is sensible to build diagnostic models in which the basic hypotheses are much more detailed than the disease-level hypotheses of PIP and INTERNIST. Kulikowski, Weiss, and their colleagues have built such a system based on the causal modeling of the disease glaucoma. Their system is called CASNET, and it is in principle a general tool for building causal models with which well-known diseases may be diagnosed and treated (Weiss, 1974).

Presentation

CASNET defines a causal network of *dysfunctional states* and a set of *tests* that provide evidence about the likelihood of the existence of those states in the patient under consideration. States represent detailed dysfunctions of physiology, not complete diseases; thus the determination of disease is separated from the question of what, in detail, is going wrong in the patient.

The network consists of a set of nodes, some of which are designated as *starting states,* meaning that they are etiologically primary, and some as *final states,* meaning that they have no dysfunctional consequences. All causal relationships are represented by a link between two nodes, with a link strength that is interpreted as the frequency with which the first node causes the second. Starting states are given a prior frequency. No cycles are allowed in the network. Almost all nodes are representations of real physiological disorders. Although logical combinations of physiological states may be represented by a single node (for example, to express joint causation), this technique is discouraged. Further, "the resolution of states should be maintained only at a level consistent with the decision-making goal. A state network can be thought of as a streamlined model of disease that unifies several important concepts and guides us in our goal of diagnosis. It is not meant as a complete model of disease" (Weiss, 1974).

Two separate probabilistic measures are computed for every state in the network. A node's *status* is an estimate of its likelihood from the results of directly relevant tests. The status determines whether a node is *confirmed* or *disconfirmed.* A node's *weight* is an essentially independent estimate of its likelihood that derives from the strength of causal association between the node and its nearest confirmed and disconfirmed relatives. The weight computation ignores test results that affect the node's own status but is sensitive to results that establish the confirmation status of its causal relatives.

All tests are binary and are entered with an evaluation of the cost of each. If a positive or negative test result is reported, a set of links from the test to nodes of the network implies the presence or absence, respectively, of the corresponding nodes. Each link is labeled with a confidence measure for both positive and negative results, separately. A test may represent a simple observation of the patient, or it may be a logical combination of specific results of other tests. Only the results of simple tests are directly asked of the user of the program—the others are computed from the results of simple tests.

The status of each node is measured in the same units that are used to report the confidence measures of the implications of tests. Every time the result of a test is reported, the status of every node to which that test is linked is recomputed: if the result of the test has less confidence (i.e., is smaller in magnitude) than the status of the node, no change occurs. If the test result has greater confidence, the node's status is changed to that value. If they are equal, but of opposite sign, the node's status is set to zero, and a contradiction is noted for the user. One threshold, T, is defined such that if the status of a node is less than $-T$, the node is *denied,* and if the status exceeds $+T$, the node is *confirmed.*

The use of a maximum-confidence value for status and the ability to define a high-confidence test as the conjunction of two lower-confidence tests are in the fuzzy set tradition. This approach sidesteps the problem of the interpretation of mutually dependent test results, as they arise in a Bayesian formulation, by requiring the designer of the data base to define explicitly a new test for any combination of tests that jointly support the same node. Weiss argues that in his application domain this is perfectly appropriate, because when tests of varying confidence are available, only the results of the strongest should be counted (Weiss, 1974). One may question, however, whether this approach could be extended to wider medical areas, especially where many tests are available but only a consistent reading on most of them is enough to confirm a hypothesis.

Both for selecting a "most informative" test and for interpreting the pattern of status values among nodes of the network as a coherent disease hypothesis, CASNET defines an *acceptable path* in the network as a sequence of nodes that includes no denied nodes. A *forward weight* is computed for every node in the network, which represents the likelihood of that node when considering the degree to which its confirmed causal antecedents should cause it. Consider each admissible path that leads to node n_j and starts either at a starting node or at a closest confirmed node. CASNET computes the likelihood of causation along each such path by multiplying the link strengths along it (and the prior frequency for a starting state). The forward weight, w_j, of node n_j is defined to be the sum of the weights along each such path.

An *inverse weight,* representing the degree to which the presence of a node is implied by the presence of its causal consequents, is also computed.[19] CASNET then takes the maximum of the forward and inverse weights as the *total weight,* which is interpreted as a frequency measure of the degree to which the node is expected to be confirmed or disconfirmed from circumstantial causal evidence. Obviously, nodes with a high total weight and a status score near zero are excellent candidates for testing, since we might expect them to be confirmed. Conversely, nodes with low total weight are also candidates for testing, since we expect them to be denied. CASNET permits a number of different testing strategies to be used, based in part on the expected information implied by the weights and in part on the costs of the various tests.[20]

One should interpret the status of various nodes in the network as measures of the likelihood of subparts of a coherent disease. Based on the notion of the acceptable path, CASNET defines a number of different kinds of disease pathways, depending on which starting nodes are acceptable for such a path and on what criteria are used to terminate the path. It can compute those paths that are *most likely* to account for all the *confirmed* nodes in the network, all those that are *potential* explanations, and those that are not contradicted by a denied starting node (called *global*). Once the start of a disease path is selected, its termination criterion determines the type of path. An acceptable path that ends on a confirmed node is *confirmed.* An acceptable path ending on an undenied node is *possible.* A

[19] We cannot describe all of the computational mechanisms of CASNET here. An excellent presentation of the algorithms and a thorough justification for the particular choices made are in Weiss's thesis (1974).

[20] At present, the program is used with a fixed sequence of tests because an attempt is being made to gather a large, uniform data base about glaucoma patients. Thus the test selection function and this interesting weighting function are not in use (Weiss, 1976).

path that ends on a final state, even if it includes denied nodes, is *predictive*. Depending on the intent of the user, any combination of starting and termination criteria for a disease path may be selected. For example, the most likely starting criterion taken with the confirmed termination criterion will yield the "best estimate" diagnosis of the patient's current state. Selecting the global starting criterion and the predictive stopping criterion produces essentially all pathways through the network.

The most likely starting nodes are used to establish the probable causal mechanisms (the diseases) that account for the patient's difficulties. The ends of disease pathways give an estimate of the extent of the diseases. Together, these can be used to identify the primary disorder, to select a therapy for it, and to make prognostic judgments.

In a very clever manner, the determination of the effectiveness of therapy is handled by application of the same techniques used for diagnosis. A new causal network is constructed, in which the various therapies are the starting states and other nodes represent either complications of the treatments themselves or disorders not alleviated by the treatments. All of the above techniques are then available to assess whether any confirmed disorders are left after treatment and, if so, by what causal paths they could come about.

Discussion

At the level of testing, confirmation, and denial of nodes of the causal network, virtually all of CASNET's reasoning is probabilistic, based on the fuzzy set formalism for test interpretation and a probability interpretation for propagating causal frequency. The ability to define a hierarchy of tests (where higher tests summarize logical combinations of results of lower ones) and the simple confidence interpretation of node status provide a mechanism in which categorical rules for deciding node status are easily embedded.

The selection of a diagnosis and an associated therapeutic plan depends principally on the network designer's categorical understanding of the possible causal pathways through the net and on his or her definition of just which paths are subsumed by a given disease. In fact, if forward and inverse weights were not calculated, the elimination of any causal links that are not part of an identified disease path would result in no net effect on the operation of the program.

Weiss emphasizes that perfect accuracy in diagnosis by his program is not an unrealistic goal (presumably, without significant cost limitations on its testing strategy). This is to be contrasted to statistical classification schemes that would likely remain imperfect even with the addition of large quantities of new data. In CASNET, this confidence is justified because an error in the program's classification of a patient must ultimately indict some part of the causal model. In response, it may be necessary to add more tests to help distinguish the erroneous case, or the network may need to be disaggregated in selected places to give a more detailed model of some aspect of the disease. In the typical statistical approach, where the unit hypothesis is the disease, such local refinement is less feasible.

The glaucoma program works so well because its domain is narrow and the pathophysiology is well understood. Especially when compared with the domain of all of internal medicine (INTERNIST) or renal disease (PIP), the level of detail that is medically known and that it is practical to include in the glaucoma program is great. In fact, we speculate that the program could be recast as a categorical reasoning program. Given a fixed flow chart for test selection, we might consider in turn each of the roughly 50 starting states. From each, we might imagine a discrimination network that traces those diseases that start with that starting node. The discrimination net would branch, based on the crudely quantized confidence measure (status) of each successor node. That same measure could be used to determine the end of the disease path and thus the degree of progression of the disease and its possible therapies. Of course, such a technique may be too rigid to use in a changing environment or may not capture some capabilities of the original program (e.g., it could not compute all possible causes of some dysfunction). We hasten to mark this as pure speculation, but it suggests that perhaps more powerful categorical decision-making techniques could equally well solve the glaucoma problem, and thus that the probabilistic appearance of the CASNET solution is perhaps unnecessary.

A causal model is, nevertheless, attractive. We have seen physicians create (occasionally incorrectly) causal explanations for phenomena that they associate with diseases even though such a causal model played no important role in their interpretation of the phenomena. People seem happier if they understand why something happens than if they merely know that, under given circumstances, it does. Causal models for diagnosing dysfunction have been implemented for simple physical devices (Rieger, 1975) and proposed for medicine (Smith, 1978). In both these approaches, causality is taken as a categorical, not a probabilistic, connection. Reasoning about likelihood is often quantified only in the very fuzzy sense of IMPOSSIBLE, UNLIKELY, POSSIBLE, PROBABLE, and CERTAIN, and distinctive rules rather than a uniform numerical computation are used to combine data with different degrees of likelihood.

9.3.4 Production Rules—MYCIN and Inference Nets

The final AIM program whose reasoning component we shall describe is MYCIN, which is being developed to advise physicians and medical students in the appropriate treatment of infections (Shortliffe and Buchanan, 1975) (see also Chapter 5).

IF: 1) The stain of the organism is gram positive and
2) The morphology of the organism is coccus and
3) The growth confirmation of the organism is chains
THEN: There is suggestive evidence (.7) that the identity of
the organism is streptococcus

FIGURE 9-5 A typical MYCIN rule.

Presentation

MYCIN's knowledge is expressed principally in a number of independently stated rules of deduction, a typical example of which is shown in Figure 9-5. MYCIN's highest-level goal is to determine if the patient is suffering from a significant infection that should be treated, and if he or she is, to select the appropriate therapy. It uses a backward-chaining deduction scheme in which all applicable rules are tried: if a condition in the IF *(antecedent)* part of a rule is decidable from the data base, that is done; if the condition can be asserted by the THEN *(consequent)* part of some other rules, they are applied; otherwise, MYCIN asks the user. Thus the rule of Figure 9-5 might be applied in the following chain of reasoning:

1. To decide if the patient needs to be treated, we must decide if he or she has a significant infection.
2. We must know the likely identity of the infecting organism to decide if the infection is significant.
3. The rule of Figure 9-5 can determine the identity of the organism.

Because conditions in the rules may include logical disjunctions as well as conjunctions, the deduction forms an AND/OR tree.

When the methodology of MYCIN was applied to the simple domain of bicycle troubleshooting, a small set of categorical rules of this type was sufficient to give the program some interesting behavior. The complication in MYCIN arises from the uncertainty with which a medical rule implies its consequences, the applicability of several uncertain rules to suggest the same consequence, and the need to apply rules even when their antecedents are to some degree uncertain.

MYCIN associates a *certainty factor* (CF) with each rule, which is a number between 0 and 1, representing the added *degree of belief* that the rule implies for its consequent. With each fact in the data base is a *measure of belief* (MB) and a *measure of disbelief* (MD), both numbers between 0 and 1 that summarize all the positive and negative evidence that has been imputed for this datum by the application of rules that conclude about the datum. The measures of belief and disbelief are maintained separately for each item, and the certainty factor of the fact is their difference. Thus the CF of a fact is a number between -1 and 1.

Arguing that the rule "A implies B with probability X" should not be inverted in the traditional probabilistic sense to entail "A implies not B with probability $(1-X)$," Shortliffe defines a *confirmation* formalism for computing the certainty of facts (Shortliffe and Buchanan, 1975). In its simplest form, it says the following: assume that we are told (perhaps by some rule $S1$) the fact H with certainty $\text{MB}_{H|S1}$. Later, we discover that another source of information, $S2$, tells us H again, this time with certainty $\text{MB}_{H|S2}$. Instead of using a maximum, as CASNET would, we would like to feel more confident in H after having received two reports in its favor than after having received either one by itself. MYCIN's scheme means that every new report of the truth of H reduces the difference between 1 and H's measure of belief by the fraction that is the certainty of the new report. For example, if $\text{MB}_{H|S1} = 0.4$ and $\text{MB}_{H|S2} = 0.6$, then the combined result is $\text{MB}_{H|S1,S2} = 0.76$. This process is defined separately for positive and negative reports, and we have

$$\begin{aligned} \text{MB}_{H|S1,S2} &= 0 \qquad \text{if } \text{MD}_{H|S1,S2} = 1 \\ &= \text{MB}_{H|S1} + \text{MB}_{H|S2}\,(1 - \text{MB}_{H|S1}) \qquad \text{otherwise} \end{aligned} \tag{6}$$

and

$$\begin{aligned} \text{MD}_{H|S1,S2} &= 0 \qquad \text{if } \text{MB}_{H|S1,S2} = 1 \\ &= \text{MD}_{H|S1} + \text{MD}_{H|S2}\,(1 - \text{MD}_{H|S1}) \qquad \text{otherwise} \end{aligned} \tag{7}$$

where $S1$ and $S2$ are the two reports. The measures of belief and disbelief combine to give a certainty factor for each fact:

$$\text{CF}_H = \text{MB}_H - \text{MD}_H$$

This, then, defines MYCIN's method of summarizing the certainty of a hypothesis when the application of several rules has contributed evidence for it.

To compute the measure of belief (or disbelief) contributed by a particular rule, MYCIN multiplies the CF of the rule by the MB (or MD) of the rule's antecedent. A fuzzy set strategy of maximizing for OR and minimizing for AND is adopted to compute the belief measures of the antecedent from the belief measures of its components. This approach is presented and justified in Shortliffe and Buchanan (1975) and Shortliffe (1976). An alternative formulation of separate measures of belief and disbelief is to be found in Shafer (1976).

Discussion

In MYCIN, the question of just what connections exist among different facts in the data base is not explicitly addressed. In addition to the rules that we have mentioned above, MYCIN also includes a *context hierarchy,* which plays a smaller but still important role in the program's operation. For example, the facts that "there are cultures associated with infections" and that "cultured organisms are associated with cultures" are embedded in no rules, but rather in this additional mechanism.[21] Turning MYCIN inside out, that context mechanism could be viewed as the principal organizational facility of the diagnosis program. In such a view, the underlying reasoning activity is filling in a *frame* for the patient by directly asking for some information (e.g., age and sex) and by instantiating and recursively filling in other frames (e.g., cultures and operations). The productions and their associated certainty factors are then seen as a set of procedurally attached heuristics to help fill in those frames. We conjecture that this methodology, which underlies the operation of the GUS program (Bobrow et al., 1977), would provide a reasonable alternative way of implementing the MYCIN system.

MYCIN's categorical knowledge is encoded in three ways. First, the presence of each rule implicitly establishes a categorical, inferential connection between those facts in its consequent and those it uses in its antecedent. The MYCIN control structure, which is a nearly purely categorical backward-chaining deduction scheme, is based on these relationships. Second, the context tree explicitly,defines what objects may exist in MYCIN's universe of discourse and how they may relate. Such categorical information would underlie a GUS-like implementation of MYCIN. Third, many other relationships, which record such data as how to ask a question and what answers are acceptable, are also categorical in nature. MYCIN's probabilistic reasoning resides in its use of the measures of belief and disbelief about each fact and the certainty factors associated with each rule. Although this probabilistic method has important consequences for the assessment of the relative likelihoods of the various infecting organisms under consideration, it appears that it affects the program's questioning behavior only slightly. Except in the case where a line of reasoning is pursued because of the joint effect of several very weak independent inferences, which we suspect is rare, the particular numbers used make little difference except in the final diagnosis (and thus therapy). We note that the context tree that is built for each patient depends for its structure mainly on information that is always asked of the patient, such as what cultures have been taken, what operative procedures have been performed,

[21] Note that, because of interposed levels of complexity such as the existence of cultures, the example "traceback" we presented above of how MYCIN would decide to apply the rule of Figure 9-5 is overly simplistic.

and what drugs are being used in treatment. Even dramatic changes in the probabilistic component of MYCIN's reasoning strategy would not alter this behavior.

MYCIN has also inspired the creation of a more uniform inference scheme, in which every potential fact in the data base is viewed as a node in a large inferential network.[22] In such a network, the reasoning rules form the connections among the fact nodes, and we think of propagating some measures of likelihood among the nodes so that the impact of directly observable facts may be reflected on the diagnostic consequences of ultimate interest. This is the approach taken by Duda, Hart, and Nilsson (1976) in their *inference net* formalism. The propagation scheme used there is Bayesian in its heritage, but suffers from the typical distortions (see above) that the Bayesian methodology can introduce.

Of course, it is natural to compare the inference net to the causal net. The difference is primarily in the semantic interpretation of what a node and a link represent. In CASNET, the node is a dysfunctional state, and the link represents causality in the application domain. In the inference net, nodes are essentially arbitrary facts about the world, and rules are arbitrary implications among those facts. Much of Weiss's reasoning in justifying the particular propagation algorithms he has chosen rests on his specific interpretation of the network. Because the semantics of the inference net are less clearly (or constantly) defined, we must be more skeptical when evaluating the acceptability of the approximations introduced by the propagation formulas.

9.4 Another Look at the Problems of Diagnosis

Compared to the expert physician, our best AIM programs still have many deficiencies. We catalog a few of the more significant ones:

1. Programs that deal with relatively broad domains, such as INTERNIST and PIP, have inadequate criteria for deciding when a diagnosis is *complete.* There is no sense of when the major diagnostic problems have been resolved and only the "loose ends" remain: the programs continue exploring less and less sensible additional hypotheses until the user tires of the consultation. For example, PIP only stops if no active hypotheses remain or if every finding of every active hypothesis has been explored already.

[22] Uniformity is not necessarily an advantage for a reasoning scheme. For example, the particular structures used by MYCIN are cleverly exploited by Davis in building an interesting knowledge-acquisition module (Davis, 1976). In a uniform system of representation, it would be more difficult for his programs to decide just where new knowledge is to be added.

2. Because the initial strategy of the programs is to use every significant new finding as a clue to raise the possibility of associated disorders and because this strategy remains throughout the programs' operation, new hypotheses are continually being activated. Thus, when the program asks about an expected finding for one of its leading hypotheses and the finding is present, that finding often suggests new hypotheses as well, even though it is perfectly consistent with the diagnosis being pursued. Obviously, some such sensitivity is necessary or the program would remain committed to its first hypothesis, but we now feel that it would be preferable if new hypotheses were triggered only by evidence that contradicts a current belief.
3. Part of the routine developed by clinicians is an appropriate order for acquiring information systematically. Computer diagnosticians tend to enforce such an order either too strictly (e.g., the flow charts and MYCIN, which cannot accept out-of-sequence information in any useful way) or not at all (e.g., INTERNIST or PIP, where a global computation after the report of each fact may, in the worst case, change the program's focus to an entirely new topic for each question).
4. The programs rely on a global likelihood assessment scheme, but they use a semantics that is too weak for the states over which they try to compute approximate probabilities. For example, none of the programs can dynamically distinguish among the aggregate hypotheses

   a. A and B, both together, when in fact A has caused B,

   b. A and B co-occurring but apparently unrelated, and

   c. A or B but not both.

   Yet there are therapeutic and strategic decisions that hinge on just such distinctions. For example, it may be sufficient to treat only for A in the first case, but not in the second; trying to discriminate between A and B makes sense in the third case, but not in the others. PIP and INTERNIST might eliminate some of these hypotheses by noting those causal or associational links that are disallowed by the data base, but in no sense are these hypotheses generally distinguishable. MYCIN might include some rules that could, for example, reduce the possibility of hypothesis c, but it also lacks any mechanism to take up the problems of dependence. Although CASNET does allow the proper handling of this problem, it must do so by the creation of joint states, which is its weakest semantic ability.

9.4.1 Possible Improvements

The practice of clinical medicine offers some clues to the proper solution of some of these difficulties. Questions of the appropriate termination of the diagnostic process and control over the proliferation of hypotheses may be resolved by considering two factors. First, the diagnosis needs to be only as precise as is required by the next decision to be taken by the doctor. Thus, if all the remaining possible diagnoses are irrelevant or equivalent in their implications for therapy or test selection, then nothing is lost by postponing their consideration. Bayesian programs that explicitly compare the cost of new information to its expected benefit will achieve this saving (Gorry et al., 1973), but none of the programs discussed here includes such a computation.

Second, the simple passage of time, "creative indecision," often provides the best diagnostic clues because the evolution of the disorder in time adds a whole new dimension to the other available information. Whereas MYCIN, CASNET, and the Digitalis Therapy Advisor all use changes over time as diagnostic clues, none of the programs exploits the possibility of deferring its own decisions with a deliberate eye to waiting for disease evolution. Such a strategy is also applicable on the much shorter time scale of the diagnostic session. In taking the present illness, for example, the doctor knows that a physical examination and a review of symptoms will soon provide additional information. Therefore, consideration of unlikely leads and small discrepancies can be deferred, leaving a coherent structure of problems to work with at the moment.

The ability to lay aside information that does not fit well with the current hypotheses is also a good mechanism for limiting the rapid shifts of focus caused by consideration of newly raised but unrelated hypotheses. In addition, however, the programs must have a sense of the orderly process by which information is normally gathered. The attempts in PIP to characterize a finding fully before proceeding and the attempts in INTERNIST and CASNET to ask summarizing questions (not described here) before launching on a series of similar, detailed questions are attempts to reflect such an order. We might, as Miller suggests (1975), go much further. We could, for example, incorporate a strategy that says, "When investigating a suspected chronic disease, insist on a chronological description of all the patient's relevant history." If such a strategy were followed, the program would not quickly jump at a "red herring" uncovered during the acquisition of those historical data. For example, consider a patient with a long history of sickle cell anemia who now complains of acute joint pain. Although that complaint would ordinarily raise the issue of rheumatoid arthritis, in this case we (and the program) should realize that the joint pain is a reasonable consequence of an already known disease process and should not evoke an immediate attempt to create elaborate additional explanations. Maintaining a richer semantic structure of just what the current hypothesis is and allowing that structure to control the program's focus of attention should also stabilize the program's behavior.

Another possible mechanism for controlling the logic of diagnosis is suggested by the following example. Consider the earliest stages in the diagnosis of chest pain, a symptom of potentially grave consequence. With a disaggregated structure of relationships between findings and hypotheses, chest pain might suggest angina pectoris, aortic stenosis, pneumonia, tuberculosis, pericarditis, costo-chondritis, depression, hiatus hernia, pancreatitis, esophagitis, gastric ulcer, fractured rib, pulmonary embolism, etc.—a long list of significantly different low-level hypotheses. Once those are all active, we must evaluate and compare all of them to choose a best hypothesis. On the other hand, we can say that, initially, we will only use the finding of chest pain to choose a somewhat specific diagnostic area for our further focus; specifically, we would like to choose one of these generic hypotheses: the pain is due to cardiac, pulmonary, gastrointestinal, psychogenic, or muscular-skeletal causes. We ask only the age and sex of the patient and three of the most important descriptors of the chest pain, its character, provocation, and duration. Obtaining a rank order for the five categories from each descriptor and combining them by a very simple arithmetic formula, we get a reasonably robust estimate of what is the best diagnostic area to pursue.

No simple scheme like the one suggested here is, of course, a panacea. However, we have been surprised at how effective rather crude heuristic techniques can be when they are tailored to a specific problem. To illustrate the necessity of that tailoring, it should be pointed out that the same technique appears *not* to be effective at the next level of diagnosis, for example, in sorting out the various possible cardiac causes of chest pain.

In summary, our analysis of the reasoning mechanisms of current AI programs leads us to these conclusions:

1. If possible, a carefully chosen categorical reasoning mechanism that is based on some simple model of the problem domain should be used for decision making. Many such mechanisms may interact in a large diagnostic system, with each being limited to its small subdomain. Many of the intuitively appealing observations made above can probably be implemented by the use of such techniques.
2. When complex problems need to be addressed—which treatment should be selected, how much of the drug should be given, etc.—then causal or probabilistic models are necessary. The essential key to their correct use is that they must be applied in a limited problem domain where their assumptions can be accepted with confidence. Thus it is the role of categorical methods to discover what the central problem is and to limit it as strongly as possible; only then are probabilistic techniques appropriate for its solution.

9.4.2 Postscript

As we interact with our medical colleagues at work, we are sometimes amazed by two observations:

1. They are often extremely reluctant to engage in any numerical computation involving the likelihood of a diagnosis or the prognosis for a

treatment. Even when official blessing is bestowed upon Bayesian techniques, we have seen both experienced and novice physicians acknowledge and then ignore them. Doctors certainly have a strong impression of their confidence in the diagnosis or treatment, but that impression must arise more from recognizing a typical situation or comparing the present case to their past experiences rather than from any formal computation of likelihoods.

2. An experienced physician can be pushed, in his or her domain of expertise, to give arbitrarily many complex potential explanations for a patient's condition. Especially in the teaching hospital environment with which we are most familiar, this serves the useful pedagogical purpose of discouraging pat answers from students. Because so many diagnostic possibilities appear to be available for the expert to consider, we suspect that the rapid generation and equally rapid modification or elimination of many explicit hypotheses play a significant role in his or her reasoning.

These observations reinforce our beliefs that somewhat more careful approaches to diagnosis are needed, ones that apply the most successful available techniques to each component of the diagnostic process. Although probabilistic techniques will be best in some well-defined domains, they should not be applied arbitrarily to making other decisions where the development of precise categorical models could lead to significantly better performance. The development and aggregation of a number of different approaches, both categorical and probabilistic, into a coherent program that is well suited to its application area remains a fascinating and difficult challenge.

When thinking about the effectiveness of a computerized medical consultant, it is essential to recognize the difference between impressive expertlike and truly expert behavior. A vehement critic of early work in artificial intelligence accused the practitioners of this "black art" of trying to reach the moon by climbing the tallest tree at their disposal (Dreyfus, 1972). We must be somewhat concerned that the initial successes of the current programs should not turn out to be merely the improved view from a lofty branch.

ACKNOWLEDGMENTS

This research was supported by the Department of Health, Education and Welfare (Public Health Service) under grant no. 1 R01 MB 00107-03 and by Dr. Pauker's Research Career Development Award (1 K04 GM 00349-01) from the General Medical Sciences Institute, National Institutes of Health.