Toward Normative Expert Systems: The Pathfinder Project[*]

David E. Heckerman
Medical Computer Science Group
Section on Medical Informatics
Stanford University School of Medicine
Stanford, California 94305

Eric J. Horvitz
Medical Computer Science Group
Section on Medical Informatics
Stanford University School of Medicine
Stanford, California 94305

Bharat N. Nathwani
Department of Pathology
School of Medicine
University of Southern California
Los Angeles, California 90033

[*] Core research on the Pathfinder project has been supported by the National Library of Medicine under Grant RO1LM04529. Components of the research have been supported by the National Aeronautics and Space Administration under Grant NCC-220-51, by the National Science Foundation under Grant IRI-8703710, and by the Josiah Macy, Jr., Foundation. Computational support has been provided by the SUMEX-AIM resource under National Institutes of Health Grant LM05208.

Abstract

We describe key issues and achievements on the Pathfinder expert-system project. Pathfinder research has extended over 5 years at the University of Southern California and at Stanford University. Our investigation of automated reasoning has spanned several paradigms for representing and reasoning with expert medical knowledge. The Pathfinder team has concentrated on the construction of ***normative expert systems***, which are expert systems based on the principles of decision theory. Within the normative paradigm, we have developed useful techniques for acquiring, representing, manipulating, and explaining complex medical knowledge. A large component of such knowledge is made available only through careful assessment of subjective probabilistic relationships. Although the bulk of Pathfinder research has been carried out within the domain of lymph-node pathology, the insights and techniques have relevance to a wide variety of complex biomedical domains. We introduce the project, describe the significance of normative reasoning in medicine, and review theoretical and empirical Pathfinder developments. We describe the techniques for constructing and managing a large probabilistic knowledge base, the use of a normative hypothetico-deductive architecture, the application of alternative information-acquisition strategies, the explanation of probabilistic inference, and the heuristic and normative evaluation of the Pathfinder reasoning system.

1 Introduction

For over 6 years, we have worked to construct an expert system, called *Pathfinder*, to assist general pathologists with diagnosis in the specialty area of hematopathology [1,2]. A major component of our research has addressed the acquisition and representation of expert pathology knowledge, as well as the manipulation and explanation of that knowledge. We have found that it is crucial to represent and reason with ***uncertain*** knowledge, and that the capture and manipulation of uncertain knowledge is fundamentally different from corresponding tasks for knowledge held with certainty. Consequently, we have had to explore carefully a variety of techniques proposed for reasoning under uncertainty, and to develop new approaches to grapple with complex relationships among evidence and hypotheses.

Methodologies for reasoning under uncertainty have played a central role in the history of medical informatics. Medical reasoning is typically dominated by great uncertainties: the complexity of pathophysiology typically overwhelms the abilities of people to understand all relevant details about the predicaments of their patients. Over the last 3 decades, medical-informatics investigators explored several computer-based reasoning methodologies for representing and manipulating uncertain biomedical knowledge for use in automated medical reasoners. Researchers have hoped that expert reasoning systems, based on one or more of these uncertain-reasoning methodologies, will one day serve as common sources of expert advice when human experts are not available.

Beginning with early failures to represent complex uncertain knowledge successfully with production rules and with nonprobabilistic scoring schemes, the Pathfinder team has concentrated on decision-theoretic methods for diagnosis. Decision theory, which includes probability theory and the maximum expected utility principle, provides a set of desirable rules that people believe they should follow or wish they could follow when confronted with a confusing, high-stakes decision. That is, the principles are viewed traditionally as *normative*. Psychologists have found, however, that people, including experts, deviate from these principles in stereotypic fashion [3,4]. Thus, we believe that the development of *normative expert systems*—expert systems based on the principles of decision theory—will lead to improvements in the delivery of expert knowledge.

A chief problem with the development of normative expert systems is the complexity of traditional representations of knowledge in a decision-theoretic framework. This complexity has dampened interest in applying decision theory in computer-based reasoning systems. Pathfinder research has developed a set of techniques that can make normative expert systems practical to develop and to use. In addition to the fundamental work involving the efficient acquisition and refinement of large normative knowledge bases and the development of tractable reasoning strategies, we have explored the use of human-oriented classification hierarchies for tailoring the system's behavior to different users, and for the explanation of the system's recommendations. We have also developed and tested techniques for the evaluation of the accuracy of diagnosis. Our work to solve problems of representing, manipulating, and explaining uncertain knowledge is relevant to problems in artificial intelligence (AI) that extend well beyond medical informatics.

Our presentation is organized in three main parts. In Part I, we provide background information that is necessary to understand the new research presented in this paper. In Sections 2 and 3, we introduce the domain of lymph-node pathology and we discuss problems associated with diagnosis in this domain. In Section 4, we introduce the Pathfinder expert system. In this section, we provide a history of the development of Pathfinder, and describe its basic operation through the use of a simple example. In Section 5, we introduce the hypothetico-deductive approach to reasoning, and discuss its application to expert systems research. In Sections 6 through 8, we describe our use of both decision-theoretic and non-decision-theoretic to represent uncertain medical knowledge, and discuss the events that led to our eventual use of decision theory in Pathfinder.

In Part II, we present our new research inspired by the Pathfinder project. In Section 9, we describe a representation that greatly facilitates the capture and representation of uncertain knowledge within the probabilistic framework. In Sections 10 and 11, we discuss our use of decision theory to identify pieces of evidence that are cost effective for narrowing a differential diagnosis, and our methods for explaining such recommendations.

Finally, in Part III, we present a detailed evaluation of the diagnostic accuracy of two versions of Pathfinder. In one version of Pathfinder, we assume that all pieces of evidence are conditionally independent. The assumption of global conditional independence has been made by almost all medical-informatics researchers in the past. In the other version, we avoid this assumption. Instead, we represent accurately the probabilistic dependencies among the pieces of evidence represented by the system. Our evaluation demonstrates dramatically that it is important to represent such dependencies in the lymph-node domain.

Part I
Background

2 Diagnosis in Surgical Pathology

As portrayed in Figure 1, surgical pathologists perform diagnosis primarily through the identification of a set of relevant microscopic features that appears in a section of tissue. A pathologist applies his knowledge about the features on a slide to determine the likelihood of alternative diseases. The pathologist's diagnosis is relayed to an oncologist, who directs a patient's cancer therapy, based on this recommendation. The quality of a patient's therapy therefore is greatly dependent on the accuracy of the pathologist's diagnosis. Unfortunately, surgical pathology diagnoses can be extremely difficult to make. For example, as is suggested in Figure 1, within many areas of pathology, diseases with widely different therapies and prognostic courses may resemble one another closely.

The task of a pathologist includes (1) identifying features and quantifying them; (2) constructing a ***differential diagnoses***, or set of diseases consistent with the observations; (3) deciding what additional features to evaluate; (4) deciding what costly tests or stains to employ; and (5) rendering a diagnoses. Thus, we must consider these components of diagnosis in building a useful expert system for pathologists.

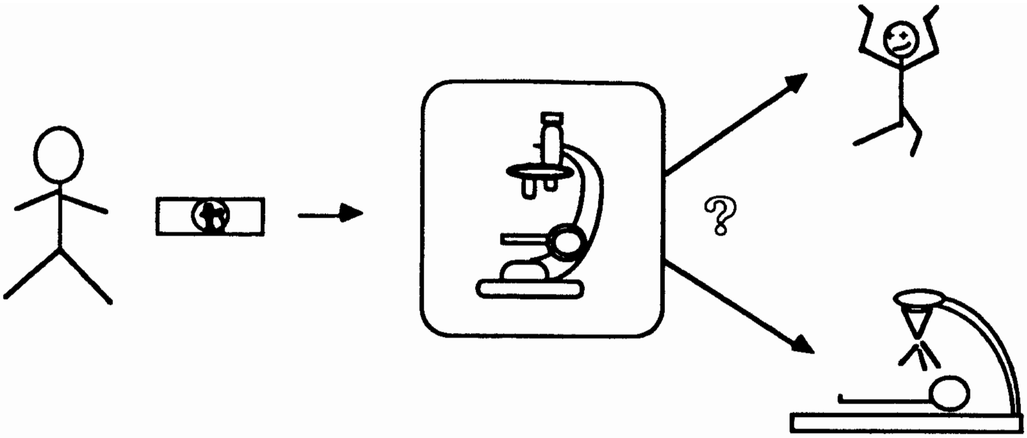


Figure 1: A central component of surgical pathology diagnosis involves the identification of microscopic features visible on a tissue-biopsy section. The pathologist's diagnosis is used by oncologists in the patient's therapy. Within lymph-node pathology, diseases with widely different outcomes may resemble one another closely. Thus, accurate diagnosis is required for appropriate therapy.

There are approximately 40 subspecialty areas in surgical pathology, representing expertise in the diagnosis of pathology in tissues from the different human organ systems. Some pathologists specialize in the diagnosis of disease in one or more subspecialties. These expert pathologists tend to focus their attention on difficult diagnostic problems within the subspecialties.

Unlike a specialist, a general pathologist typically performs diagnosis on sections from a wide range of organ systems. In each area of pathology, a pathologist may identify pieces of evidence from among hundreds of distinct features. These features include colors and patterns discernable at low magnifications, and assemblies of cells, cell structure, and organelle structures seen at higher magnifications. If a general pathologist has difficulty with diagnosis, he frequently refers the case to a subspecialist. This referral process usually incurs both a delay in diagnosis and an extra cost. Sometimes, the delay in diagnosis is unacceptable, and the pathologist cannot refer the case to a subspecialist. For example, surgeons often rely on pathologists for the timely diagnosis of disease in frozen tissue, taken from patients under anesthesia [5,6].

3 Problems in Hematopathology

The Pathfinder team has explored the feasibility of providing computer-based expertise for diagnosis, based on the status of features observed in a section of lymph-node tissue. The pathology of the lymph nods poses difficult diagnostic problems. There are approximately 30 different types of primary and secondary malignant hematopoietic diseases of the lymph node. These malignant diseases have to be distinguished from approximately 30 different benign diseases, many of which closely simulate malignant lymphomas. For epidemiologic and therapeutic reasons, it is important that the benign diseases be differentiated from malignant conditions and that a precise classification of malignant lymphoma be established, so that the patient receives the most appropriate form of radiation therapy or chemotherapy [7,8,9].

The diagnosis of lymph node diseases, based on morphologic presentation, is one of the most difficult and problematic tasks of surgical pathology [10,11,12,13]. Several cooperative oncology studies have documented that, although experts show agreement with one another, the diagnosis rendered by a community-hospital pathologist may have to be changed by expert hematopathologists in as many as 50 percent of the cases [14].

In response to the diagnostic problems in lymph-node pathology, the National Cancer Institute (NCI) created the Lymphoma Task Force over 2 decades ago. The task force is now called the Repository Center and the Pathology Panel for Lymphoma Clinical Studies. The main function of this panel of expert pathologists is to confirm the diagnosis of the general pathologists and to ensure that the pathologic diagnoses are made uniformly from one center to another. Without uniformity in diagnosis, the results of multiple clinical therapeutic trials could not be compared. Unfortunately, the panel is useful in only a small percentage (3 percent) of cases; the Pathology Panel annually reviews only 1000 cases, whereas more than 30,000 new cases of lymphomas are reported each year.

Our goal is to close the wide gap between the quality of diagnoses at community hospitals and diagnoses by scarce experts in pathology specialty fields with Pathfinder. This normative expert system promises to increase the accuracy of in-house pathology diagnoses, to reduce the frequency and cost of referrals, and to assist operating-room pathologists who cannot rely typically on the security blanket of an expert opinion.

4 The Pathfinder System

The Pathfinder expert system reasons about approximately 60 malignant and benign diseases of lymph nodes, constructing differential diagnoses through the consideration of evidence about the status of over 100 morphologic and nonmorphologic features visible in lymph-node tissue. In Pathfinder, *features* are each structured into a set of two to ten mutually exclusive and exhaustive *values*. These values typically represent the degree of severity of a particular feature (e.g., necrosis may be absent, present, or prominent).

4.1 Implementation History

In the course of our research, we have implemented several versions of the probability-based Pathfinder system. We constructed the earliest Pathfinder expert system with the Maclisp language on a Digital Equipment Corporation DEC-2060. Later, we transferred the program into the Portable Standard LISP (PSL) language and moved it to the Hewlett-Packard 9836 LISP workstation. Two years ago, we reimplemented the program in Macintosh Programmers' Workshop (MPW) Object Pascal on the Macintosh II. We have continued to refine and to test the knowledge base within the Macintosh II environment. As we shall explain, we implemented a knowledge-acquisition program in the same system designed to operate as a parallel application, enabling an engineer to cycle easily between knowledge-base refinement and expert-system testing.

4.2 System Functionality

The Pathfinder system allows a user to enter values for one or more salient features of a lymph-node section. Given these feature–value pairs, the system displays a differential diagnosis ordered by likelihood of diseases. In response to a query from the user, Pathfinder recommends a set of features that are the most cost effective for narrowing the differential diagnosis. The pathologist can answer one or more of the recommended questions. This process continues until the differential diagnosis is a single disease, there are no additional tests or questions, or a pathologist determines the informational benefits are not worth the costs of further observations or tests.

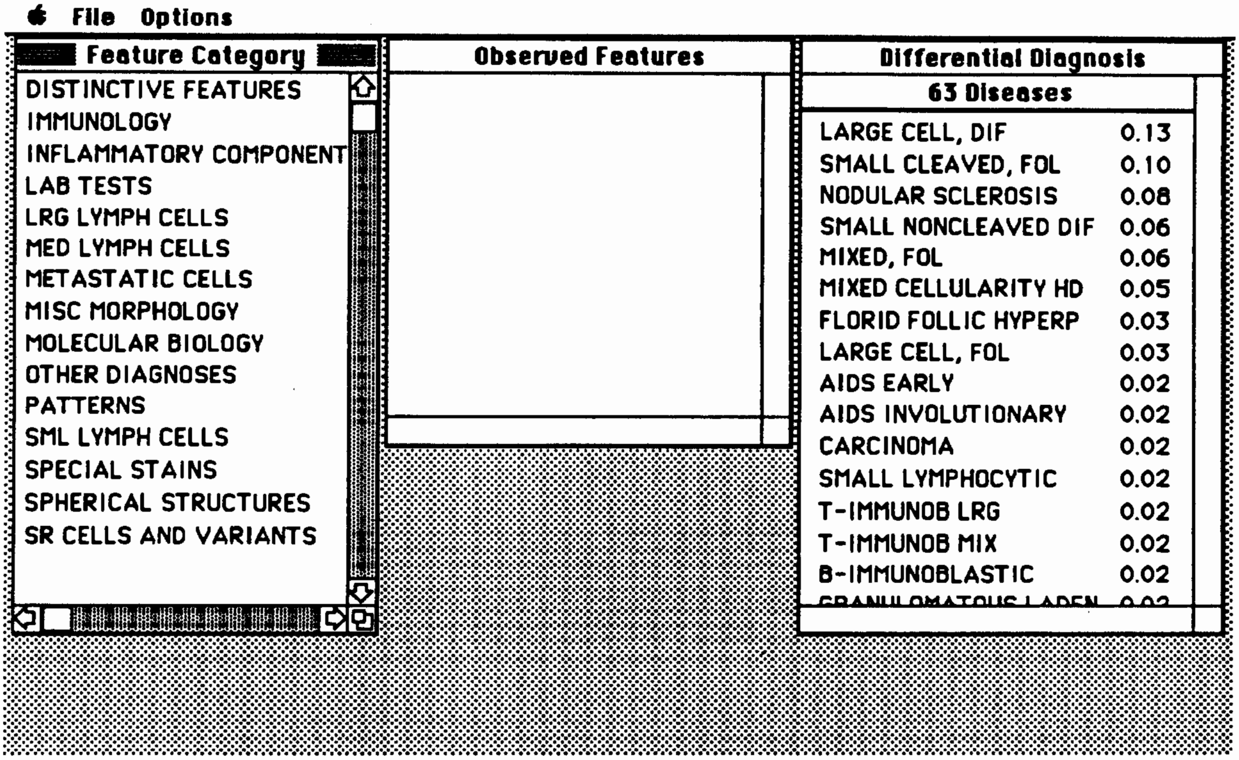
The operation of the latest version of Pathfinder is illustrated in Figures 2 through 8. Figure 2 shows the initial Pathfinder screen. The FEATURE CATEGORY window displays the categories of features that are known to the system; the OBSERVED FEATURES window displays evaluated features; and the DIFFERENTIAL DIAGNOSIS window displays the list of diseases with their associated probabilities. As there are no features observed at the outset of a case, the probabilities shown are the prior probabilities of disease.

If the user double-clicks on the feature category SPHERICAL FEATURES, then Pathfinder displays a list of atomic features for that category, as shown in Figure 3. To enter a particular feature, the user double-clicks on that feature, and then selects one of the mutually exclusive and exhaustive values for that feature. For example, Figure 4 shows what happens when the user selects the feature F % AREA (percent area of the lymph-node section that is occupied by follicles). In the figure, a third window appears that lists the values for this feature: NA (not applicable), 1–10%, 11–50%, 51–75%, 76–90%, and >90%. Figure 5 shows the result of selecting the last value for this feature. In particular, the feature–value F % AREA: >90% appears in the middle column, and the differential diagnosis is revised, based on this observation.

As we mentioned, the user can continue to enter any number of features of his own selection. Figure 6 shows the Pathfinder screen after the user has reported that follicles are in a back-to-back arrangement and show prominent polarity. Alternatively, the user can ask the program to recommend additional features for observation. Figure 7 shows that the most cost effective feature to evaluate, given the current differential diagnosis, is monocytoid cells. If the user observes that monocytoid cells are prominent, then we obtain the differential diagnosis in Figure 8. In this case, the four features in the middle column have narrowed the differential diagnosis to a single disease: the early phase of AIDS.

5 Hypothetico-Deductive Reasoning

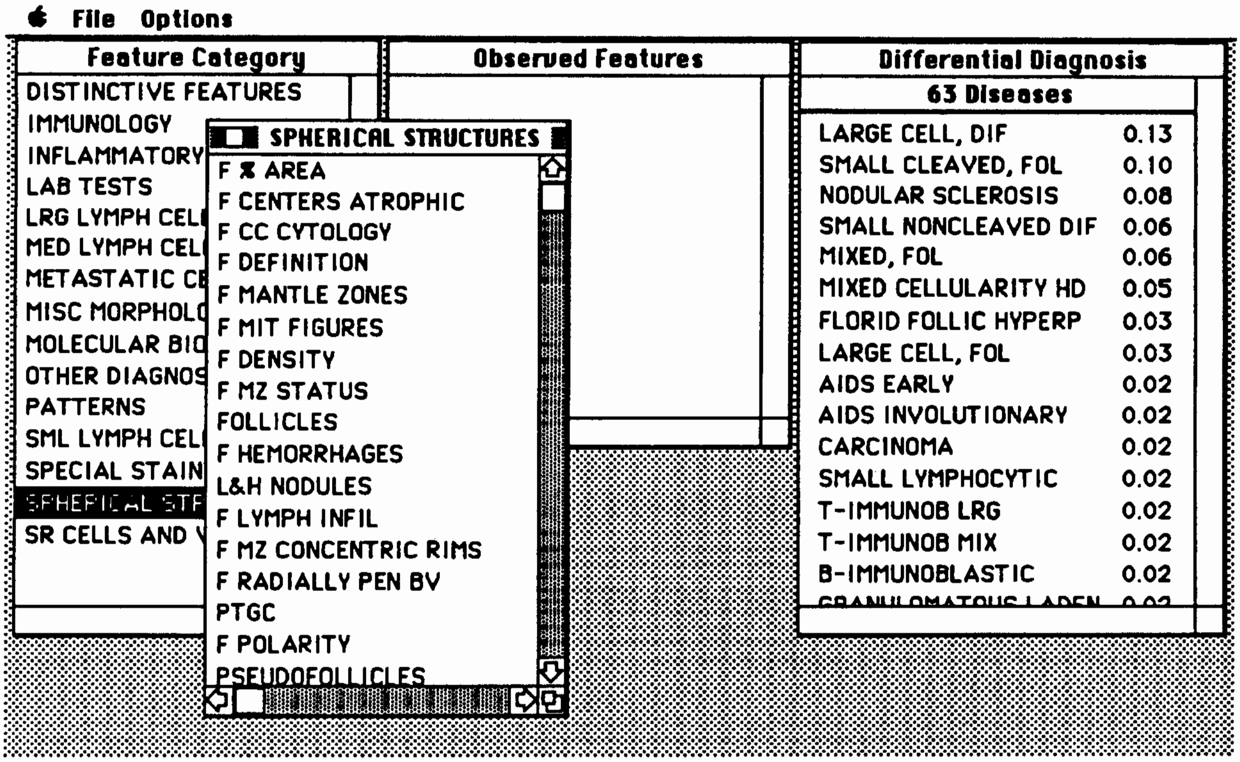
The basic reasoning architecture of Pathfinder is referred to as *hypothetico-deductive reasoning*. A flow-chart representation of the general method is shown in Figure 9. In hypothetico-deductive reasoning, a set of observations or test information is used to build a list of plausible hypotheses. The list of hypotheses is then examined in the process of determining the best



DISTINCTIVE FEATURES
IMMUNOLOGY
INFLAMMATORY COMPONENT
LAB TESTS
LRG LYMPH CELLS
MED LYMPH CELLS
METASTATIC CELLS
MISC MORPHOLOGY
MOLECULAR BIOLOGY
OTHER DIAGNOSES
PATTERNS
SML LYMPH CELLS
SPECIAL STAINS
SPHERICAL STRUCTURES
SR CELLS AND VARIANTS

LARGE CELL, DIF	0.13
SMALL CLEAVED, FOL	0.10
NODULAR SCLEROSIS	0.08
SMALL NONCLEAVED DIF	0.06
MIXED, FOL	0.06
MIXED CELLULARITY HD	0.05
FLORID FOLLIC HYPERP	0.03
LARGE CELL, FOL	0.03
AIDS EARLY	0.02
AIDS INVOLUTIONARY	0.02
CARCINOMA	0.02
SMALL LYMPHOCYTIC	0.02
T-IMMUNOB LRG	0.02
T-IMMUNOB MIX	0.02
B-IMMUNOBLASTIC	0.02
GRANULOMATOUS LADEN	0.02

Figure 2: The Pathfinder interface, displaying the FEATURE CATEGORY, OBSERVED FEATURES, and DIFFERENTIAL DIAGNOSIS window. Diseases are listed initially by their prior probabilities in the current clinical setting.

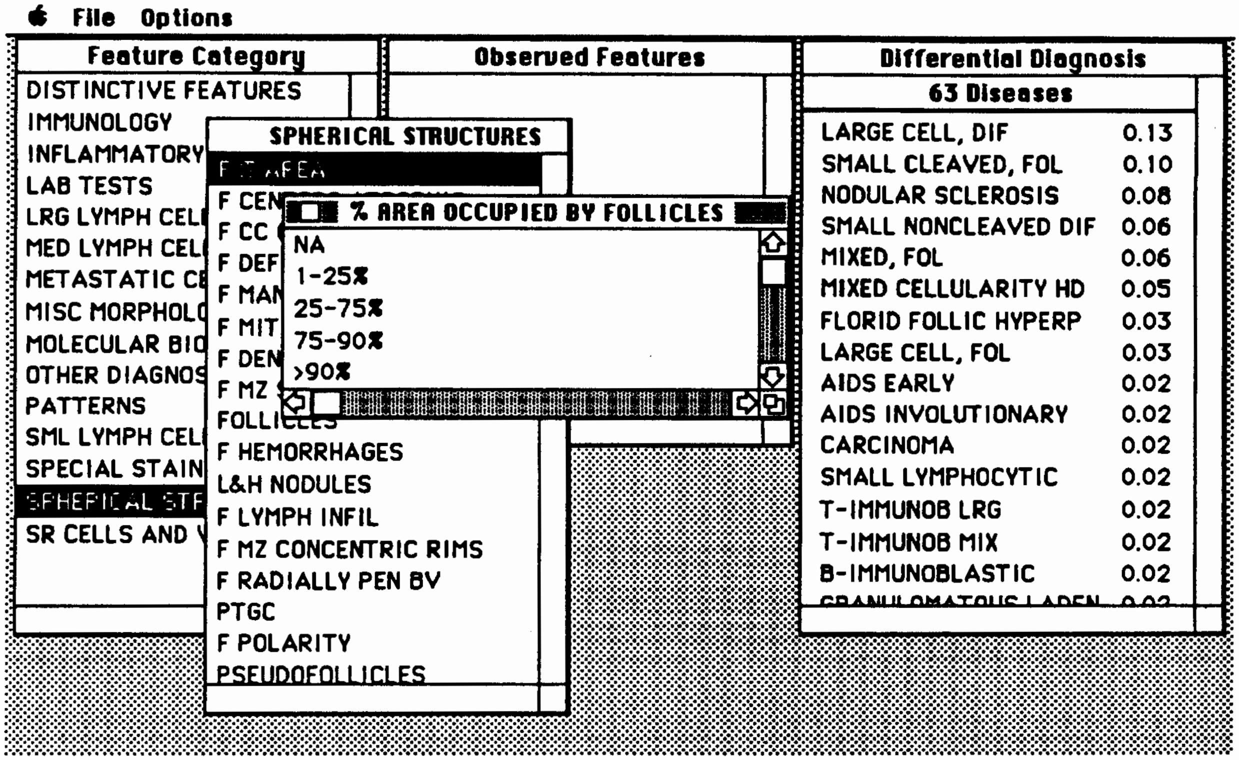


DISTINCTIVE FEATURES
IMMUNOLOGY
INFLAMMATORY
LAB TESTS
LRG LYMPH CEL
MED LYMPH CEL
METASTATIC C
MISC MORPHOL
MOLECULAR BI
OTHER DIAGNOS
PATTERNS
SML LYMPH CEL
SPECIAL STAIN
SPHERICAL STF
SR CELLS AND

F % AREA
F CENTERS ATROPHIC
F CC CYTOLOGY
F DEFINITION
F MANTLE ZONES
F MIT FIGURES
F DENSITY
F MZ STATUS
FOLLICLES
F HEMORRHAGES
L&H NODULES
F LYMPH INFIL
F MZ CONCENTRIC RIMS
F RADIALLY PEN BV
PTGC
F POLARITY
PSEUDOFOLLICLES

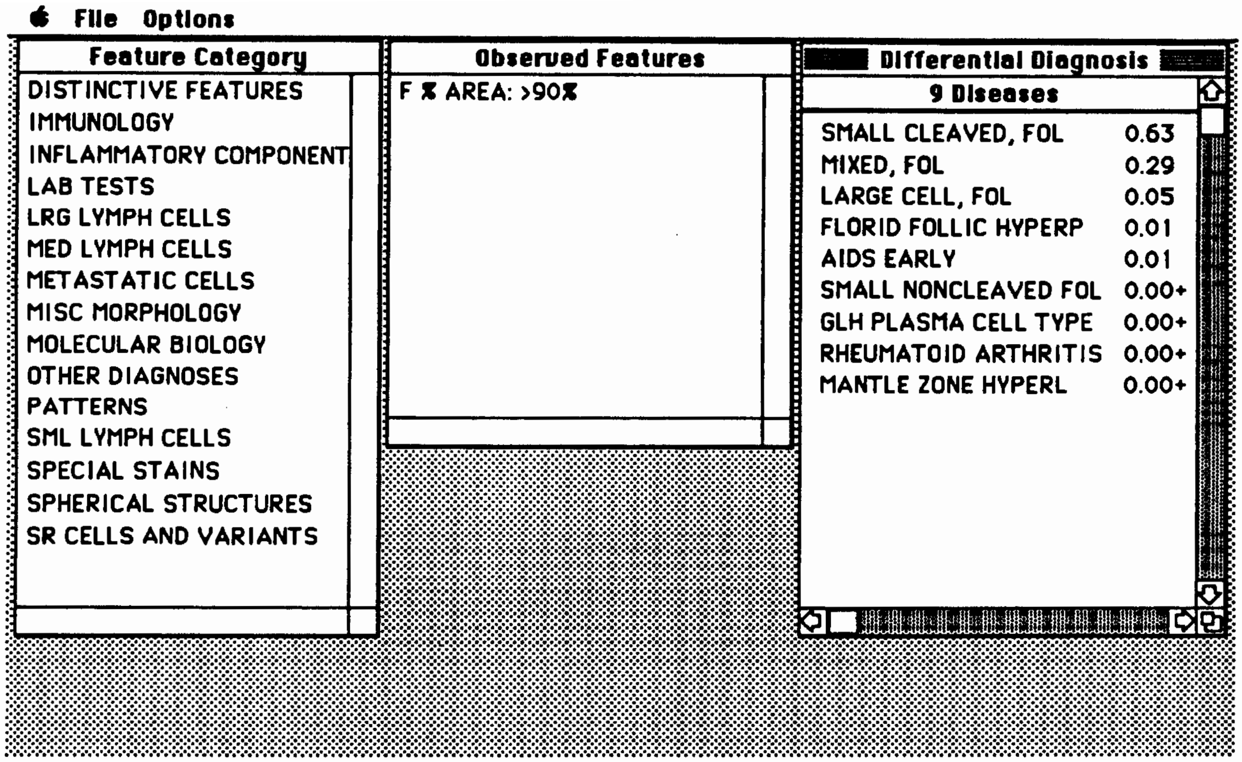
LARGE CELL, DIF	0.13
SMALL CLEAVED, FOL	0.10
NODULAR SCLEROSIS	0.08
SMALL NONCLEAVED DIF	0.06
MIXED, FOL	0.06
MIXED CELLULARITY HD	0.05
FLORID FOLLIC HYPERP	0.03
LARGE CELL, FOL	0.03
AIDS EARLY	0.02
AIDS INVOLUTIONARY	0.02
CARCINOMA	0.02
SMALL LYMPHOCYTIC	0.02
T-IMMUNOB LRG	0.02
T-IMMUNOB MIX	0.02
B-IMMUNOBLASTIC	0.02
GRANULOMATOUS LADEN	0.02

Figure 3: The expansion of a category of features, named SPHERICAL FEATURES, into a list of histologic features.



63 Diseases	
LARGE CELL, DIF	0.13
SMALL CLEAVED, FOL	0.10
NODULAR SCLEROSIS	0.08
SMALL NONCLEAVED DIF	0.06
MIXED, FOL	0.06
MIXED CELLULARITY HD	0.05
FLORID FOLLIC HYPERP	0.03
LARGE CELL, FOL	0.03
AIDS EARLY	0.02
AIDS INVOLUTIONARY	0.02
CARCINOMA	0.02
SMALL LYMPHOCYTIC	0.02
T-IMMUNOB LRG	0.02
T-IMMUNOB MIX	0.02
B-IMMUNOBLASTIC	0.02
GRANULOMATOUS LADEN	0.02

Figure 4: Entry of the feature F % AREA. This figure portrays the response of Pathfinder to a double-click of the mouse button on the feature describing the percent area of the lymph-node section occupied by follicles. The program displays a long version of the feature name and the list of the mutually exclusive and exhaustive values for the feature.



Feature Category

DISTINCTIVE FEATURES
IMMUNOLOGY
INFLAMMATORY COMPONENT
LAB TESTS
LRG LYMPH CELLS
MED LYMPH CELLS
METASTATIC CELLS
MISC MORPHOLOGY
MOLECULAR BIOLOGY
OTHER DIAGNOSES
PATTERNS
SML LYMPH CELLS
SPECIAL STAINS
SPHERICAL STRUCTURES
SR CELLS AND VARIANTS

Observed Features

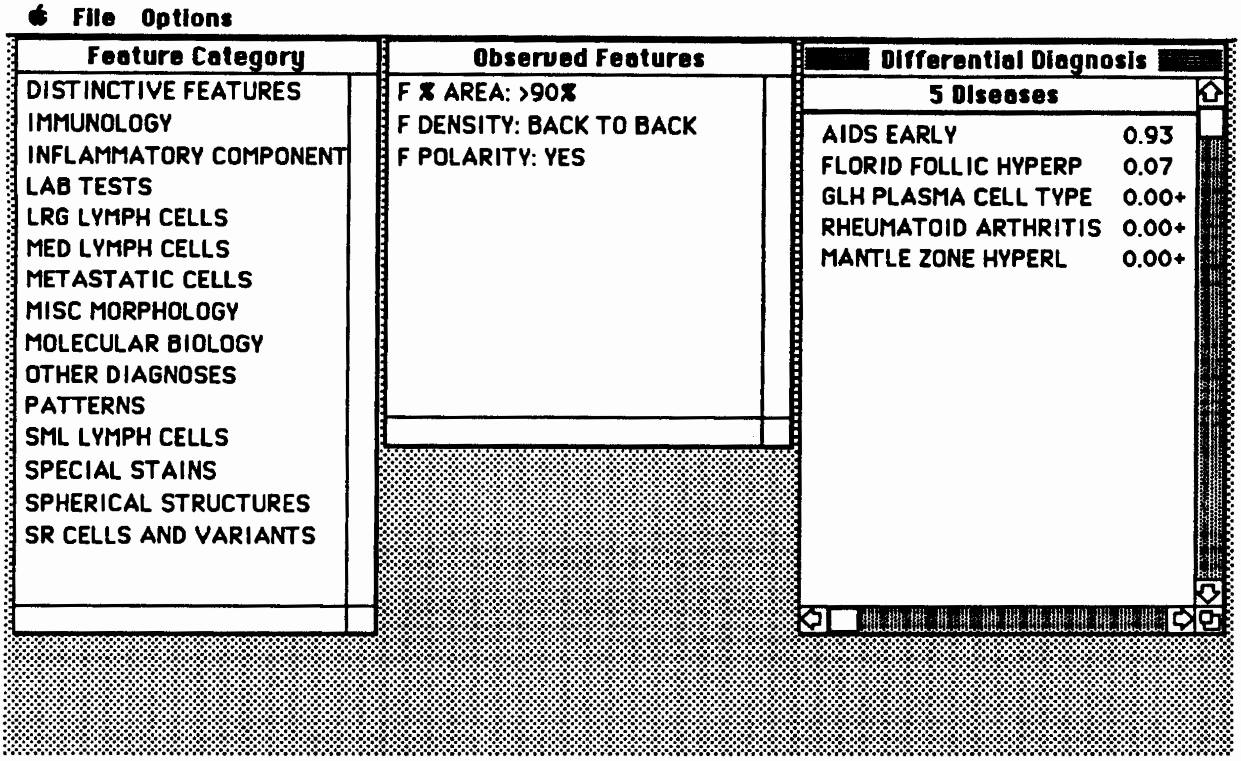
F % AREA: >90%

Differential Diagnosis

9 Diseases

SMALL CLEAVED, FOL	0.63
MIXED, FOL	0.29
LARGE CELL, FOL	0.05
FLORID FOLLIC HYPERP	0.01
AIDS EARLY	0.01
SMALL NONCLEAVED FOL	0.00+
GLH PLASMA CELL TYPE	0.00+
RHEUMATOID ARTHRITIS	0.00+
MANTLE ZONE HYPERL	0.00+

Figure 5: The result of entry of the feature F % AREA. The feature and the entered value—>90%—appear in the middle column. Based on this feature, the program revises the differential diagnosis in the right-hand column.



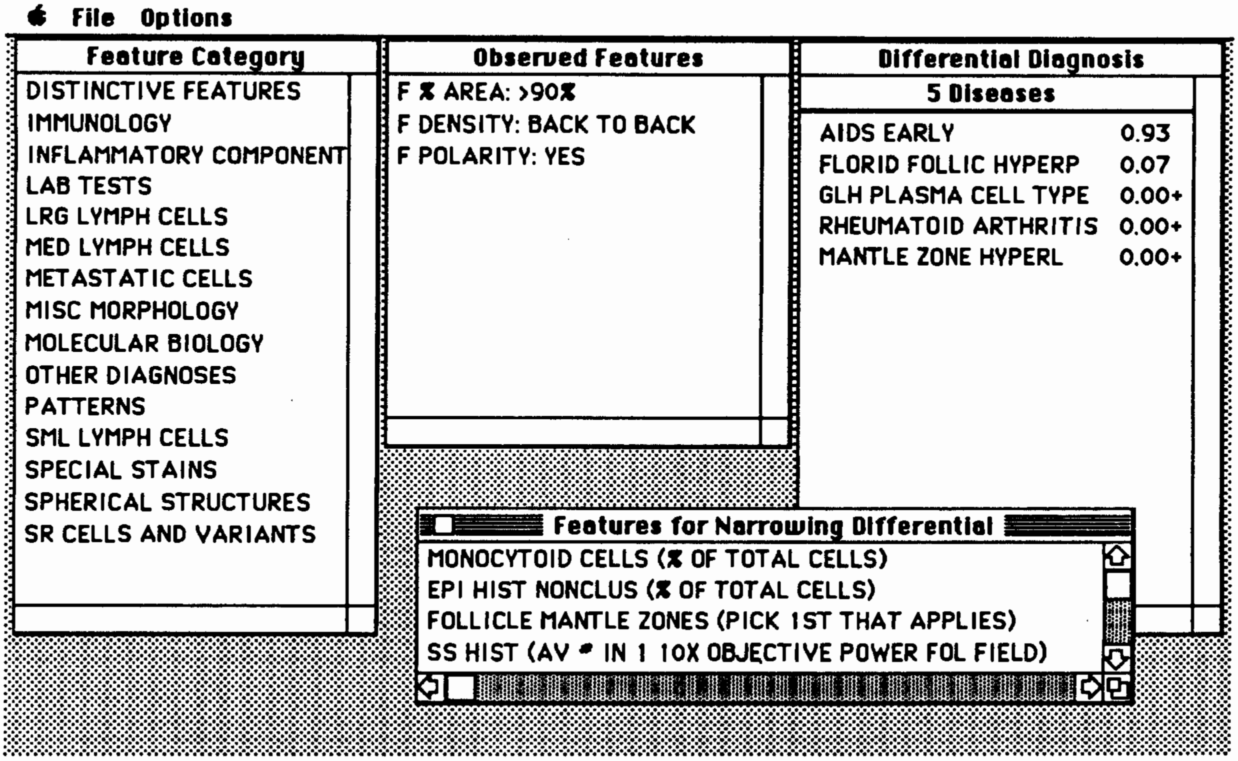
DISTINCTIVE FEATURES
IMMUNOLOGY
INFLAMMATORY COMPONENT
LAB TESTS
LRG LYMPH CELLS
MED LYMPH CELLS
METASTATIC CELLS
MISC MORPHOLOGY
MOLECULAR BIOLOGY
OTHER DIAGNOSES
PATTERNS
SML LYMPH CELLS
SPECIAL STAINS
SPHERICAL STRUCTURES
SR CELLS AND VARIANTS

F % AREA: >90%
F DENSITY: BACK TO BACK
F POLARITY: YES

5 Diseases

AIDS EARLY	0.93
FLORID FOLLIC HYPERP	0.07
GLH PLASMA CELL TYPE	0.00+
RHEUMATOID ARTHRITIS	0.00+
MANTLE ZONE HYPERL	0.00+

Figure 6: The display after two additional features have been entered. The observations that follicles are in a back-to-back arrangement (F DENSITY: BACK-TO-BACK) and show prominent polarity (F POLARITY: PROMINENT) appear in the middle column. The revised differential diagnosis appears in the right-hand column.



Feature Category

DISTINCTIVE FEATURES
IMMUNOLOGY
INFLAMMATORY COMPONENT
LAB TESTS
LRG LYMPH CELLS
MED LYMPH CELLS
METASTATIC CELLS
MISC MORPHOLOGY
MOLECULAR BIOLOGY
OTHER DIAGNOSES
PATTERNS
SML LYMPH CELLS
SPECIAL STAINS
SPHERICAL STRUCTURES
SR CELLS AND VARIANTS

Observed Features

F % AREA: >90%
F DENSITY: BACK TO BACK
F POLARITY: YES

Differential Diagnosis

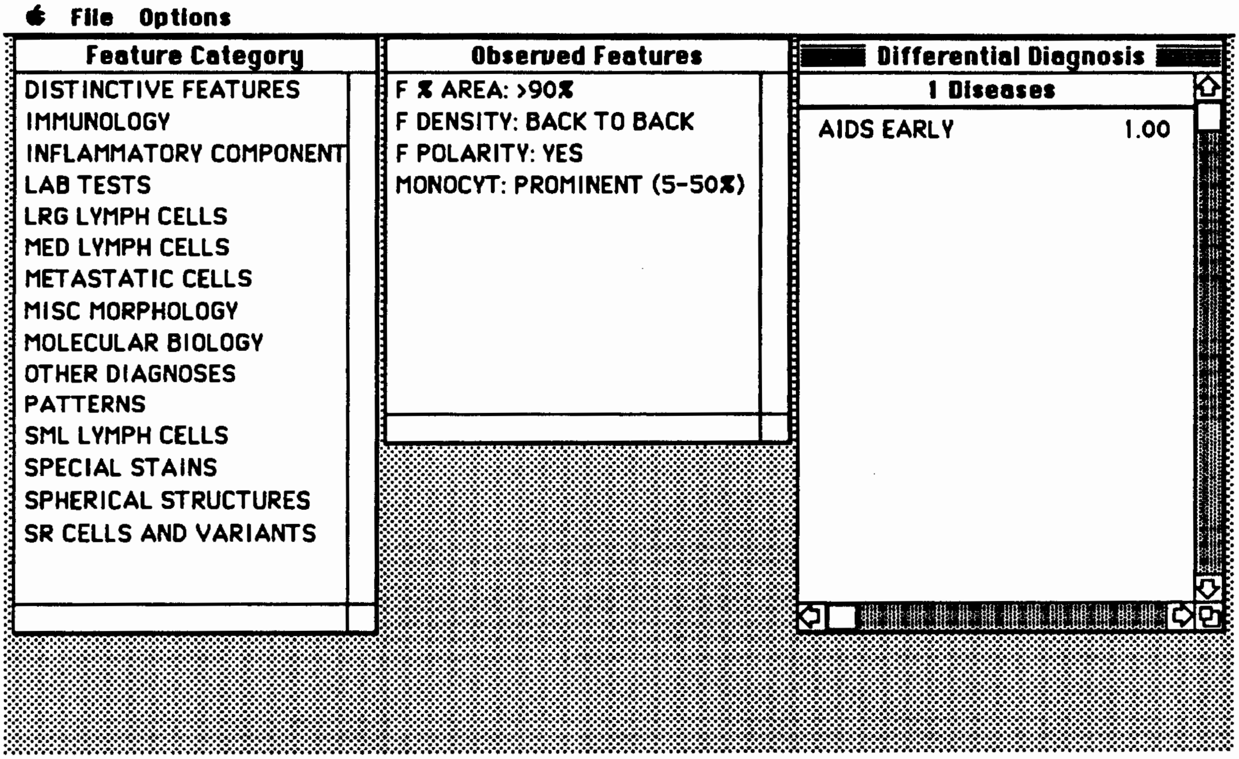
5 Diseases

AIDS EARLY	0.93
FLORID FOLLIC HYPERP	0.07
GLH PLASMA CELL TYPE	0.00+
RHEUMATOID ARTHRITIS	0.00+
MANTLE ZONE HYPERL	0.00+

Features for Narrowing Differential

MONOCYTOID CELLS (% OF TOTAL CELLS)
EPI HIST NONCLUS (% OF TOTAL CELLS)
FOLLICLE MANTLE ZONES (PICK 1ST THAT APPLIES)
SS HIST (AV # IN 1 10X OBJECTIVE POWER FOL FIELD)

Figure 7: The display after the user requests feature recommendations. Given the current differential diagnosis, Pathfinder displays the four most valuable features for the pathologist to observe next. The best feature is monocytoid cells.



DISTINCTIVE FEATURES
IMMUNOLOGY
INFLAMMATORY COMPONENT
LAB TESTS
LRG LYMPH CELLS
MED LYMPH CELLS
METASTATIC CELLS
MISC MORPHOLOGY
MOLECULAR BIOLOGY
OTHER DIAGNOSES
PATTERNS
SML LYMPH CELLS
SPECIAL STAINS
SPHERICAL STRUCTURES
SR CELLS AND VARIANTS

F % AREA: >90%
F DENSITY: BACK TO BACK
F POLARITY: YES
MONOCYT: PROMINENT (5-50%)

AIDS EARLY 1.00

Figure 8: Pathfinder determines that only a single disease—AIDS EARLY (the early phase of AIDS)—is consistent with the four observations.

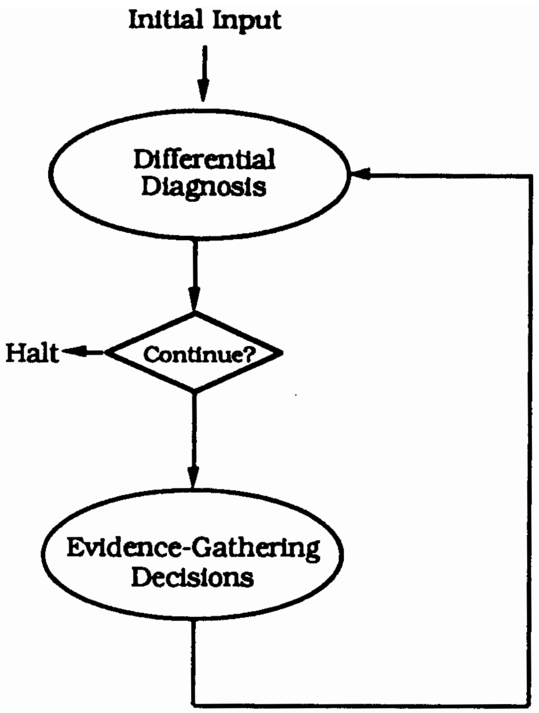


Figure 9: Hypothetico-deductive reasoning. First, several salient manifestations are identified and input to the system. Based on these manifestations, the system constructs a differential diagnoses—a list of hypotheses and likelihoods associated with these hypotheses. Next, the system analyzes the current differential diagnosis to identify the next best manifestations for the user to observe. The process cycles until the differential diagnosis is narrowed to a single disease, there are no additional tests or questions, or the user determines the informational benefits are not worth the costs of further observations or tests.

next observations to make. The process cycles to refine the differential diagnosis. Cognitive psychologists have found that physicians frequently employ hypothetico-deductive reasoning in performing clinical diagnosis [15,16]. The hypothetico-deductive strategy may also be important in the consultation process among colleagues, defining the communication between expert and referring physician during a consultation. The commonality of this strategy in clinical medicine suggests that the behavior of expert systems that perform hypothetico-deductive reasoning would be compatible with human diagnosticians.

As portrayed in Figure 9, there the two fundamental components of hypothetico-deductive reasoning. First, we must apply a method for determining the likelihood of each disease, given a set of observations. Second, we must identify those features or tests whose observation would be most useful in refining the differential diagnosis. Both of these tasks must be performed under uncertainty.

6 Overview of Paradigms for Uncertain Reasoning

We shall now examine normative and nonnormative approaches to these tasks, in a historical context. Then, we shall examine normative reasoning in Pathfinder in more detail.

6.1 Probability and Decision Theory

Probability theory has roots, over 3 centuries ago, in the work of Bernoulli, Laplace, Fermat, and Pascal. There has been a tremendous amount of theoretical and practical research on the theory and application of the principles of probability following the original pioneering studies of the early probabilists. Probability theory is a set of axioms that defines measures of belief in events or distinctions and describes how such measures can be made consistent or combined to infer measure of belief in related events. Decision theory extends the language of probability, and allows us to define and reason about *alternatives* or the actions available to a decision maker, the *outcomes* of each possible action, and the *preferences* of a decision maker for each possible outcome. In particular, as we have mentioned, decision includes probability theory and the maximum expected utility (MEU) principle [17].

6.1.1 Probability as Personal Belief

The prevalent conception of the probability for event x is that it is a measure of the frequency with which that event occurs, when repeated many times. A more general notion, however, is that the probability for event x represents the degree of belief held by a person that the event will occur. In this definition, we do not need to refer to the repetition of any event. A probability of 1 for x indicates that the person who owns that probability believes that the event will occur with certainty; a probability of 0 for x indicates that the persons believes that the event definitely will not happen. Intermediate values correspond to various degrees of uncertainty that the individual may have about the occurrence of x. Although this interpretation of a probability differs from that of the frequency interpretation, both measures obey the same set of axioms.

The conception of probability as a measure of personal belief is central to research on the use of probability and decision theory for representing and reasoning with expert knowledge in computer-based reasoning systems. There is usually no alternative to acquiring from experts the bulk of probabilistic information used in an expert system. Gathering a significant portion of frequencies through empirical study would entail much time and great expense. For example, there are over 75 thousand probabilities in the latest version of Pathfinder; and some of these probabilities are on the order of 10^{-6}. Furthermore, even when statistical studies have been preformed in some domain, we often cannot employ the frequencies that these studies produce, because the specific diseases, features, and contexts used in an expert system for that domain may not match those distinctions used in the studies. Nonetheless, probability theory provides for the gradual integration of appropriate statistical data into an expert system as it becomes available [18,19].

6.1.2 Normative Reasoning in Medicine

Decision theory prescribes widely accepted principles for rational belief and action under uncertainty. This is why people or computer systems that hold beliefs and make decisions that are consistent with probability and decision theory often are referred to as *normative* reasoners. Many investigators have studied the justification for decision theory. Their analyses show the equivalence between the axioms of decision theory and a set of fundamental properties about manipulating beliefs and choosing actions that are found to be intuitive and persuasive [20,17,21,22,23].

Research demonstrating differences between the *descriptive* behavior of humans and the recommendations of normative decision models has highlighted the potential value of normative reasoning systems for physicians. Studies have found that people are plagued with judgment biases [3]. More specifically, such biases have been demonstrated in medical problem solving [4,16]. Under the typical situations of high stakes in medicine, physicians may make decisions that, from the perspective of a normative analysis, incur great costs to patients. The goals of Pathfinder research are to codify expert judgments in a normative representation and to provide to nonexpert clinicians conclusions drawn from these beliefs. As we shall see in Part III, such an approach can be shown to deliver knowledge that is superior to knowledge provided by unassisted experts.

6.1.3 Early Normative Systems

Early discussions and research projects on the automation of medical reasoning centered on the implementation of normative expert reasoners [24]. Investigators worked to build systems that might some day provide to physicians conclusions drawn from knowledge that is encoded and manipulated in accordance with normative principles. Several medical diagnostic systems were constructed. These systems include Warner's system for the diagnosis of heart disease [25] and deDombal's system for the analysis of acute abdominal pain [26].

Two of the early expert-systems research projects applied hypothetico-deductive reasoning. The earliest work on hypothetico-deductive reasoning was performed by Gorry [27]. Gorry applied this methodology to Warner's heart-disease knowledge base, and, later, to a knowledge base for renal failure [28]. The approach, called the *method of sequential diagnosis*, was an improvement on older programs that required that all relevant findings in a patient case be present before they could make a diagnosis. As there are often hundreds of possible clinical findings to consider, the nonsequential approach has been considered less suitable for application in a clinical setting than are systems using the method of sequential diagnosis.

To build a differential diagnosis, Gorry's systems (and other early probabilistic systems) used probability theory. Further, to generate questions in his systems, Gorry applied a value-of-information analysis based on decision theory. We shall describe this analysis in Section 8. For both phases of the hypothetico-deductive analysis, Gorry's systems incorporated the assumptions that diseases are mutually exclusive and exhaustive, and that all pieces of evidence are conditionally independent given the disease present. He, and other investigators exploring normative reasoning, believed that it was necessary to make such assumptions to make knowledge acquisition, diagnostic inference, and explanation tractable.

6.2 Nonnormative Reasoning Methodologies

Although the principles of decision theory are well understood and are widely accepted as a gold standard or *normative* theory for decision making, medical-informatics researchers became interested in nonprobabilistic approaches to reasoning in the early-1970s. Growing perceptions of the inadequacy of the early decision-theoretic reasoning systems led to diminished interest in probability-based computational decision support [25,27,26]. Major problems cited with the probabilistic and decision-theoretic approaches were the complexity of building, representing, and manipulating knowledge bases. Investigators highlighted the inadequacy of making global assumptions of conditional independence, and of mutual exclusivity and exhaustivity, to regain tractability. Beyond problems with computational intractability, critics of the decision-theoretic reasoning in expert systems have cited limited expressiveness of normative representations, dwelling on the apparent differences between the quantitative approach of probabilistic inference and the informal, qualitative nature of human reasoning [29,30,28].

6.2.1 Heuristic Scoring Schemes

Several projects examined nonprobabilistic quantitative approaches within the hypothetico-deductive framework. A well-known example of the use of heuristic scoring schemes for capturing medical expertise is the Internist-1 project [31], and its descendents, the Quick Medical Reference (QMR) [32] and Caduceus [33] projects. Internist-1 and QMR are based on a hypothetico-deductive reasoning framework, similar to the strategy employed earlier by Gorry in his work with simple probabilistic reasoning systems. Unlike Gorry, however, the Internist-1 group elected not to use probability and decision theory, and instead applied heuristics for constructing the differential diagnosis and an ad hoc scoring scheme for assigning belief to competing entities. The Internist-1 team created several heuristic strategies to generate recommendations on new evidence to acquire, given the current differential diagnosis. These systems acquired and combined several classes of numeric weighting measures to build differential diagnoses and to direct evidence-gathering strategies.

6.2.2 Production-Rule Paradigm

In the early 1970s, several groups of medical informatics researchers, concerned with problems with the expressiveness of the restricted forms of probabilistic inference and with the complexity of attempts to relax the simplifying assumptions, became interested in the use of newer logic-based reasoning techniques, developed in AI research on theorem proving. The application of production rule systems to clinical medicine heralded the start of a new subdiscipline of AI, called artificial intelligence in medicine (AIM). AIM researchers sought to replace what were viewed as inappropriately complex quantitative approaches with methods for performing more abstract symbolic reasoning.

AIM researchers, interested in logical reasoning methods, sought to apply logical rules of the form *IF* A, *THEN* B—called *production rules*—to medical diagnosis. In a production-rule system, rules are chained together through the logical interaction of their antecedents and consequents, forming a directed graph. Production rules were appealing in that they were considered to be a modular approach to representing expert knowledge [34]. The traditional production rule was interpreted as a straightforward logical implication. Unlike the probabilistic and nonnormative hypothetico-deductive approaches, production-rule approaches generate questions based on a relatively fixed traversal of the directed graph formed by the set of rules.

Attempts to apply production rules to reasoning about complex medical problems stimulated the AIM community to introduce techniques for representing uncertainty to representations built on the foundations of logical chaining. That is, in applying production rules to real-world diagnosis, some investigators saw a need to modify the true-or-false nature of rules, to capture uncertainty about implication. Researchers working on the Mycin project introduced the quasiprobabilistic *certainty-factor* (CF) model, a numeric scheme for representing the degree of confirmation or disconfirmation of the consequent of a rule, given the rule's antecedent [34].

7 Pathfinder: A Return to Decision Theory

The first versions of the Pathfinder expert system were based on production rules. The rule-based versions of Pathfinder were implemented in the Meta-Level Representation System (MRS) [35]. Early on, we encountered two related problems with the behavior of rule-based systems for reasoning in pathology. First, we found that the production-rule system's sequences of questions appeared inflexible and inappropriate to our expert. In particular, the rule-based system did not make use of the current differential diagnoses to determine the most useful features to observe. Second, our rule-based approach did not consider uncertainty. Given the closely related symptomology of many lymph-node diseases, there is often uncertainty in pathology diagnosis, associated with a set of observations. For example, several diseases might be possible given a set of histologic features. In such cases, it is important to consider the different likelihoods of the diseases.

The *descriptive inadequacy* of the MRS-based Pathfinder system stimulated us to implement a hypothetico-deductive approach. Also, we experimented with three belief-combination schemes to construct differential diagnoses: (1) the quasiprobabilistic Dempster–Shafer [36] approach, (2) the Mycin CF approach, and (3) simple probabilistic inference, assuming conditional independence among features given diseases. We evaluated the behavior of our system informally with these alternative combination schemes.

7.1 Early Empirical Results

During masked studies, we noticed a significant improvement in the system's diagnostic accuracy when the combination scheme was converted from the quasiprobabilistic approaches to simple probabilistic inference, assuming conditional independence. When we returned to a methodology similar to the approach explored 2 decades earlier by Warner, Gorry, and other medical-informatics pioneers, Pathfinder performed noticeably better than did the other approaches. Later, in a formal study, we showed that the diagnostic accuracy of the simple conditional-independence probabilistic method was superior to that of the Dempster–Shafer and CF updating schemes [37]. The increased performance was quantified with an evaluation scheme incorporating decision-theoretic and ad hoc measures. We shall describe these metrics in detail in Part III.

Like many AIM researchers at the time, we believed that probability and decision theory were inadequate for representing medical expertise. Yet, our empirical evidence highlighted the promise of these techniques. Our best-available method for reasoning was a highly constrained form of probabilistic inference. We believed that we could produce even more accurate diagnostic behavior by relaxing the conditional-independence restrictions.

We were familiar with the literature in AI and medical informatics that warned about the intractability of relaxing the assumption of independence. This fear of intractability was based on a sense that a move beyond a simple conditional-independence model would necessarily encounter massive interdependence among pieces of evidence. Nonetheless, we believed that we could capture and manage all dependencies that could be identified as relevant by an expert. We speculated that cognitive limitations might constrain the complexity of our normative computer-based models for diagnosis; even in cases where there is, in reality, great interdependency among evidence and diseases, experts may attend to only the most salient dependencies.[1] In cases where our models were so complex as to require intractable representational and computational effort, we could seek to reason with decision theory about tradeoffs associated with alternative approximations [38,39].

The acquisition of a comprehensive dependency model promised to capture the knowledge of experts with fidelity. We set out to examine the nature of dependencies in pathology, and to develop techniques for managing the complexity associated with acquiring, representing, and reasoning with these dependencies. We also undertook in-depth study of the foundations of belief and action under uncertainty and of the relationship of the nonprobabilistic schemes to probability theory.

[1] Of course, we may be able to extend a normative model, built to capture an expert's view of the world, with new dependencies introduced through statistical study.

We are not alone in the investigation of probabilistic and decision-theoretic inference in medical expert systems; several other recent or ongoing research projects have explored normative reasoning in medical expert systems. Projects include the Nestor system for reasoning about endocrinology disorders [40], the Glasgow Dyspepsia expert system for assisting in gastroenterology diagnosis [41], the Neurex system for diagnosis of neurological findings [42], the Medas system for assisting physicians in emergency medicine [43], and the Munin system for diagnosis of muscular disorders [44].

7.2 Theoretical Study of Nonnormative Methods

Following our empirical analyses, we embarked on theoretical work addressing the relationships of probability and decision theory to alternative methods. In some of our earliest theoretical research, we identified commonalities in several nonprobabilistic reasoning methodologies. A number of the popular belief-updating schemes employed by AIM researchers, including the Mycin CF approach and the Internist-1 scoring scheme, could be shown to be based on a stereotypical *modular-updating paradigm* [45]. Such schemes combine independent measures of *change* in belief, or *belief updates*. We worked to develop an understanding of what the foundations of belief-updating schemes are and of how alternative approaches were related to probability theory [46,47]. We showed how AIM researchers had without justification assumed modularity among belief updates [48], and we demonstrated that uncertain beliefs are fundamentally less modular than are beliefs held with certainty. We discussed how the nonmodularity of uncertain knowledge frequently makes the rule-based calculi unusable or intractable for reasoning with uncertainty in a coherent manner [49].

A detailed analysis showed that the original Mycin CF model is inconsistent with probability theory and that we could make it consistent with a relatively simple modification [47]. The analysis clearly showed that the consistent version of the CF model implied independence constraints on evidence that were even stronger than had been assumed by the medical-informatics researchers who had examined probabilistic updating with the assumption of conditional independence; that is, investigators using the CF model were implicitly assuming that evidence was conditionally independent given diseases hypotheses, as well as given their negation [47].

We also found that the Mycin CF approach, and several other ad hoc belief-combination schemes, confused many AI researchers and medical informaticists about belief-updates versus absolute beliefs [47,50]. Certainty factors were defined to be measures of the degree of change in belief in a hypothesis, given evidence. Yet many people referred to—and assessed—these quantities as measures of absolute belief.

More recent discussions, building on the earlier analyses, have also described the inadequate handling of prior probabilities, or prevalence rates, in rule-based expert systems [51,52]. When a CF-based updating scheme makes a recommendation, it uses measures of belief assigned to competing diseases without consideration of priors. Thus, in effect, a system based on this scheme treats all diseases as having equal prior probabilities.

In other work on the analysis of the relationship of probability theory to heuristic uncertainty-management schemes, we have looked at the rationale that motivated researchers

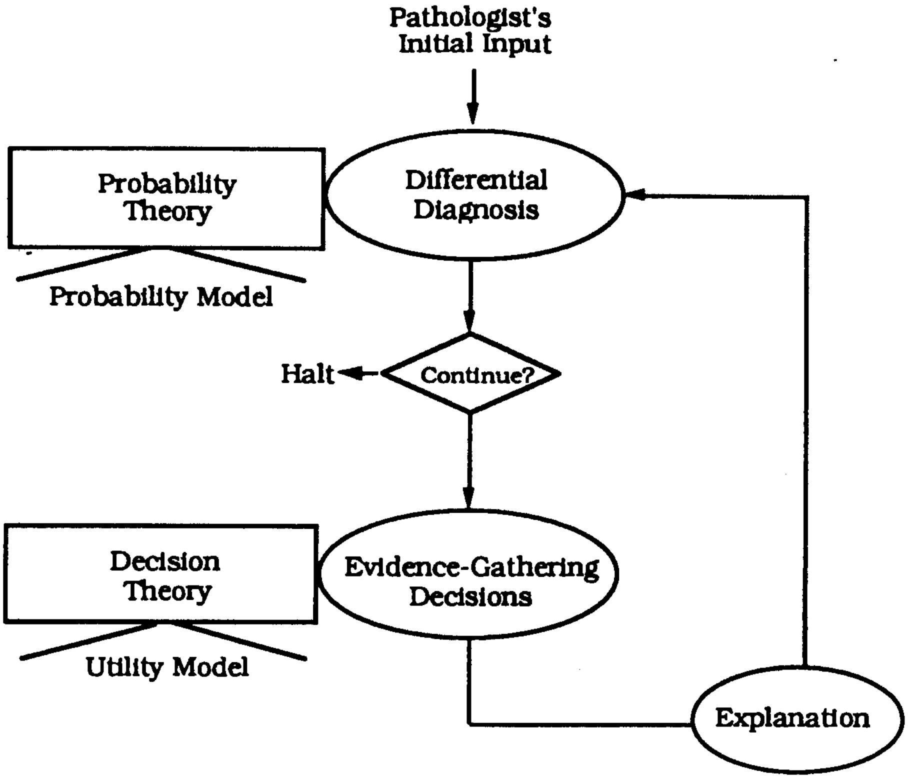


Figure 10: Overview of Pathfinder's hypothetico-deductive architecture. First, the pathologist enters salient features. Next, a probabilistic reasoner determines a differential diagnosis, based on those features. Then, a value-of-information analysis identifies features that are useful for narrowing the differential diagnosis. At the end of each cycle, the system makes available explanations of its recommendations.

to create different nonprobabilistic schemes [51,53]. We have also examined the relationship of the Internist-1 scoring scheme to probability theory [46]. In this work, performed in collaboration with the QMR team, we found relationships between numeric quantities used in the Internist-1 system and probabilistic quantities.[2]

8 Decision-theoretic Inference in Pathfinder

In this section, we discuss details of the decision-theoretic computations employed in the hypothetico-deductive approach of Pathfinder. We also contrast these techniques to those used in earlier normative medical expert systems.

As shown in Figure 10, we use probability theory to construct the differential diagnosis from a set of observations. This analysis is performed on a probability model that captures the diagnostic prowess of the expert pathologist. Next, we use a decision-theoretic calcula-

[2] This work has led to ongoing research on the transformation of the QMR knowledge base to a probabilistic model, and on the concomitant development of a normative version of the QMR system [54,55].

tion to identify morphologic features or laboratory tests that are useful for narrowing this differential diagnosis. In this phase, we examine both the costs and the benefits of observing these features or tests.

8.1 Computation of the Differential Diagnosis

As we discussed in Section 4, distinctions in the Pathfinder model include disease states, features, and a set of mutually exclusive and exhaustive values for each feature. Let m and n denote the number of diseases and features, respectively. Also, let $d_1, d_2, \ldots, d_m$ denote the disease entities. For the moment, let us suppose that each disease d_i may be present or absent. Let D_j denote some *instance* of diseases. That is, D_j denotes some assignment of present or absent to each of the diseases $d_1, d_2, \ldots, d_m$. For example, D_4 may represent the state in which a patient has both AIDS and a nonHodgkin's lymphoma, but no other disease. Further, let f_1, f_2, $\ldots$, f_n denote the features, and let v_i denote an observed value for the ith feature. Now imagine that we have observed values for q features. To simplify the notation, let us renumber the n features so that we have observed values for the first q features. In Pathfinder, we are interested in determining the probability of each disease instance, given the observations f_1v_1, f_2v_2, $\ldots$, f_qv_q. This quantity for disease instance D_j is known as the *posterior probability* of D_j, and is denoted

$$p(D_j|f_1v_1, f_2v_2, \ldots, f_qv_q)$$

Thus, the number of probabilities we seek to determine is exponential both in the number of observed features and in the number of diseases.

In principle, we can assess directly the posterior probabilities from an expert. Aside from the intractable nature of this task, most physicians are more comfortable assessing probabilities in the opposite direction. That is, they are more comfortable assessing the probabilities that the set of observations f_1v_1, f_2v_2, $\ldots$, f_qv_q will appear given a particular disease instance D_j, denoted

$$p(f_1v_1, f_2v_2, \ldots, f_qv_q|D_j)$$

Using Bayes' theorem, we can compute from these probabilities and *prior probability* of diseases instances $p(D_j)$, the desired posterior probabilities

$$p(D_j|f_1v_1, f_2v_2, \ldots, f_qv_q) = \frac{p(f_1v_1, f_2v_2, \ldots, f_qv_q|D_j)\ p(d_j)}{\sum_{D_k}\ p(f_1v_1, f_2v_2, \ldots, f_qv_q|D_k)\ p(D_k)}$$

where the sum over D_k runs over all disease instances. Unfortunately, this approach to the problem is also intractable, because the number of probabilities of the form $p(f_1v_1, f_2v_2, \ldots, f_qv_q|D_j)$ is exponential both in the number of diseases and in the number of features.

To manage the complexity of the general case, researchers, in the past, made two assumptions. First, they supposed that *all* findings were *conditionally independent*, given any disease state. That is, they assumed that, if the true disease state of the patient was known, then the likelihood of seeing any observation $f_i v_i$ did not depend on observations made about any other features. Thus,

$$p(f_i v_i | D_j, f_1 v_1, \ldots, f_{i-1} v_{i-1}, f_{i+1} v_{i+1}, \ldots, f_q v_q) = p(f_i v_i | D_j)$$

Given this assumption, it follows from the axioms of probability [56] that

$$p(f_1 v_1, f_2 v_2, \ldots, f_q v_q | D_j) = p(f_1 v_1 | D_j)\ p(f_2 v_2 | D_j)\ \ldots,\ p(f_q v_q | D_j)$$

Second, investigators supposed that the traditional disease entities were mutually exclusive and exhaustive. That is, they assumed that each disease instance corresponded to a situation where only one disease was present. Given these two assumptions, the posterior probabilities of disease were determined from the tractable computation

$$p(d_j | f_1 v_1, f_2 v_2, \ldots, f_q v_q) = \frac{p(f_1 v_1 | d_j)\ p(f_2 v_2 | d_j)\ \ldots,\ p(f_q v_q | d_j)\ p(d_j)}{\sum_{d_k} p(f_1 v_1 | d_k)\ p(f_2 v_2 | d_k)\ \ldots,\ p(f_q v_q | d_k)\ p(d_k)}$$

where d_j represents the disease instance in which only disease d_j is present. Thus, only the conditional probabilities $p(f_i v_i | d_j)$ and the prior probabilities $p(d_j)$ are required for the computation.

In Pathfinder, the assumption that diseases are mutually exclusive is appropriate, because co-occurring diseases almost always appear in different lymph nodes or in different regions of the same lymph node, and a user can analyze each area of pathology separately. Also, the large scope of Pathfinder makes reasonable the assumption that the set of diseases is exhaustive. The assumption of global conditional independence, however, is highly inaccurate. For example, given certain diseases, the finding that follicles are abundant in the tissue section increases greatly the chances that sinuses in the interfollicular areas will be partially or completely destroyed. Thus, Pathfinder has provided an excellent testbed to study, in isolation, one of the central difficulties in knowledge acquisition for normative expert systems.

8.2 Computation of Recommendations for Evidence Gathering

Suppose we have observed the set of features $F = \{f_1, f_2, \ldots, f_q\}$. Given these observations, should we examine some other set of features F_{new}? In the decision-theoretic framework, we can determine the appropriate action—to observe or not to observe the additional features—by computing the *value of information* of observing F_{new}, and by comparing this value to the cost of observing F_{new}. We should make the additional observations if and only if the value of information exceeds the cost. If there is no set of features for which this criterion holds, we should halt the hypothetico-deductive cycle, and make a diagnosis.

To compute the value associated with observing a set of additional features, we need to know the decision maker's preference or *utility* for each possible clinical outcome. Let U_{d_i,d_j} denote the utility of the situation in which a patient has disease d_i and is diagnosed as having d_j. In Section 10, we discuss how to assess these quantities. Given a matrix of all U_{d_i,d_j}, we first determine the optimal diagnosis for the patient, under the assumption that we observe no additional features. To determine this diagnosis, we use the MEU principle, which states that the optimal diagnosis is the one that maximizes the expected utility of the patient. Formally, let ϕ denote the set of feature–value pairs f_1v_1, f_2v_2, ..., f_qv_q that we have observed thus far. The optimal diagnosis, given observations ϕ, denoted $dx(\phi)$, is then given by

$$dx(\phi) = \mathrm{argmax}_{d_j} \left[\sum_{d_i} p(d_i|\phi)\ U_{d_i,d_j} \right] \tag{1}$$

where the function $\mathrm{argmax}_{d_j}[\cdot]$ returns the diagnosis that maximizes its argument. We can also compute the expected utility of this diagnosis, given the observations ϕ. This quantity, denoted $EU(dx(\phi)|\phi)$, is given by

$$EU(dx(\phi)|\phi) = \sum_{d_i} p(d_i|\phi)\ U_{d_i,dx(\phi)} \tag{2}$$

Now imagine that we observe an additional set of features F_{new}. Let F' denote the union of these features and the features in F, and let ϕ' denote the set of observations for the features in F'. Given these observations, we can compute the optimal diagnosis $dx(\phi')$ and its expected utility $EU(dx(\phi')|\phi')$:

$$dx(\phi') = \mathrm{argmax}_{d_j} \left[\sum_{i} p(d_i|\phi')\ U_{d_i,d_j} \right] \tag{3}$$

$$EU(dx(\phi')|\phi') = \sum_{d_i} p(d_i|\phi')\ U_{d_i,dx(\phi')} \tag{4}$$

We can also compute the expected utility of the original diagnosis, given observations ϕ', denoted $EU(dx(\phi)|\phi')$

$$EU(dx(\phi)|\phi') = \sum_{d_i} p(d_i|\phi')\ U_{d_i,dx(\phi)} \tag{5}$$

The quantity $EU(dx(\phi)|\phi')$ is never greater than the measure $EU(dx(\phi')|\phi')$, because, by definition, the diagnosis $dx(\phi')$ maximizes the expected utility of the patient, given the observations ϕ'. The difference between $EU(dx(\phi')|\phi')$ and $EU(dx(\phi)|\phi')$ represents the value of observing the set of features F_{new}. To determine this value, in general, we must average over all the possible observations associated with the new features. Let $EV(F_{\mathrm{new}}|\phi)$ denote the value of observing the features in F_{new}, given that we have already made observations ϕ about the features in F. We obtain

$$EV(F_{\mathrm{new}}|\phi) = \sum_{\phi'} p(\phi'|\phi) \left[EU(dx(\phi')|\phi') - EU(dx(\phi)|\phi') \right] \tag{6}$$

The computation of $EV(F_{\mathrm{new}}|\phi)$ is exponential in the number of features in the set F_{new}. Consequently, researchers have adopted a *myopic policy*, in which the elements of F_{new} are restricted to one or a small number of features. In principle, this assumption could affect the diagnostic accuracy of an expert system, because evidence gathering might be halted prematurely.[3] Nonetheless, Gorry has shown that the use of an approximation where F_{new} is restricted to a single feature does not diminish significantly the diagnostic accuracy of an expert system for congenital heart disease [27]. In Pathfinder, we similarly restrict F_{new} to a single feature.

Part II
New Research

9 Construction of a Normative Knowledge Base

We shall now focus on one of the most problematic areas of using decision theory in diagnosis: the construction of a probability model for computing differential diagnoses. The phrase "knowledge-acquisition bottleneck" has been used frequently among researchers in medical informatics and in artificial intelligence to express frustration about the difficult process of encoding knowledge. Indeed, the acquisition of knowledge has been considered a major impediment to constructing genuinely useful computer aids. The knowledge acquisition problem has been especially salient in the case of probability-based reasoning systems.

The construction of a normative knowledge base entails three phases: (1) identify important distinctions, (2) acquire relevant dependencies among these distinctions, and (3) quantify the probabilistic strengths of the dependencies. Our decomposition of the model-construction process into distinct phases should not imply a rigid separation among the components of model building. We have found that there is typically a great deal of interaction among the phases, because work on the development of one aspect of model construction tends to lead to the incremental refinement of other aspects of the model.

9.1 The Identification and Refinement of Distinctions

The first phase of knowledge-base construction involves the identification of the disease entities, the features relevant to diagnoses, and the set of mutually exclusive and exhaustive values associated with each feature. One of the most difficult components of this portion of knowledge acquisition is making the distinctions unambiguous. To accomplish this task, we use a technique from decision analysis called the *clarity test.* Consider the event *large cells are abundant.* Using the clarity test, we ask the expert: "Would an omniscient being be able to determine the presence of abundant large cells with certainty in a particular lymph-node section?" If the expert answers "no," the distinction does not pass the clarity test. We must then work to make the definition of the distinction more precise. This task is typically an iterative one. In our example, the expert might settle on the more precise

[3] For example, suppose that only two feature remain unobserved. In this case, a value-of-information analysis on each feature alone might indicate that neither feature is cost effective for observation, yet the value of information for the feature pair could exceed the cost of their observation. Furthermore, the observation of these two features could change the diagnosis significantly.

event of ***lymphoid cells greater than 20 microns in diameter occupy more than 70 percent of the total cell population in nonfollicular areas.***

Pathfinder's diseases, features, and feature-value pairs were determined in several meetings among four hematopathology experts (Drs. Costan Berard, Jerome Burke, Ronald Dorfman, and Bharat Nathwani). The initial consensus model was developed through spirited communication among the experts and knowledge engineers. During this process, the knowledge engineers noted that many features traditionally considered as independent pieces of evidence in pathology could be shown to have overlapping or interdependent relevance to diseases. Attempts to clarify and redefine features, so as to optimize the features' independent diagnostic relevance, were typically received with enthusiasm by the pathologists. We found that many features that had been described traditionally by pathologists had not before been scrutinized from an informational perspective.

9.2 Capture and Representation of Probabilistic Dependencies

Two difficulties have provided major challenges to investigators attempting to construct normative expert systems in medicine: (1) more than one disease may be present at one time; and (2) features are probabilistically dependent on one another, even when diseases that are present in a patient are known with certainty. As we discussed in Section 8.1, Pathfinder has provided an excellent testbed to study the latter problem in isolation.

9.2.1 Belief Networks

We have addressed the problem of probabilistic dependencies by using a representation developed in the decision-science community called *belief networks* [57,58]. The belief network is a graphical knowledge-representation language that encodes probabilistic dependencies among propositions and events. The representation rigorously describes probabilistic relationships, yet has a human-oriented qualitative structure that facilitates communication between the expert and the probabilistic model. Moreover, the representation can represent any probabilistic-inference problem.[4]

A belief network is a directed acyclic graph that contains nodes that represent distinctions, and arcs that represent dependencies among those distinctions. Figure 11 shows a belief network for the problem of distinguishing NS (ordinary nodular sclerosis Hodgin's disease) from CP (cellular phase of nodular sclerosis Hodgin's disease). The node DISEASE represents the two possible diseases, and the nodes CAP THICKENING (capsule thickening of 10 or more lymphocyte diameters), FCB (fibrocollagenous bands), and FIBROSIS represent the features that are relevant to the discrimination of these two diseases. Each node in the belief network is associated with a set of mutually exclusive and exhaustive values. For this inference problem, the values of the node DISEASE are NS and CP, and the values for each of the features nodes are T (true or present) and F (false or absent).

In a belief network, an arc from node x to node y reflects an assertion by the builder of that network that the probability distribution for y may depend on the value assigned to x.

[4] In this paper, we address only the representation of probabilistic-inference problems with belief networks. Nonetheless, an extension of belief networks called *influence diagrams* [57] can represent any decision problem.

We say that the *x conditions y*. For example, in Figure 11, the arcs from the disease node to the feature nodes reflect the expert's belief that the probability of observing a particular value for each feature may depend on the disease that is present. In addition, the arc from CAP THICKENING to FCB reflects the expert's assertion that the probability distribution for FCB may depend on whether or not there is capsule thickening, even when the identity of the disease is known. Conversely, the lack of arcs in a belief network reflect assertions of conditional independence. In Figure 11, there is no arc between CAP THICKENING and FIBROSIS nor is there an arc between FCB and FIBROSIS. The lack of these arcs encode the expert's assertion that FIBROSIS is conditionally independent of CAP THICKENING and FCB, given the identity of the patient's disease.

Each node in a belief network is associated with a set of probability distributions. In particular, a node has a probability distribution for every instance of its conditioning nodes. For example, in Figure 11, FIBROSIS is conditioned by DISEASE. Thus, FIBROSIS has two probability distributions (shown below the belief network in Figure 11): the probability distribution for observing fibrosis given that a patient has NS, and the distribution for observing fibrosis given that a patient has CP. Similarly, CAP THICKENING has two probability distributions. In contrast, FCB is conditioned by both DISEASE and CAP THICKENING. Consequently, this node has four distributions corresponding to the instances where DISEASE is NS or CP, and where CAP THICKENING is T or F. Finally, DISEASE has only one distribution—the prior probability of disease—because it is not conditioned by any nodes.

Given any belief network, we can construct a joint probability distribution for the entire domain of that network. We can build this distribution from the probability distributions associated with each node in the network, and from the assertions of conditional independence reflected by the lack of arcs in the network. For example, as shown in Figure 11, let d, c, b, and f represent the variables DISEASE, CAP THICKENING, FCB, and FIBROSIS, respectively. From the product rule for probabilities [56], we know that the joint probability distribution for these variables is given by

$$p(d,c,b,f) = p(d)\ p(c|d)\ p(b|c,d)\ p(f|b,c,d) \tag{7}$$

In addition, we know that FIBROSIS is conditionally independent of CAP THICKENING and FCB, given DISEASE. That is,

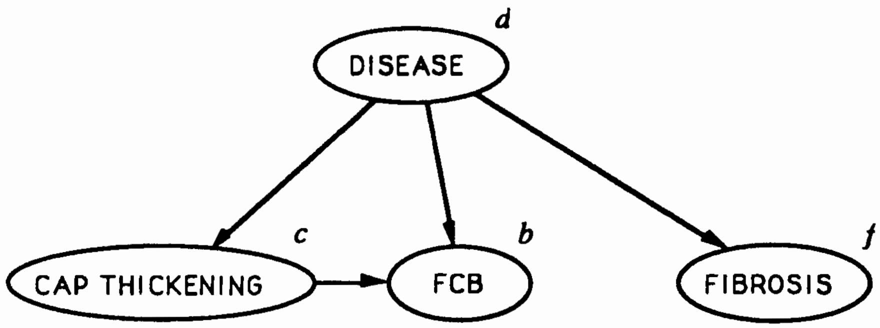
$$p(f|b,c,d) = p(f|d) \tag{8}$$

Combining Equations 7 and 8, we have

$$p(d,c,b,f) = p(d)\ p(c|d)\ p(b|c,d)\ p(f|b) \tag{9}$$

The four sets of probability distributions on the right-hand side of Equation 9 are exactly those distributions associated with the nodes in the belief network.

Thus, using a belief network, knowledge engineers can greatly simplify the capture of probabilistic dependencies, without the need to sacrifice a precise probabilistic representation nor the need to make erroneous assumptions of conditional independence. If an expert believes that—for example—CAP THICKENING and FCB are conditionally independent, he



$$p\ d = \text{NS}) = 0.95$$

$$p(c = \text{T} \mid d = \text{NS}) = 0.9$$
$$p(c = \text{T} \mid d = \text{CP}) = 0.7$$

$$p(f = \text{T} \mid d = \text{NS}) = 0.05$$
$$p(f = \text{T} \mid d = \text{CP}) = 0.00$$

$$p(b = \text{T} \mid c = \text{F}, d = \text{NS}) = 0.05$$
$$p(b = \text{T} \mid c = \text{T}, d = \text{NS}) = 0.35$$
$$p(b = \text{T} \mid c = \text{F}, d = \text{CP}) = 0.00$$
$$p(b = \text{T} \mid c = \text{T}, d = \text{CP}) = 0.00$$

Figure 11: A belief network for the discrimination of the diseases NS (ordinary nodular sclerosis Hodgkin's disease) and CP (cellular phase nodular sclerosis Hodgkin's disease). The features relevant to this diagnostic problem are CAP THICKENING (capsule thickening of 10 or more lymphocyte diameters), FCB (fibrocollagenous bands), and FIBROSIS. The arcs from the disease node to the feature nodes reflect the expert's belief that the likelihood of observing each feature may depend on the disease that has manifested in the lymph node. The arc from CAP THICKENING to FCB represents the expert's assertion that the probability of FCB may depend on whether or not there is capsule thickening, given disease. Conversely, the lack of arcs from CAP THICKENING and FCB to FIBROSIS represent the expert's belief that FIBROSIS is conditionally independent of the other two features, given disease. The probability distributions associated with each node are shown below the belief network (T denotes true or present, and F denotes false or absent).

can represent this dependency explicitly. On the other hand, if he believes that the features are conditionally independent, he can represent this assertion. In either case, we can construct a joint probability distribution for the domain.

Furthermore, builders of expert systems can use belief networks to simplify probabilistic inference—in the case of Pathfinder, the computation of the probability of disease given observations. Researchers have developed several algorithms that exploit the assertions of conditional independence embedded in a belief network for this computation [59,58,60,61]. In Pathfinder, we use a special case of the algorithm described in [60].

9.2.2 Similarity Networks for Focusing Attention

Figure 12 illustrates the complete belief network for Pathfinder. In the figure, we have omitted the arcs from DISEASE to the feature nodes to highlight the conditional dependencies among the features. The belief network is complex. In fact, we were unable to construct this network directly. Instead, we developed a representation, called a *similarity network*, that allowed us to decompose the construction of this belief network into a set of tasks of manageable size [51,62,63].

A similarity network consists of a similarity graph and a collection of local belief networks. A *similarity graph* is an undirected graph whose vertices (nodes) represent the mutually exclusive diseases, and whose edges connect diseases that an expert considers to be similar or difficult to discriminate in practice. Figure 13 shows the similarity graph for Pathfinder. The edge between INTERFOLLICULAR HD (interfollicular Hodgkin's disease) and MIXED CELLULARITY HD (mixed-cellularity Hodgkin's disease), for example, reflects the expert's opinion that these two diseases are often mistaken for each other in practice.

Associated with each edge in a similarity graph is a *local belief network*. The local belief network for an edge is a belief network that contains only those features that are relevant to the differential diagnosis of the two diseases that are connected by that edge. The local belief networks are typically small, because the disease pairs for which they are constructed are similar. For example, the belief network in Figure 11 is the local belief network for the edge between CELLULAR PHASE NSHD (cellular phase nodular sclerosis Hodgkin's disease) and NODULAR SCLEROSING HD (ordinary nodular sclerosis Hodgkin's disease) in the similarity graph. The local belief network contains only the features CAP THICKENING, FCB, and FIBROSIS. Thus, the expert believes that only these features are relevant to the differential diagnosis of these two types of nodular sclerosis Hodgkin's disease.

As another example, the local belief network in Figure 14 is associated with the edge between L&H DIFFUSE HD (lymphocytic and histiocytic diffuse Hodgkin's disease) and L&H NODULAR HD (lymphocytic and histiocytic nodular Hodgkin's disease) in the similarity graph. The local belief network contains only the features L&H NODULES (lymphocytic and histiocytic nodules) and PTGC (progressively transformed germinal centers). The lack of the arc between the two features reflects the expert's assertion that the two features are conditionally independent, given the presence of either of these two diseases.

Given the similarity graph and all its associated local belief networks, we can construct the belief network for the entire domain of lymph-node pathology—called the *global belief network*—with a simple procedure. In particular, we construct the *graph union* of all the

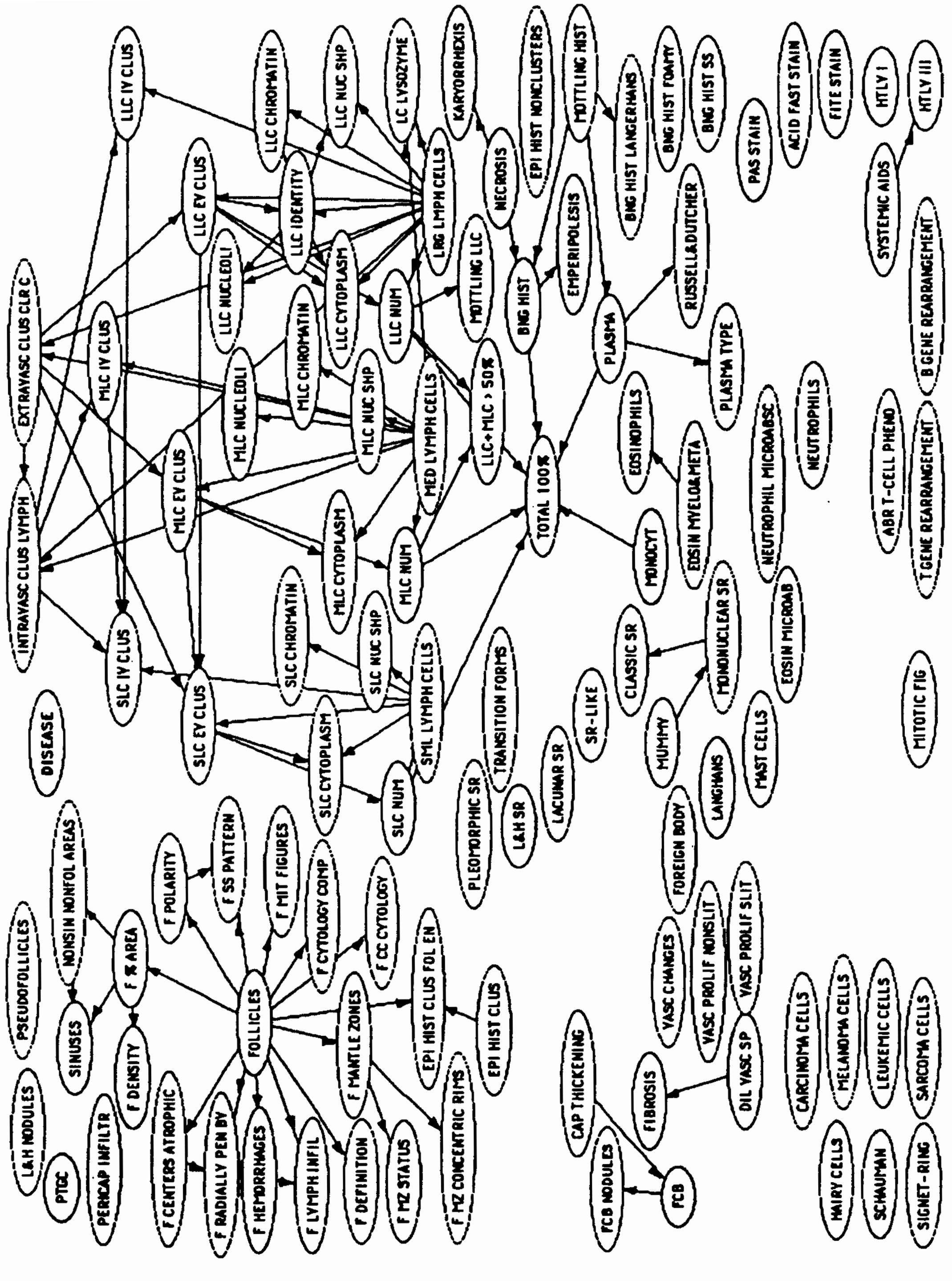


Figure 12: The complete belief network for Pathfinder. The node DISEASE contains over 60 lymph-node diseases. The conditioning arcs from DISEASE to other nodes are not shown so that the conditional dependencies among features are highlighted. The Appendix contains a key to the feature and disease abbreviations.

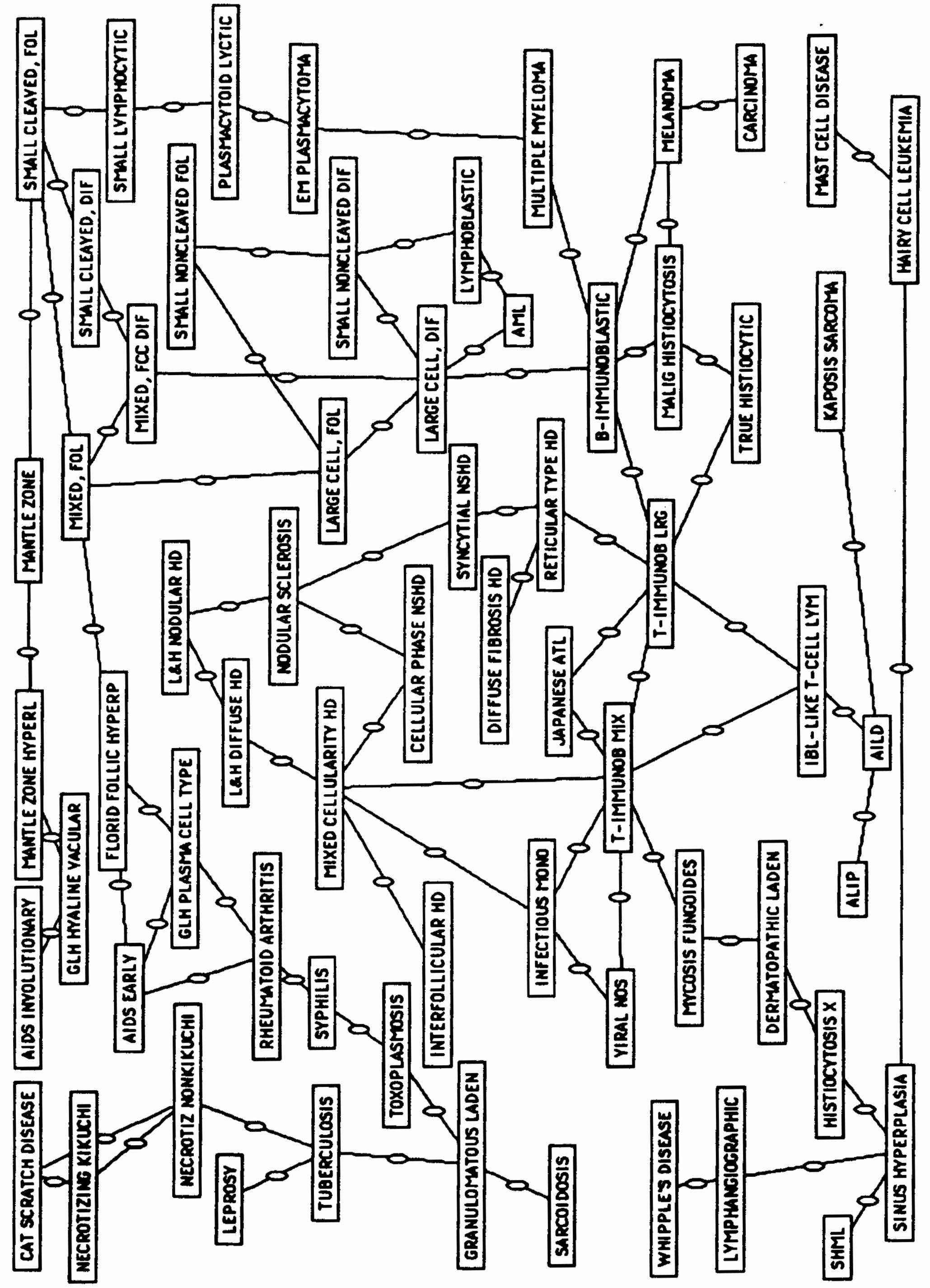


Figure 13: The similarity graph for Pathfinder. The nodes in the graph represent the mutually exclusive diseases that can manifest in a lymph node. Edges connect diseases that the expert considers to be similar.

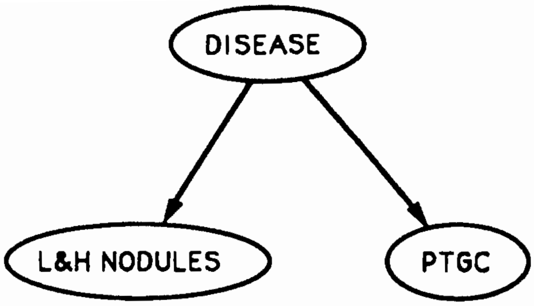


Figure 14: The local belief network for the problem of differentiating L&H DIFFUSE HD (lymphocytic and histiocytic diffuse Hodgkin's disease) and L&H NODULAR HD (lymphocytic and histiocytic nodular Hodgkin's disease).

local belief networks. The operation of graph union is straightforward. The nodes in the graph union of a set of graphs is the simple union of the nodes in the individual graphs. Similarly, the arcs in the graph union of a set of graph is the simple union of the arcs in the individual graphs. That is, a node (or arc) appears in the graph union, if and only if there is such a node (or arc) in at least one of the individual graphs. The belief network in Figure 12 is the belief network formed by this procedure.

Under relatively weak conditions, the global belief network constructed is this manner is *valid* [63]. That is, any joint distribution that satisfies the assertions of conditional independence implied by the local knowledge maps also satisfies the assertions of conditional independence implied by the global knowledge map. We say that the construction of the global knowledge map from the similarity network is *sound*. Thus, the similarity-network representation greatly facilitates the construction of large belief networks. A similarity network allows an expert to decompose the task of building a large belief network into modular and relatively small subtasks. Using a similarity network, an expert can focus his attention on relatively small diagnostic problems that correspond to actual clinical dilemmas.

Several important features of the similarity-network representation are discussed in [62] and [63]. For example, similarity networks can be extended to include local belief networks for sets of hypotheses that contain two or more elements. Essentially, we need only to replace the similarity graph with a similarity hypergraph. (A hypergraph consists of nodes and edges among node sets of arbitrary size.) The representation also can be used in situations where diseases are not mutually exclusive.

A similarity network derives its power from its ability to represent assertions of conditional independence that are not conveniently represented in an ordinary belief network. To illustrate such an assertion, we let variable d represent the mutually exclusive and exhaustive diseases $d_1, d_2, \ldots, d_n$. Further, let $d_{\subseteq}$ denote a proper subset of diseases. If d and feature f are independent, given that one of the elements of $d_{\subseteq}$ is present, we say that f is not relevant to $d_{\subseteq}$. Formally, a feature f ***is not relevant to*** the set $d_{\subseteq}$, if and only if

$$p(d_i|fv, d_{\subseteq}) = p(d_i|d_{\subseteq}) \tag{10}$$

for all values v of variable f, and for all diseases d_i in $d_{\subseteq}$. In Equation 10, the set $d_{\subseteq}$, which conditions both probabilities, denotes the disjunction of its elements. We call the form of conditional independence represented by Equation 10 *subset independence.* Using Bayes' theorem, we can derive an equivalent criterion for subset independence. In particular, we can show that a feature f is not relevant to the set of diseases $d_{\subseteq}$, if and only if

$$p(fv|d_i) = p(fv|d_j) \tag{11}$$

for all pairs $d_i, d_j \in d_{\subseteq}$, and for all values v of feature f.

Assertions of subset independence are *asymmetric.* In general, an assertion of conditional independence is asymmetric if it holds for only some instances of its variables. Assertions of subset independence, in particular, hold for only proper subsets of the disease variable d.

We cannot easily encode subset independence or other forms of asymmetric conditional independence in an ordinary belief network [63]. In contrast, such assertions are represented naturally by local knowledge maps. In particular, if we omit the feature f from the local knowledge map for the diseases d_i and d_j, then we are asserting that f is not relevant to the set $\{d_i, d_j\}$. In the next section, we examine how to exploit subset independence for probability assessment.

9.3 Quantification of Probabilistic Relationships

In Section 9.2.1, we saw that each node in a belief network is associated with a set of probability distributions. In Figure 11 we represented these distributions simply as a table of numbers. We can, however, represent such distributions in a similarity network. For example, consider the feature PTGC (progressively transformed germinal centers). In the global knowledge map (see Figure 12), this feature is conditioned by DISEASE. Thus, we need to assess the probability distribution for PTGC, given each disease. Figure 15 shows how we can represent these assessments, using the Pathfinder similarity graph. In the figure, only the portion of the similarity graph for Hodgkin's diseases is shown. To simplify the presentation, we shall restrict our attention to these diseases in the remainder of this discussion. The rounded rectangle labeled with the feature name contains the mutually exclusive and exhaustive values for the feature: ABSENT and PRESENT. The two numbers under each disease are the probability distribution for the feature given that disease. For example, the probability that PTGC is PRESENT, given L&H NODULAR HD, is 0.1.

As another example, consider the feature CLASSIC SR (classic Sternberg–Reed cells). In the global knowledge for Pathfinder, the node DISEASE and the node MONONUCLEAR SR (mononuclear variants of Sternberg–Reed cells) are the parents of CLASSIC SR. Consequently, we need to assess distributions for CLASSIC SR, given both DISEASE and all possible values of MONONUCLEAR SR. Figure 16 shows how we can encode these assessments, using Pathfinder's similarity graph. Figure 16(a) contains distributions for classic Sternberg–Reed cells, given each Hodgkin's disease, for the case where mononuclear variants of Sternberg–Reed cells are rare (less than three cells in a 4–square–centimeter section of the lymph node). Similarly, Figure 16(b) contains such distributions for the case where mononuclear variants of Sternberg–Reed cells are present (three to 20 cells in a 4–square–centimeter section of the lymph node). These two sets of distributions represent quantitatively the expert's be-

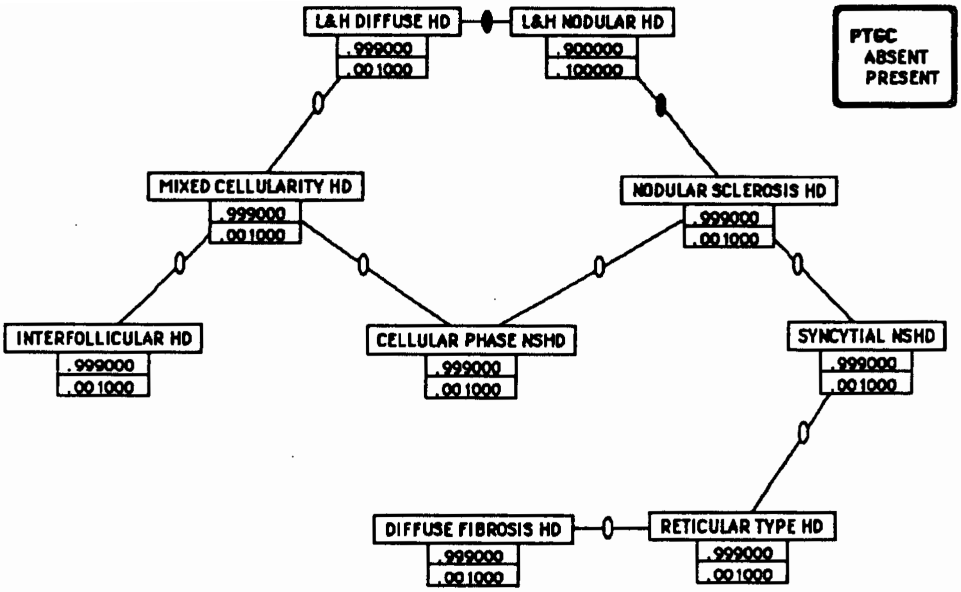


Figure 15: Probability assessment using a similarity network. The probability distributions for the feature PTGC, given the various types of Hodgkin's disease, are shown. The rounded rectangle labeled with the feature name contains the mutually exclusive and exhaustive values for the feature: ABSENT and PRESENT. The numbers below each disease are the probability distribution for PTGC given that disease. For example, the probability that PTGC is PRESENT, given L&H NODULAR HD, is 0.1.

lief that the chances of seeing classic Sternberg–Reed cells are greater when there are more mononuclear variants of Sternberg–Reed cells.

We can use the assertions of subset independence encoded in the similarity network to simplify probability assessment. Let us again consider the assessments for PTGC in Figure 15. A black oval on an edge in the similarity graph reflects that the feature PTGC is present in the local knowledge map corresponding to that edge. Conversely, a white oval on an edge represents that this feature is absent from that local knowledge map. As shown in the figure, when a feature is omitted from a local belief network, the conditional probability distributions on either side of an edge are equal. This observation follows from Equation 11 and from the fact that any feature omitted from a local belief network cannot be relevant to the two diseases associated with that map. Consequently, for the feature PTGC, we need to assess probability distributions given only L&H DIFFUSE HD and L&H NODULAR HD. The remaining distributions must be equal to the distribution for L&H DIFFUSE HD.

A problem with this approach to assessment is illustrated in Figure 17. Specifically, the probability distributions for the feature CAP THICKENING (capsule thickening) given INTERFOLLICULAR HD and DIFFUSE FIBROSIS HD are equal. Because we did not connect these diseases in the similarity graph, however, the equality of these distributions remains hidden until we assess the actual probabilities. We can remedy this difficulty by composing a local belief network for every pair of diseases. For domains such as Pathfinder's that contain many diseases, however, this alternative is impractical. Alternatively, we can compose a *partition* of the diseases for each feature to be assessed. In composing a partition, we place each disease into one and only one set. We place two or more diseases in the same set only if the nondistinguished variable associated with the partition is not relevant to those hypotheses in the set. After composing the partition for a given feature, we assesses probability distributions for the feature, given each disease. Given Equation 11, however, we need to assess only one probability distribution for each set in the partition.

A partition for the feature CAP THICKENING is shown in Figure 18. In this partition, the diseases are divided into four sets: the singleton sets containing NODULAR SCLEROSIS HD, SYNCYTIAL NSHD, and CELLULAR PHASE NSHD, and the set labeled HODGKIN'S that contains the remaining diseases. The partition reflects the assertion that the feature CAP THICKENING is relevant to none of the six diseases in the set HODGKIN'S. That is, if the expert knew that the true disease was in the set HODGKIN'S, then his observation of the status of the lymph-node capsule would not change his relative probabilities of the diseases in that set. Consequently, we need to assess only four probability distributions. These distributions, shown below the sets in Figure 18, are the same as those shown in Figure 17. By using this partition, we uncover equalities among the distributions for CAP THICKENING before we assess probabilities; we thereby avoid the assessment of three distributions.

Using partitions to assess the joint probability distribution for Pathfinder, we decreased the time to assess a belief network by more than a factor of five. At first, this observation may seem surprising, given that a partition must be composed for each conditioning instance of every feature. Two factors contributed to the efficiency of the approach. First, the task of composing a single partition is straightforward. Apparently, as is the case with assertions of symmetric conditional independence, people find it easy to make judgments of subset inde-

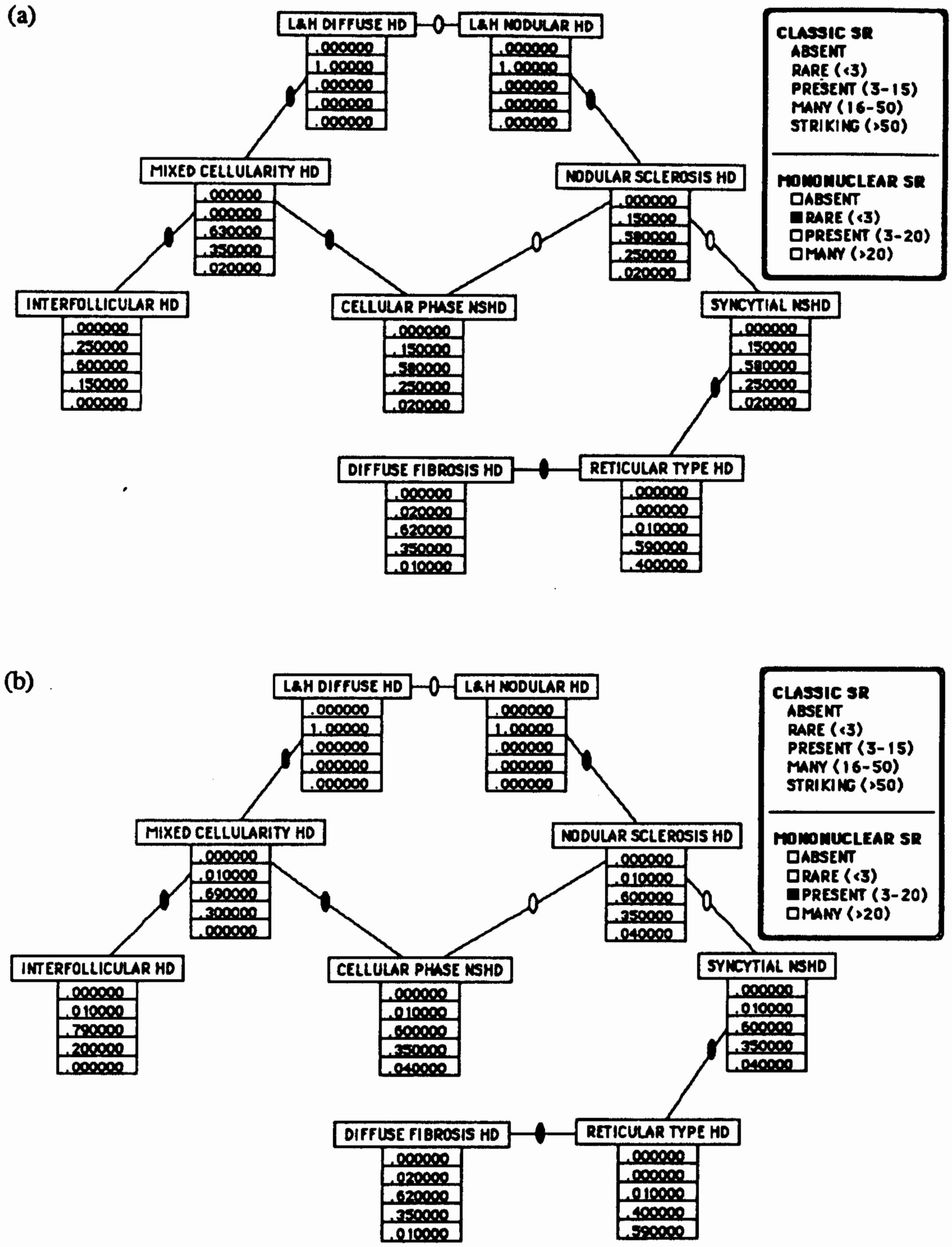


Figure 16: Probability assessment for dependent features. (a) The probability distributions for the feature CLASSIC SR (classic Sternberg–Reed cells), given each Hodgkin's disease, and given that MONONUCLEAR SR (mononuclear variants of Sternberg–Reed cells) are RARE (less than three cells in a 4–square–centimeter section of the lymph node). (b) Similar distributions given that mononuclear variants of Sternberg–Reed cells are PRESENT (three to 20 cells in a 4–square–centimeter section of the lymph node).

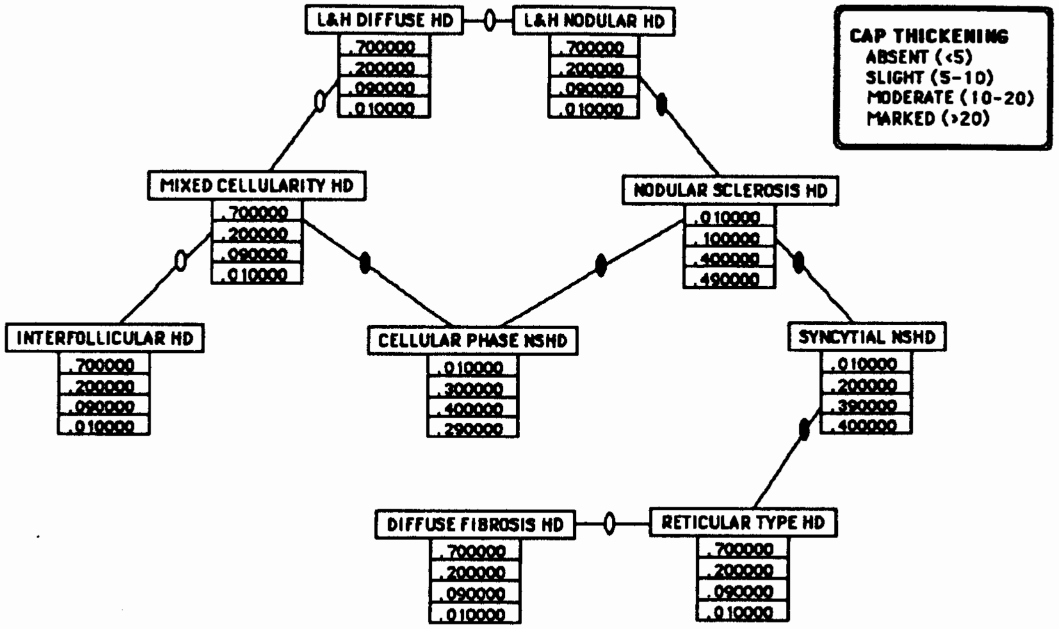


Figure 17: Hidden equivalence in a similarity network. The assessment of the feature CAP THICKENING is shown. The values for the feature are ABSENT (<5 lymphocyte diameters), SLIGHT (5–10 lymphocyte diameters), MODERATE (10–20 lymphocyte diameters), and MARKED (>20 lymphocyte diameters). Although the distributions for INTERFOLLICULAR HD and DIFFUSE FIBROSIS HD are equal, this equality is hidden until the actual assessments are made, because the two diseases are not connected in the similarity graph.

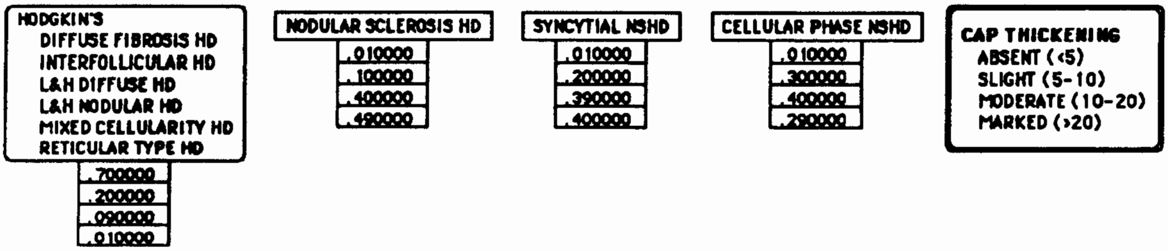


Figure 18: Assessment of probabilities using a partition. The partition contains four sets of diseases consisting of three singleton sets and the set labeled HODGKIN'S. We need to assess only one probability distribution for each set.

pendence without assessing the probabilities underlying such judgments. Second, partitions often are identical or related from one feature to another. In constructing partitions, we can use this close relationship to avoid constructing each partition from scratch.

9.4 A Graphical Knowledge-Acquisition Tool

A crucial step in the development of the probabilistic-dependency model for the lymph-node system was the construction of a computer-based knowledge-acquisition tool, called SimNet, based on the similarity-network representation [63]. All the figures of belief networks and similarity networks shown in this paper were created with SimNet.

The tool allows similarity networks to be constructed and edited on a large bit-mapped display. In practice, an expert first uses the system to create a similarity graph. The expert then selects an edge of interest, and the program automatically sets up a belief-network template (containing only the disease node) from which the expert can construct the local belief network. As the belief networks are created by the expert, the global belief network is constructed automatically. An expert can then use SimNet to assess the probabilities associated with the global belief network. Using the program, the expert can encode probabilities directly in the similarity graph, or in partitions.

10 Decision Theory for Evidence Gathering

One of the major components of the hypothetico-deductive cycle is the identification of features that are cost-effective for narrowing the differential diagnosis. As we discussed in Section 8.2, such features are identified by the computation of value of information. These computations, in turn, require a utility model for diagnosis. In this section, we discuss Pathfinder's utility models, and describe extensions to traditional value-of-information computations.

10.1 A Utility Model for the Patient

The utility U_{d_i,d_j} summarizes the preferences of the decision maker for the situation in which a patient has disease d_i, but is diagnosed as having disease d_j. Factors that influence such preferences include the length of the patient's expected life, the pain associated with treatment and with the disease itself, the psychological trauma to the patient and his family, and the monetary cost associated with treatment and with disability.

The most fundamental question in utility assessment is: Who is the decision maker? From our perspective, a pathologist is only a provider of information. Thus, the U_{d_i,d_j} in the utility model of a computer-based diagnostic system should reflect the patient's preferences. For example, consider the situation where a pathologist believes, after reviewing a case, that the probability of the benign infection mononucleosis is 0.9, and that the probability of Hodgkin's disease is 0.1. Should the patient be treated for Hodgkin's disease now, or should he wait for more definitive diagnostic signs to develop? Delaying treatment of Hodgkin's disease decreases the chances of long-term survival if the patient has this condition. On the other hand, the treatment for Hodgkin's disease is highly invasive and thus carries significant risk. In addition to suffering the acute trauma of chemotherapy and radiotherapy, a patient is likely to become sterile and is put at increased risk for developing other cancers. The decision about therapy will depend on how the *patient* feels about the alternative outcomes. Different patients may have dramatically different preferences.

In practice, pathologists do not acquire detailed knowledge about the preferences of patients for each case. Instead, pathologists make a best guess about the preferences of their patients. Pathologists traditionally assume that their best guess about patient utility will suffice, and that patient-specific variations are not significant enough to warrant the cost of acquiring patient preferences. Indeed, the standards of practice in pathology, and throughout medicine, center on the role of the physician as the ultimate decision maker. The development of efficient techniques for acquiring and representing individual patient preferences could make the use of such knowledge more feasible, and thus more common. Several researchers have investigated the dynamic assessment of patient preferences [64,65,66]. In one approach, researchers encode several prototypical utility models, and use attributes of the patient's personality to choose the most appropriate model for that patient.

For the Pathfinder utility model, we asked our expert to imagine that he was a patient, and to provide the U_{d_i,d_j} accordingly. We hope to extend our current approach with techniques for custom–tailoring utilities to individual patients.

A difficulty in creating the utility model was developing a unit of preference that could be used to measure the utilities associated with both major and minor misdiagnoses. A version of Howard's *worth-numeraire* model [67] provided a solution to this problem. In this model, utilities associated with major misdiagnoses are measured in terms of life-and-death gambles, whereas utilities associated with minor diagnoses are measured in terms of dollars. To measure the utility of a major misdiagnosis, for example, we asked the expert to imagine that he had—say—Hodgkin's disease, and that he had been misdiagnosed as having mononucleosis. We then asked him to imagine that there was a magic pill that would rid him of this disease with probability $1-p$, but would kill him, immediately and painlessly, with probability p. The expert then provided the value of p that made him indifferent between his current situation and the situation in which he takes the pill. To measure the utility of a minor misdiagnosis—say, the diagnosis of cat-scratch disease in a patient with viral lymphadenitis—we simply asked the expert how much he would be willing to pay to be cured if he faced such a misdiagnosis.

The worth-numeraire model of Howard provides a means to convert utilities expressed in monetary terms to small probabilities of immediate, painless death. The lymph-node expert, for example, had a conversion rate of $20 per micromort. A *micromort* is a one–in–1-million chance of immediate, painless death. The conversion makes possible the direct comparison of utilities for minor and major misdiagnoses.

Like many of its predecessors, the model determines what an individual would have to be paid to assume some chance of death, and what he would be willing to pay to avoid a given risk. Also like many of its predecessors, the model shows that, for small risks of death (typically, $p < 0.001$), the amount someone would be willing to pay (or would have to be paid) to avoid (or to assume) such a risk is linear in p. That is, for small risks of death, an individual acts like an expected-value decision maker with a finite value attached to his life. For significant risks of death, however, the model deviates strongly from linearity. For example, the model shows that there is a maximum probability of death, beyond which an individual will accept no amount of money to risk that chance of death. Many people find this result to be intuitive.

10.2 A Utility Model for Research and Education

We designed Pathfinder primarily to help the community pathologist to make decisions about patient cases. Nonetheless, our program would be useful in a research and education setting as well. In these settings, the patient utility model may not be ideal. For example, academic pathologists seek to increase their understanding of the clinical significance of the most subtle disease distinctions. In particular, they are interested in identifying as many subtypes of diseases as possible, in the hope that medical researchers will develop more specific and effective therapies. Thus, academic pathologists are interested in discriminating diseases that currently have the same therapy and prognostic course. In this context, a utility model should treat all distinctions as being equally important; that is, $U_{d_i,d_i} = 1$ and $U_{d_i,d_j} = 0$ for $d_i \neq d_j$. We call this utility model a *discrimination model.* A discrimination model is also appropriate in an educational setting because pathologists in training should be sensitized to all available distinctions. The use of a patient utility model in this setting could obscure useful distinctions. Thus, Pathfinder makes available both the patient utility model and the discrimination model for reasoning about the best observations to make. That is, Pathfinder performs value-of-information computations on each model separately to provide recommendations for evidence gathering.

In the computations based on the discrimination model, we employ an efficient approximation to the computation of value of information [68]. This approximation makes use of a measure of information called *entropy.* The higher the entropy of a differential diagnosis, the greater our uncertainty about the identity of the patient's illness. In this approach, features are ranked by the change in expected entropy of the differential diagnosis that results from their observation. Ben-Bassat has shown theoretically that the value of information of a feature, given a discrimination model, is approximately closely by the expected change in entropy of the differential diagnosis. Other researchers have used this approximation in medical expert systems [28].

10.3 Integration of Heuristic Abstraction

So far, we have discussed evidence-gathering strategies that consider only the most specific disease distinctions available to Pathfinder. We have found that pathologists tend to work at levels of abstraction higher than the level of analysis provided by these detailed decision-theoretic computations [1,69].

We and other researchers have observed that the pathologists' evidence-gathering strategies often can be described by the traversal of disease hierarchies [1,70]. One such disease hierarchy is shown in Figure 19. When using this hierarchy, a pathologist first considers features that discriminate between only benign and malignant diseases. If the differential

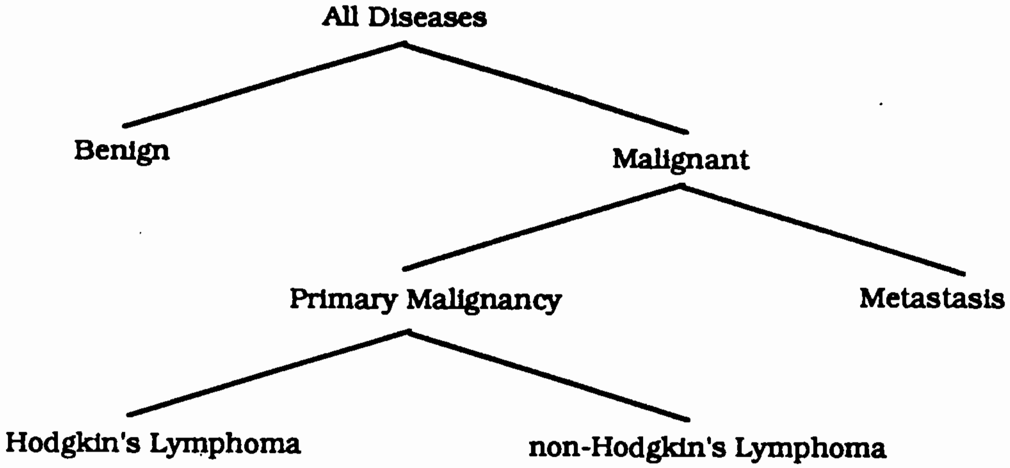


Figure 19: A heuristic problem-solving hierarchy that shows a pathologist categorizes diseases into a sequence of abstraction classes to manage the complexity of diagnostic inference.

diagnosis is narrowed to only malignant diseases, the pathologist then discriminates between primary and metastatic diseases. If metastatic diseases are ruled out, the pathologist considers features that discriminate between non-Hodgkin's and Hodgkin's. Finally, when all diseases under consideration are in the same group, a pathologist discriminates among these diseases individually.

The use of abstraction hierarchies by our domain expert, and other pathologists participating in Pathfinder research, can be viewed as a coarsening of the utility models. Rather than apply the value-of-information calculations to single diagnostic entities, we apply the calculations to groups, considering the probability of each group, and the utility of misdiagnoses among groups of diseases. That is, we allow the pathologist to consider the utility U_{G_m,G_n}, associated with misdiagnosing patients who have a disease in group G_m, instead assigning a disease in group G_n. The utility of misdiagnosis among diseases within each group is considered to be zero.

The utilities U_{G_m,G_n} can be assessed off line and can be stored as grouped utility matrices. Alternatively, during a case analysis, a physician may wish to define groups of hypotheses that reflect his current perspective on the problem. For these situations, we can approximate the utilities dynamically from the utilities U_{d_i,d_j} for single diseases, available in the Pathfinder's ungrouped utility models. For example, we can approximate the utility of misdiagnosis among two groups G_1 and G_2 to be the average of utilities U_{d_i,d_j}, where d_i is an element of G_1, and where d_j is an element of G_2; or, we can take the cost of misdiagnosis between groups G_1 and G_2 to be the maximum utility U_{d_i,d_j}, such that d_i and d_j are elements of G_1 and G_2, respectively.

We have developed a heuristic abstraction facility that allows a user to define such groups of diseases [68]. The current capability is based on a grouped-discrimination model. In this model, we set $U_{G_m,G_m} = 0$ for all G_m, and we set $U_{G_i,G_j} = U_{G_m,G_n}$ for all $G_i \neq G_j$ and $G_m \neq G_n$. Multiple windows—each representing a different perspective on the same problem—may be displayed simultaneously. By clicking on one of the windows, a user activates the perspective. He then can ask the system to generate a recommendation about the best way to discriminate among the leading disease classes represented in the window.

By using evidence-gathering strategies based on grouped utility models, we can introduce human-oriented flexibility to recommendation generation. We were not the first investigators to discover such abstraction in clinical problem solving. Cognitive psychology studies have found that clinicians, in a variety of specialty areas of medicine, frequently make use of abstraction [15,16]. We join the cognitive psychologists in conjecturing that decomposing the task of diagnosis into familiar discriminatory subproblems may be useful for managing the complexity of clinical problem solving for humans.

11 Explanation of Pathfinder Recommendations

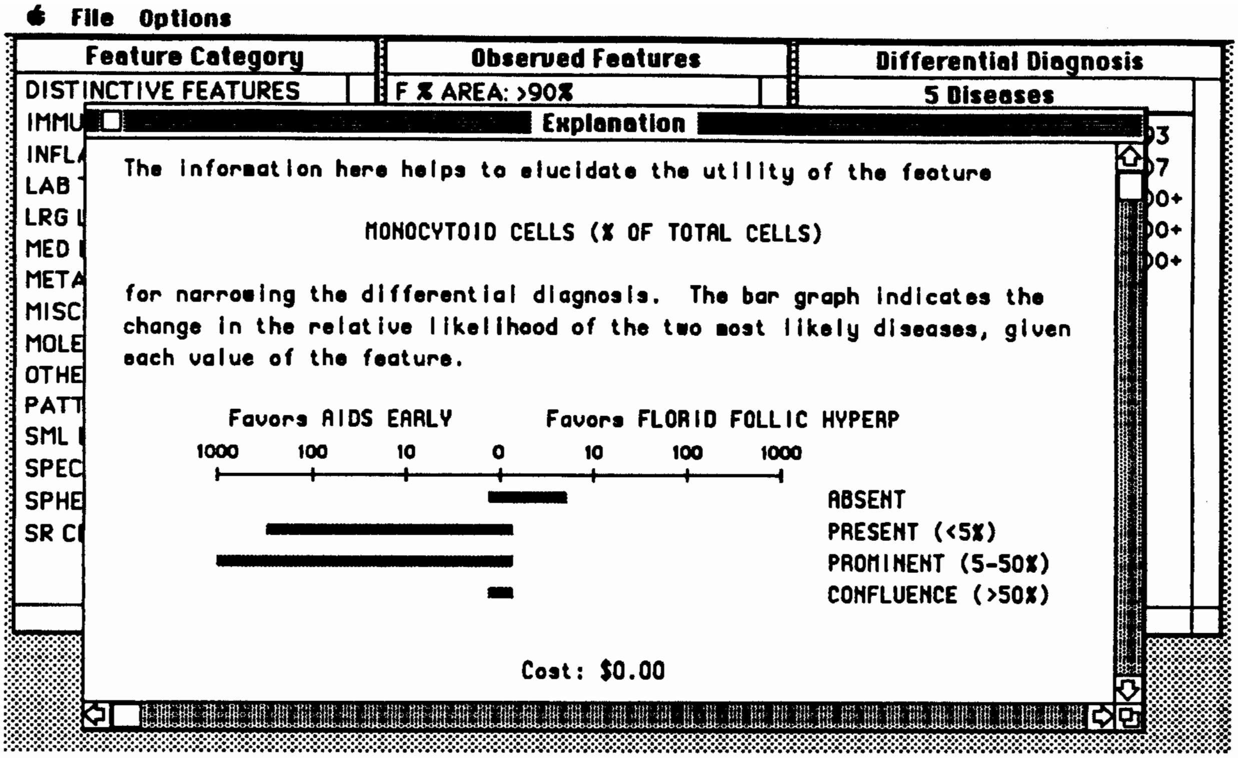
Most medical-informatics researchers agree that the comprehension of automated reasoning by users is an important factor in the acceptance of advice from expert systems [71,72]. Several researchers have criticized probabilistic reasoning systems, saying that the advice they generate is difficult to explain [29,73]. We believe that the difficulty of explaining decision-theoretic inference is related to the inescapable complexity of normative analysis. Our approach to explanation is to trade off the opacity of a complete explanation with the transparency of simpler incomplete summaries of the discriminatory power of a feature.

Recall that there are two determinants of whether or not a feature is useful for observation: the value of information of observing that feature, and the cost of observing that feature. The latter component is easy to display to the user. For example, in one version of Pathfinder, we display the monetary cost associated with observing the feature (see the bottom of Figure 20). In other versions, we display other components of cost, including estimates of the time it takes to observe a feature, and the degree of tedium associated with such a task.

Communicating to the user the details of value-of-information computations, however, is more difficult. In principle, the program should display the effect of entering each value of a feature on every disease on the differential diagnosis, as well as the likelihood that each value will be observed, given the differential diagnosis. In Pathfinder, however, we show the effect of entering each value of a feature on only two groups of diseases in the differential diagnosis. In particular, for two groups of diseases G_1 and G_2, we generate a set of likelihood ratios for each possible value v of the feature

$$\frac{p(fv|G_1)}{p(fv|G_2)}$$

where $p(fv|G_i)$, $i = 1, 2$, denotes the probability that fv is observed given that the true disease state of the patient lies in G_i. We then display graphically the logarithm of each likelihood ratio, known as the *weight of evidence* of the feature–value pair fv in favor of G_1 relative to G_2 [21]. Several other probability-based expert-system projects have used likelihood ratios and weights of evidence to explain the relevance of evidence to hypotheses under consideration [41,40,42,43].



The information here helps to elucidate the utility of the feature

MONOCYTOID CELLS (% OF TOTAL CELLS)

for narrowing the differential diagnosis. The bar graph indicates the change in the relative likelihood of the two most likely diseases, given each value of the feature.

Cost: $0.00

Figure 20: A graphical justification for the recommendation of MONOCYTOID CELLS. For each value of the feature, the length and direction of a bar reflects the change in the probability of AIDS EARLY relative to the change in the probability of FLORID FOLLIC HYPERP, given the observation of that feature–value pair. The justification also includes the monetary cost of observing the feature.

We experimented with qualitative text and graphical displays of weights of evidence. We found a graphical approach to be the most effective. A bit map of Pathfinder's graphical justification of the diagnostic utility of the feature monocytoid cells is displayed in Figure 20. In the figure, G_1 is the singleton set that corresponds to AIDS EARLY (the early phase of AIDS); G_2 is the singleton set that corresponds to the benign disease FLORID FOLLIC HYPERP (florid reactive follicular hyperplasia). For each value of the feature, a bar is displayed that extends toward either AIDS EARLY or FLORID FOLLIC HYPERP. For a given value of the feature, the length of the bar is proportional to the absolute value of the weight of evidence of that feature–value in favor of AIDS EARLY relative to FLORID FOLLIC HYPERP; if the bar extends toward AIDS EARLY, then the weight of evidence is positive, and that feature–value favors AIDS EARLY; conversely, if the bar extends toward FLORID FOLLIC HYPERP, then the weight of evidence is negative, and that feature–value favors FLORID FOLLIC HYPERP. By glancing at the graph, we can determine that the absence of monocytoid cells favors FLORID FOLLIC HYPERP, whereas the presence or prominence of these cells favors AIDS EARLY. These facts suggest immediately that the feature MONOCYTOID CELLS is a useful feature to evaluate, as the cost of observing this feature is negligible.

Part III
System Evaluation

12 Previous Evaluations

Medical expert systems can be evaluated according to their performance across several different dimensions. For example, we can evaluate the diagnostic accuracy of an expert system, the effect that use of such a system can have on the accuracy of diagnoses rendered in practice, or the value to the physician of making such a system available for use in a clinical setting. Several investigators have discussed the problems with determining the value of expert systems in clinical decision making [71,72,74,34]. To date, most evaluations have not attempted to characterize the frequency with which physicians will use an available expert systems and the difference in diagnostic performance between an unassisted and assisted clinician. Instead, studies have examined the diagnostic accuracy of expert systems relative to the performance of physicians or to gold-standard diagnoses.

Investigators have shown that simple probabilistic systems can perform at a level comparable to that of experts, and sometimes at a considerably higher level [28,75,76]. The system of deDombal averaged greater than 90 percent correct diagnoses of acute abdominal pain; expert physicians averaged only 65 to 80 percent correct with the same cases [75]. Researchers interested in the efficacy of nonprobabilistic approaches have undertaken several formal and informal evaluations of the performance of rule-based and quasiprobabilistic systems. Evaluation of the rule-based Mycin system [77] studied the performance of the system against an expert gold-standard [34]. The evaluation showed that Mycin's recommendations were approved by a majority of experts in approximately 75% of cases. Investigators also studied the usefulness of the rule-based Oncocin system [78]—a program for managing oncology protocols—for improving data collection [79]. Researchers analyzed the completeness of clinical trial data before and after Oncocin was introduced into a clinical setting. Data completeness was increased significantly with the introduction of the system. The study showed that data recorded about observations and test ordering are significantly improved with use of the system.

13 Evaluation of Pathfinder's Diagnostic Accuracy

Over the last year, we performed an evaluation of Pathfinder that compared the diagnostic accuracy of inference with a knowledge base containing conditionally independent features, and inference with a more sophisticated dependency model. Most significant in our evaluation is the development of a novel combination of ad hoc and decision-theoretic metrics for exploring the value of different systems. The Pathfinder team is planning to perform a similar study to analyze the value of making available normative expert systems to community pathologists.

In our evaluation, we compared the accuracy of the current version of Pathfinder, in which probabilistic dependencies are represented, with that of an earlier version of the program, in which all features are assumed to be conditionally independent [63]. In the evaluation study, 53 cases were selected in sequence from a large library of referrals. For each case, a community pathologist reported salient morphologic features to both versions of Pathfinder. Often, the pathologist reported (to both systems) additional features that were recommended for evaluation by one or both systems. For each case, the pathologist entered features until she believed that no additional observations were relevant to that case. In most instances, she stopped entering features when neither program had features to recommend. For several cases, however, although one or both systems recommended features, she did not observe any of these features, because she believed that they would not have a significant effect on the differential diagnosis. For several other cases, where neither system had features to recommend, she identified features on her own that she though might be relevant to the case, and entered those features.

We gauged the diagnostic accuracy of the probability distributions produced by Pathfinder by assigning the distributions a score based on two metrics: an expert-rating metric, and a formal decision-theoretic metric. The two approaches were complementary in their ability to identify components of the system that affected diagnostic accuracy [37].

13.1 Expert-Rating Approach

In the expert-rating approach, the lymph-node expert was asked to rate directly, on a subjective scale, the quality of the probability distributions produced by each version of Pathfinder. For each case, he was shown the features reported by the nonexpert, as well as the probability distributions produced by the two versions of the system. The expert was blinded as to the identity of the distributions, and the distributions were displayed in random order. For each probability distribution, the expert was asked, "On a scale from zero to ten—zero being unacceptable and ten being perfect—how accurately does the distribution reflect your beliefs?" The 0–to–10 scale provided an informal measure of the difference between the diagnostic accuracy of the two probabilistic models.

The expert-rating evaluation metric is useful because it is easy to apply, and because it readily exposes differences between probability distributions. Unfortunately, inferring the *importance* of differences with this metric is difficult. It is impossible to deduce from the expert-rating method, for example, whether or not the additional effort to represent dependencies among features was cost effective.

13.2 Decision-Theoretic Approach

We used the utility model described in Section 10 to measure the significance of differences in the probability distributions produced by the two versions of Pathfinder. For both versions of Pathfinder, we computed a quantity called *inferential loss*. The inferential loss reflects the difference in expected utility between the Pathfinder's diagnosis and diagnosis that we obtain from the *gold-standard* probability distribution associated with a case. Several researchers have suggested that similar decision-theoretic metrics be used to evaluate diagnostic computer systems [80,81,82]; some investigators have actually employed such metrics in evaluations [83,84,85].

The procedure for computing inferential loss for a given case is identical for both versions of Pathfinder. We present a generic analysis here. First, the expert examines the observations reported by the community pathologist for a given case. Based on these features, the expert assesses a probability distribution over the diseases—the gold-standard distribution. Based on these same features, Pathfinder also produces a probability distribution over diseases. Let ϕ denote these observations, and let $p_{\mathrm{pf}}(d_i|\phi)$ and $p_{\mathrm{gs}}(d_i|\phi)$ represent the probability of the ith disease under the Pathfinder distribution (either version) and gold-standard probability distribution, respectively.

Then, we determine the optimal diagnosis associated with the gold-standard distribution, denoted $dx_{\mathrm{gs}}(\phi)$, by identifying the diagnosis that maximizes the expected utility of the patient given the distribution. Similarly, we determine the optimal diagnosis associated with the Pathfinder distribution, denoted $dx_{\mathrm{gs}}(\phi)$. Formally, we compute

$$dx_{\mathrm{gs}}(\phi) = \mathrm{argmax}_{d_j} \left[\sum_{d_i} p_{\mathrm{gs}}(d_i|\phi)\ U_{d_i,d_j} \right]$$

$$dx_{\mathrm{pf}}(\phi) = \mathrm{argmax}_{d_j} \left[\sum_{d_i} p_{\mathrm{pf}}(d_i|\phi)\ U_{d_i,d_j} \right]$$

Next, we compute the expected utilities of both diagnoses. When computing expected utility, we use the gold-standard distribution, which reflects the assumed best distribution. That is, we compute

$$EU(dx_{\mathrm{gs}}(\phi)|\phi) = \sum_{d_i} p_{\mathrm{gs}}(d_i|\phi)\ U_{d_i,dx_{\mathrm{gs}}(\phi)}$$

$$EU(dx_{\mathrm{pf}}(\phi)|\phi) = \sum_{d_i} p_{\mathrm{gs}}(d_i|\phi)\ U_{d_i,dx_{\mathrm{pf}}(\phi)}$$

Finally, we determine inferential loss, denoted IL, for the Pathfinder distribution. We obtain

$$\mathrm{IL} = EU(dx_{\mathrm{gs}}(\phi)|\phi) - EU(dx_{\mathrm{pf}}(\phi)|\phi)$$

By construction, IL is always a nonnegative quantity. If both the Pathfinder and gold-standard distributions imply the same diagnosis, then the inferential loss is zero, a perfect score. Note that the units of inferential loss are micromorts—the same as those for the diagnostic utilities U_{d_i,d_j}.

13.3 Results of the Evaluation

The expert-rating and decision-theoretic scores for the two versions of Pathfinder are shown in Tables 1 and 2, respectively. Although the standard deviations are wide, both metrics show a significant difference using a bootstrap permutation test [86] (achieved significance level (ASL) of 0.007 for the expert-rating scores, and ASL of 0.07 for the decision-theoretic scores). The difference of 0.95 between the averages of the expert-rating scores does not carry much meaning. However, the difference in inference loss of approximately 300 micromorts

	Expert Ratings (0–10)	
Knowledge Base	mean	sd
Independence KB	7.99	2.32
Dependence KB	8.94	1.51

Table 1: Expert-rating scores comparing the diagnostic accuracy of the conditional-independence knowledge base (Independence KB) with that of the knowledge base containing dependency information (Dependency KB).

	Inferential Loss (micromorts)	
Knowledge Base	mean	sd
Independence KB	340	1684
Dependence KB	16	104

Table 2: Decision-theoretic scores comparing the diagnostic accuracy of the conditional-independence knowledge base (Independence KB) with that of the knowledge-base containing dependency information (Dependency KB).

has a clear interpretation. Assuming that a patient is willing to convert micromorts to dollars at a rate of $20 per micromort, as our expert was, the results in this metric show that it is worth approximately $6000 *per case* to the patient to have the more sophisticated Pathfinder knowledge be used instead of the earlier knowledge base that assumed global independence among features.

Although the decision-theoretic metric is superior to the expert-rating method in terms of the clarity it affords, the expert-rating method has its advantages as well. For example, measurements of inferential loss may be extremely variable from patient to patient because the diagnostic utilities may similarly vary. Also, the expert ratings tend to be more sensitive to differences in diagnostic accuracy. This is not surprising, because experts tend to be hypersensitive to errors in diagnosis, whether such errors matter to a decision maker or not, because the integrity of the expert is at stake. Of course, the decision-theoretic metric can be modified to be more sensitive. Considerations of integrity or liability, for example, can be incorporated into the diagnostic utilities. Indeed, the fact that components of preference can be made explicit and are under the direct control of the expert is one advantage of the decision-theoretic approach. Finally, the expert-rating metric is easier to apply than the decision-theoretic metric. It took approximately 60 hours of our expert's time to develop the utility model used in this evaluation. In contrast, it took the expert less than 1 minute per case to rate the distributions using the heuristic score. Overall, the two approaches are complementary.

14 Problem of Feature Identification

A pathologist can misidentify features or can fail to recognize important features. The decision-theoretic metric allowed us to measure the importance of such errors. Our experiment was straightforward. After we computed the results described in the previous section, the expert reviewed the tissue sections directly and provided a probability distribution over diseases based on the features he observed. This distribution will be called the *true* probability distribution. To measure the importance of feature identification, we compared the gold-standard distribution to the true distribution using the decision-theoretic computations similar to those described in Section 13.2. The scores derived in this manner reflected the significance of feature identification because the only difference between the true and gold-standard distributions was that the former is generated by the expert while he was looking at the features observed by the community pathologist, whereas the latter was generated by the expert while he was looking directly at the tissue sections.

The results of this experiment are startling. The scores for the gold-standard distributions relative to the true distributions average approximately 8,000 micromorts. This observation means that a patient is taking on an additional 8 in 1/1000 chance of death *per case*, if the community pathologist rather than our expert identifies features. Our community pathologist was a former fellow of our expert, and thus was more likely to be proficient at feature identification in the lymph node domain. Therefore, this approach to evaluation strongly suggests that future research help pathologists to recognize features. Of course, a similar experiment should be conducted to quantify the differences in diagnoses rendered by expert pathologists in order for us to judge the true significance of these results.

Part IV
Epilogue

15 Future Pathfinder Research

Pathology diagnosis depends on the accurate recognition of histologic features, as well as on coherent reasoning under uncertainty. We plan to undertake detailed clinical trials to determine the relative contribution of the these two tasks to inaccurate diagnosis. In particular, we shall analyze the gains expected from the computer-based construction of differential diagnoses and generation of recommendations about optimal diagnostic strategies. We shall also study the improvements in feature identification that come from the use of a videolibrary of features stored on a random-access videodisc. The evaluation of Pathfinder in the clinical setting will also focus on sociological factors. For example, it is important to understand how physicians' habits will be modified by the introduction of such a system into surgical-pathology practices. We shall seek also to understand the affect that such systems might have on patterns of case referral.

Other continuing Pathfinder research, in collaboration with investigators at Carnegie-Mellon University, involves the study of computer-based vision techniques for the automated identification of features on histopathology sections. This work has already yielded preliminary techniques for recognizing important features on lymph-node sections. The work also has identified a set of *subcognitive* features—features that are not traditionally used in pathology diagnosis, yet are well correlated with diseases and prognostic course. We are also investigating the feasibility of integrating automated microscopy with expert-reasoning systems. Such integrated systems might one day work with human experts as colleagues in feature recognition and diagnosis.

16 Summary and Conclusions

We reviewed 6 years of Pathfinder research on building expert systems that are founded on the principles of probability and decision theory. After describing Pathfinder's behavior, we introduced the hypothetico-deductive paradigm for diagnosis. We then described the axiomatic bases of probability and utility, and discussed the notion of probability as a measure of personal belief. We reviewed different paradigms for reasoning under uncertainty that have been pursued by medical-informatics investigators. We then introduced tractable methods for acquiring, representing, computing, and explaining decision-theoretic knowledge. We presented the belief network as a formal foundation for the representation of uncertain knowledge, and discussed several enhancements to the representation that make it an intuitive and tractable representation for large knowledge bases. We described how a graphical knowledge-acquisition system, based on the similarity-network representation, could speed up the knowledge acquisition process, making feasible large, probability-based knowledge bases. We then described our assessment of a patient utility model, and presented techniques for introducing flexibility to normative inference. This flexibility is achieved by representing human-oriented disease abstractions that enable users to perform discriminatory inference at arbitrary levels of abstraction. We discussed our work on the explanation of complex decision-theoretic computation. Finally, we reviewed our evaluation of Pathfinder. The evaluation used both an expert-rating metric and a decision-theoretic metric to test the diagnostic accuracy of our expert system. We used these metrics to compare a sophisticated dependency model with a simple model that embodies the assumption of global conditional independence among features. Finally, we described our plans to undertake more detailed clinical trials and to continue to study the automated recognition of features.

Pathfinder research has demonstrated that a decision-theoretic representation is sufficiently tractable and expressive to capture the important knowledge in the domain of lymph-node pathology. We hope that our experiences will inspire other medical-informatics investigators to develop normative expert systems for medicine.

Acknowledgments

We thank Greg Cooper, Lawrence Fagan, Thomas Lincoln, Randolph Miller, Ramesh Patil, Edward Shortliffe, and Clive Taylor for advice and guidance on Pathfinder research. We also thank Costan Berard, Jerome Burke, and Ronald Dorfman for serving with Bharat Nathwani as domain experts in lymph-node pathology. We are indebted to Doyen Nguyen, who served as the community pathologist in the Pathfinder evaluation. Henri Suermondt, Mark Fischinger, Marty Chavez, and especially Keung-Chi Ng assisted with programming and data management. Lyn Dupre, Greg Cooper, and Keung-Chi Ng provided useful comments on this manuscript.

Appendix: Glossary of Terms

Diseases of the Lymph Node

AIDS EARLY: AIDS, early phase
AIDS INVOLUTIONARY: AIDS, involutionary phase
AILD: Angio-immunoblastic lymphadenopathy
ALIP: Atypical lymphoplasmacytic and immunoblastic proliferation
AML: Acute myeloid leukemia
B-IMMUNOBLASTIC: Immunoblastic plasmacytoid diffuse lymphoma
CARCINOMA: Carcinoma
CAT SCRATCH DISEASE: Cat-scratch disease
CELLULAR PHASE NSHD: Cellular phase of nodular sclerosis Hodgkin's disease
DERMATOPATHIC LADEN: Dermatopathic lymphadenitis
DIFFUSE FIBROSIS HD: Diffuse fibrosis Hodgkin's disease
EM PLASMACYTOMA: Extramedullary plasmacytoma
FLORID FOLLIC HYPERP: Florid reactive follicular hyperperplasia
GLH HYALINE VACULAR: Giant lymph-node hyperplasia, hyaline vacular type
GLH PLASMA CELL TYPE: Giant lymph-node hyperplasia, plasma-cell type
GRANULOMATOUS LADEN: Granulomatous lymphadenitis
HAIRY CELL LEUKEMIA: Hairy cell leukemia
HISTIOCYTOSIS X: Histiocytosis x
IBL-LIKE T-CELL LYM: Immunoblastic lymphadenopathy-like T-cell lymphoma
INFECTIOUS MONO: Infectious mononucleosis
INTERFOLLICULAR HD: Interfollicular Hodgkin's disease
JAPANESE ATL: Japanese adult T-cell lymphoma
KAPOSIS SARCOMA: Kaposis sarcoma
L&H DIFFUSE HD: Lymphocytic and histiocytic diffuse Hodgkin's disease
L&H NODULAR HD: Lymphocytic and histiocytic nodular Hodgkin's disease
LARGE CELL, DIF: Large cell diffuse lymphoma
LARGE CELL, FOL: Large cell follicular lymphoma
LEPROSY: Leprosy
LYMPHANGIOGRAPHIC: Lymphangiography effect
LYMPHOBLASTIC: Lymphoblastic lymphoma
MALIG HISTIOCYTOSIS: Malignant histiocytosis
MANTLE ZONE: Mantle-zone lymphoma
MANTLE ZONE HYPERL: Mantle-zone hyperplasia
MAST CELL DISEASE: Mast-cell disease
MELANOMA: Melanoma
MIXED CELLULARITY HD: Mixed-cellularity Hodgkin's disease
MIXED, FCC DIF: Mixed (follicular center cell type) diffuse lymphoma
MIXED, FOL: Mixed (follicular center cell type) follicular lymphoma
MULTIPLE MYELOMA: Multiple myeloma
MYCOSIS FUNGOIDES: Mycosis fungoides

NECROTIZ NONKIKUCHI: Non-Kikuchi's necrotizing lymphadenitis

NECROTIZING KIKUCHI: Kikuchi's necrotizing lymphadenitis

NODULAR SCLEROSIS HD: Nodular sclerosis Hodgkin's disease

PLASMACYTOID LYCTIC: Small lymphocytic diffuse lymphoma with plasmacytoid features

RETICULAR TYPE HD: Reticular type Hodgkin's disease

RHEUMATOID ARTHRITIS: Rheumatoid arthritis

SARCOIDOSIS: Sarcoidosis

SHML: Sinus histiocytosis with massive lymphadenopathy

SINUS HYPERPLASIA: Sinus hyperplasia

SMALL CLEAVED, DIF: Small cleaved diffuse lymphoma

SMALL CLEAVED, FOL: Small cleaved follicular lymphoma

SMALL LYMPHOCYTIC: Small lymphocytic lymphoma

SMALL NONCLEAVED DIF: Small noncleaved diffuse lymphoma

SMALL NONCLEAVED FOL: Small noncleaved follicular lymphoma

SYNCYTIAL NSHD: Syncytial nodular sclerosis Hodgkin's disease

SYPHILIS: Syphilis

T-IMMUNOB LRG: Peripheral T-cell lymphoma, large-cell type

T-IMMUNOB MIX: Peripheral T-cell lymphoma, mixed-cell type

TOXOPLASMOSIS: Toxoplasmosis

TRUE HISTIOCYTIC: True histiocytic lymphoma

TUBERCULOSIS: Tuberculosis

VIRAL NOS: Viral lymphadenitis, not otherwise specified

WHIPPLE'S DISEASE: Whipple's disease

Features of the Lymph Node

ABR T-CELL PHENO: Abberrant T-cell phenotype in medium-sized or large lymphoid cells

ACID FAST STAIN: Acid fast stain

B GENE REARRANGEMENT: Immunoglobulin gene rearrangement

BNG HIST: Benign histiocytes not otherwise specified in the nonfollicular areas

BNG HIST FOAMY: Foamy benign histiocytes in the nonfollicular areas that do not contribute to mottling

BNG HIST LANGERHANS: Langerhans benign histiocytes in the nonfollicular areas

BNG HIST SS: Starry-sky benign histiocytes in the nonfollicular areas

CAP THICKENING: Capsule thickening (number of lymphocytes thick)

CARCINOMA CELLS: Carcinoma cells

CLASSIC SR: Classic Sternberg–Reed cells (number per 4-square-centimeter

DIL VASC SP: Vascular spaces dilated by red blood cells

EMPERIPOLESIS: Number of histiocytes showing emperipolesis

EOSIN MICROAB: Eosinophil microabscessess

EOSIN MYELO&META: Eosinophilic myelocytes and metamyelocytes

EOSINOPHILS: Eosinophils (not in microabcesses)

EPI HIST CLUS: Epithelioid histiocyte clusters

EPI HIST CLUS FOL EN: Epitheliod histiocyte clusters encroaching and/or within follicles

EPI HIST NONCLUSTERS: Epitheliod histiocyte nonclusters (percent of total cell population)

EXTRAVASC CLUS CLR C: Extravascular clusters of clear lymphoid cells

F % AREA: Percent area occupied by follicles

F CC CYTOLOGY: Cytology of follicular center cells in most follicles

F CENTERS ATROPHIC: Atrophic centers in any follicles

F CYTOLOGY COMP: Similar cells inside and outside of most follicles

F DEFINITION: Definition of follicles

F DENSITY: Follicle density

F HEMORRHAGES: Hemmorrhages in any of the follicles

F LYMPH INFIL: Lymphocyte infiltration of any follicles

F MANTLE ZONES: Follicle mantle zones in any follicles

F MIT FIGURES: Follicle mitotic figures in 10 high-power fields

F MZ CONCENTRIC RIMS: Mantle zone concentric rims in any follicles

F MZ STATUS: Follicle mantle zones

F POLARITY: Prominent polarity in any follicle

F RADIALLY PEN BV: Number of follicles showing radially penetrating blood vessels

F SS PATTERN: Follicle starry-sky histiocytes (average number in one 10X objective power)

FCB: Fibrocollagenous bands or sclerosis

FCB NODULES: Nodules formed by fibrocollagenous bands

FIBROSIS: Fibrosis

FITE STAIN: Fite stain

FOLLICLES: Follicles

FOREIGN BODY: Foreign body (number in 4-square-centimeter section)

HAIRY CELLS: Hairy cells

HTLV I: HTLV I antibody test

HTLV III: HTLV III antibody test

INTRAVASC CLUS LYMPH: Intravascular clusters of lymphoid cells

KARYORRHEXIS: Karyorrhexis

L&H NODULES: Lymphocytic and hitiocytic nodules

L&H SR: Lymphocytic and hitiocytic variants of Sternberg–Reed cells (number in 4-square-centimeter section)

LACUNAR SR: Lacunar variants of Sternberg–Reed cells (number in 4-square-centimeter section)

LANGHANS: Langhans cells (number in $4cm^2$ section)

LC LYSOZYME: Lysozyme positivity in medium-sized and/or large lymphoid cells

LEUKEMIC CELLS: Leukemic cells

LLC CHROMATIN: Chromatin of most large lymphoid cells

LLC CYTOPLASM: Cytoplasm of most large lymphoid cells

LLC EV CLUS: Large lymphoid cells in extravascular clusters of clear cells

LLC IDENTITY: Identity of most large lymphoid cells

LLC IV CLUS: Large lymphoid cells in intravascular clusters

LLC NUC SHP: Nuclear shape of most large lymphoid cells

LLC NUCLEOLI: Nucleolar features of most large lymphoid cells

LLC NUM: Number of large lymphoid cells in the nonfollicular areas (percent of total cell population)

LLC+MLC > 50%: Number of medium-sized and large lymphoid cells in the nonfollicular areas exceeds 50 percent of total cell population

LRG LMPH CELLS: Large lymphoid cells

MAST CELLS: Mast cells (number in 4cm^2 section)

MED LYMPH CELLS: Medium-sized lymphoid cells

MELANOMA CELLS: Melanoma cells

MITOTIC FIG: Mitotic figures in 10 high-power fields (nonfollicular areas)

MLC CHROMATIN: Chromatin structure of most medium-sized lymphoid cells

MLC CYTOPLASM: Cytoplasm of most medium-sized lymphoid cells

MLC EV CLUS: Medium-sized lymphoid cells in extravascular clusters of clear cells

MLC IV CLUS: Medium-sized lymphoid cells in intravascular clusters

MLC NUC SHP: Nuclear shape of most medium-sized lymphoid cells

MLC NUCLEOLI: Nucleolar features of most medium-sized lymphoid cells

MLC NUM: Number of Medium-sized lymphoid cells in the nonfollicular areas (percent of total cell population)

MONOCYT: Monocytoid cells (percent of total cell population)

MONONUCLEAR SR: Mononuclear variants of Sternberg–Reed cells (number in 4-square-centimeter section)

MOTTLING HIST: Mottling by langerhans or other histiocytes

MOTTLING LLC: Mottling by large lymphoid cells

MUMMY: Large mummified cells (number in 4-square-centimeter section)

NECROSIS: Necrosis

NEUTROPHIL MICROABSC: Neutrophil microabcessess

NEUTROPHILS: Neutrophils (not in microabcesses)

NONSIN NONFOL AREAS: Nonsinus nonfollicular areas

PAS STAIN: Strong PAS positivity in the histiocytes

PERICAP INFILTR: Pericapsular infiltration

PLASMA: Plasma cells in the nonfollicular areas (percent of total cell population)

PLASMA TYPE: Plasma cell type

PLEOMORPHIC SR: Pleomorphic variants of Sternberg–Reed cells (number in 4-square-centimeter section)

PSEUDOFOLLICLES: Pseudofollicles

PTGC: Progressively transformed germinal centers

RUSSELL&DUTCHER: Russell and/or Dutcher bodies

SARCOMA CELLS: Sarcoma cells

SCHAUMAN: Schauman cells

SIGNET-RING: Signet-ring cells

SINUSES: Sinuses

SLC CHROMATIN: Chromatin structure of most small lymphoid cells

SLC CYTOPLASM: Cytoplasm of most small lymphoid cells

SLC EV CLUS: Small lymphoid cells in extravascular clusters of clear cells
SLC IV CLUS: Small lymphoid cells in intravascular clusters
SLC NUC SHP: Nuclear shape of most small lymphoid cells
SLC NUM: Number of small lymphoid cells in the nonfollicular areas (percent of total cell population)
SML LYMPH CELLS: Small lymphoid cells
SR-LIKE: Sternberg–Reed-like cells (number in 4cm^2 section)
SYSTEMIC AIDS: Systemic AIDS
T GENE REARRANGEMENT: T-cell receptor gene rearrangement
TRANSITION FORMS: Transition forms (lymphoid cells having sizes other than the sizes of small, medium-sized, or large cells) in the nonfollicular areas
VASC CHANGES: Endarteritis or periarteritis
VASC PROLIF NONSLIT: Vascular proliferation (non-slitlike)
VASC PROLIF SLIT: Vascular proliferation slitlike

List of Symbols

$\vert$	Sheffer stroke
ϕ	The greek letter phi
$\sum_i$	The summation over the index i
$\subseteq$	Subset of
$\in$	Element of
$\neq$	Not equal to

References

[1] D.E. Heckerman, E.J. Horvitz, and B.N. Nathwani. Pathfinder research directions. Technical Report KSL-89-64, Medical Computer Science Group, Section on Medical Informatics, Stanford University, Stanford, CA, October 1985.

[2] D.E. Heckerman, E.J. Horvitz, and B.N. Nathwani. Update on the Pathfinder project. In *Proceedings of the Thirteenth Symposium on Computer Applications in Medical Care, Washington, DC*, pages 203–207. IEEE Computer Society Press, Los Angeles, CA, November 1989.

[3] D. Kahneman, P. Slovic, and A. Tversky, editors. *Judgment Under Uncertainty: Heuristics and Biases.* Cambridge University Press, New York, 1982.

[4] A.S. Elstein. Clinical judgment: Psychological research and medical practice. *Science*, 194:696–700, 1976.

[5] R.J. Rosai and L.V. Ackerman. The pathology of tumors. part II: Diagnostic techniques. *Cancer*, 29:22–39, 1979.

[6] H.S. Levin. Operating room consultation by the pathologist. *Urology Clinics of North America*, 12:549–56, 1985.

[7] S.A. Rosenberg. Non-Hodgkin's lymphoma: Selection of treatment on the basis of histologic type. *New England Journal of Medicine*, 301:924–928, 1979.

[8] E.R. Gaynor and J.E. Ultmann. Non-Hodgkin's lymphoma: Management strategies. *New England Journal of Medicine*, 311:1506–1508, 1984.

[9] S.A. Rosenberg. The low-grade non-Hodgkin's lymphomas: Challenges and opportunities. *Journal of Clinical Oncology*, 3:299–310, 1985.

[10] G.E. Byrne. Rappaport classification of non-Hodgkin's lymphoma: Histologic featues and clinical significance. *Cancer Treatment Reports*, 61:935–944, 1977.

[11] C.A. Coltman, R.A. Gams, J.H. Glick, and R.D. Jenkin. Lymphoma. In B. Hoogstraten, editor, *Cancer Research Impact of the Cooperative Groups*, pages 39–84. Masson Publishing USC, 1980.

[12] S.E. Jones, J.J. Butler, G.E. Byrne, C.A. Coltman, and T.E. Moon. Histopathologic review of lymphoma cases from the southwest oncology group. *Cancer*, 39:1071–1076, 1977.

[13] H. Kim, R.J. Zelman, M.A. Fox, J.M. Bennett, C.W. Berard, J.J. Butler, G.E. Byrne, R.F. Dorfman, R.J. Hartsock, R.J. Lukes, R.B. Mann, R.S. Neiman, J.W. Rebuck, W.W. Sheehan, D. Variakojis, J.F. Wilson, and H. Rappaport. Pathology panel for lymophoma clinical studies: A comprehensive analysis of cases accumulated since its inception. *Journal of the National Cancer Institute*, 68:43–67, 1982.

[14] E. Velez-Garcia, J. Durant, R. Gams, and A. Bartolucci. Results of a uniform histopathological review system of lymphoma cases: A ten-year study of the southeastern cancer study group. *Cancer*, 52:675–679, 1983.

[15] A.S. Elstein, M.J. Loupe, and J.G. Erdman. An experimental study of medical diagnostic thinking. *Journal of Structural Learning*, 2:45–53, 1971.

[16] A.S. Elstein, L.S. Shulman, and S.A. Sprafka. *Medical Problem Solving: An Analysis of Clinical Reasoning.* Harvard University Press, Cambridge, MA, 1978.

[17] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior.* Princeton University Press, Princeton, NJ, 1947.

[18] R.A. Howard. Decision analysis: Perspectives on inference, decision, and experimentation. *Proceedings of the IEEE*, 58:632–643, 1970.

[19] D.J. Spiegelhalter. Probabilistic reasoning in predictive expert systems. In L.N. Kanal and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 47–67. North Holland, New York, 1986.

[20] R. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14:1–13, 1946.

[21] I.J. Good. *Probability and the Weighing of Evidence.* Hafners, New York, 1950.

[22] L.J. Savage. *The Foundations of Statistics.* Dover, New York, 1954.

[23] B. de Finetti. *Theory of Probability.* Wiley and Sons, New York, 1970.

[24] R.S. Ledley and L.B. Lusted. Reasoning foundations of medical diagnosis. *Science*, 130:9–21, 1959.

[25] H.R. Warner, A.F. Toronto, L.G. Veasy, and R. Stephenson. A mathematical approach to medical diagnosis: Application to congenital heart disease. *Journal of the American Medical Association*, 177:177–183, 1961.

[26] F.T. de Dombal, D.J. Leaper, J.R. Staniland, A.P. McCann, and J.C. Horrocks. Computer-aided diagnosis of acute abdominal pain. *British Medical Journal*, 2:9–13, 1972.

[27] G.A. Gorry and G.O. Barnett. Experience with a model of sequential diagnosis. *Computers and Biomedical Research*, 1:490–507, 1968.

[28] G.A. Gorry. Computer-assisted clinical decision making. *Methods of Information in Medicine*, 12:45–51, 1973.

[29] P. Szolovits. Artificial intelligence in medicine. In P. Szolovits, editor, *Artificial Intelligence in Medicine*, pages 1–19. Westview Press, Boulder, CO, 1982.

[30] R. Davis. Consultation, knowledge acquisition, and instruction. In P. Szolovits, editor, *Artificial Intelligence In Medicine*, pages 57–78. Westview Press, Boulder, CO, 1982.

[31] R.A. Miller, E.P. Pople, and J.D. Myers. INTERNIST-1: An experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine*, 307:476–486, 1982.

[32] R.A. Miller, M.A. McNeil, S.M. Challinor, F.E. Masarie, and J.D. Myers. The INTERNIST-1/Quick Medical Reference project–status report. *Western Journal of Medicine*, 145:816–822, 1986.

[33] H. Pople. Heuristic methods for imposing structure on ill-structured problems: The structuring of medical diagnostics. In P. Szolovits, editor, *Artificial Intelligence in Medicine*, pages 119–190. Westview Press, Boulder, CO, 1982.

[34] B.G. Buchanan and E.H. Shortliffe, editors. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project.* Addison-Wesley, Reading, MA, 1984.

[35] M.R. Genesereth. A meta-level representation system. Technical Report HPP-83-28, Heuristic Programming Project, Computer Science Department, Stanford University, Stanford, CA, 1983.

[36] G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, NJ, 1976.

[37] D.E. Heckerman. An empirical comparison of three inference methods. In R. Shachter T.S. Levitt, J. Lemmer, and L.N. Kanal, editors, *Uncertainty in Artificial Intelligence 4*. North Holland, New York, in press.

[38] D.E. Heckerman and H. Jimison. A perspective on confidence and its use in focusing attention during knowledge acquisition. In L. Kanal, T. Levitt, and J. Lemmer, editors, *Uncertainty in Artificial Intelligence 3*, pages 123–131. North Holland, New York, 1989.

[39] E.J. Horvitz. Reasoning about beliefs and actions under computational resource constraints. In L. Kanal, T. Levitt, and J. Lemmer, editors, *Uncertainty in Artificial Intelligence 3*, pages 301–324. North Holland, New York, 1989.

[40] G.F. Cooper. *NESTOR: A Computer-based Medical Diagnostic Aid that Integrates Causal and Probabilistic Knowledge.* PhD thesis, Computer Science Department, Stanford University, Stanford, CA, November 1984. Rep. No. STAN-CS-84-48. Also numbered HPP-84-48.

[41] D.J. Spiegelhalter and R.P. Knill-Jones. Statistical and knowledge-based approaches to clinical decision support systems, with an application in gastroenterology. *Journal of the Royal Statistical Society*, 147:35–77, 1984.

[42] J.A. Reggia and B.T. Perricone. Answer justification in medical decision support systems based on Bayesian classification. *Computers in Biology and Medicine*, 15:161–167, 1985.

[43] M. Ben-Bassat, V.K. Carlson, V.K. Puri, M.D. Davenport, J.A. Schriver, M.M. Latif, R. Smith, E.H. Lipnick, and M.H. Weil. Pattern-based interactive diagnosis of multiple disorders: The MEDAS system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2:148–160, 1980.

[44] S. Andreassen, M. Woldbye, B. Falck, and S.K. Andersen. MUNIN—a causal probabilistic network for interpretation of electromyographic findings. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence, Milan, Italy*, pages 366–372. Morgan Kaufman, San Mateo, CA, August 1987.

[45] D.E. Heckerman. An axiomatic framework for belief updates. In J.F. Lemmer and L.N. Kanal, editors, *Uncertainty in Artificial Intelligence 2*, pages 11–22. North Holland, New York, 1988.

[46] D.E. Heckerman and R.A. Miller. Towards a better understanding of the INTERNIST-1 knowledge base. In *Proceedings of Medinfo, Washington, DC*, pages 27–31. North Holland, New York, October 1986.

[47] D.E. Heckerman. Probabilistic interpretations for MYCIN's certainty factors. In L.N. Kanal and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 167–196. North Holland, New York, 1986.

[48] D.E. Heckerman and E.J. Horvitz. The myth of modularity in rule-based systems. In J.F. Lemmer and L.N. Kanal, editors, *Uncertainty in Artificial Intelligence 2*, pages 23–34. North Holland, New York, 1988.

[49] D.E. Heckerman and E.J. Horvitz. On the expressiveness of rule-based systems for reasoning under uncertainty. In *Proceedings AAAI-87 Sixth National Conference on Artificial Intelligence, Seattle, WA*, pages 121–126. Morgan Kaufmann, San Mateo, CA, July 1987.

[50] E.J. Horvitz and D.E. Heckerman. The inconsistent use of measures of certainty in artificial intelligence research. In L.N. Kanal and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 137–151. North Holland, New York, 1986.

[51] D.E. Heckerman. Formalizing heuristic methods for reasoning with uncertainty. Technical Report KSL-88-07, Medical Computer Science Group, Section on Medical Informatics, Stanford University, Stanford, CA, May 1987.

[52] E.J. Horvitz, J.S. Breese, and M. Henrion. Decision theory in expert systems and artificial intelligence. *International Journal of Approximate Reasoning*, 2:247–302, 1988.

[53] E.J. Horvitz, D.E. Heckerman, and C.P. Langlotz. A framework for comparing alternative formalisms for plausible reasoning. In *Proceedings AAAI-86 Fifth National Conference on Artificial Intelligence, Philadelphia, PA*, pages 210–214. Morgan Kaufmann, San Mateo, CA, August 1986.

[54] M. Henrion. Towards efficient probabilistic diagnosis in multiply connected networks. In *Proceedings of the Conference on Influence Diagrams for Decision Analysis, Inference, and Prediction, Berkeley, CA*, page ?? not published, May 1988.

[55] D.E. Heckerman. A tractable algorithm for diagnosing multiple diseases. In *Proceedings of Fifth Workshop on Uncertainty in Artificial Intelligence, Windsor, ON*, pages 174–181. Association for Uncertainty in Artificial Intelligence, Mountain View, CA, August 1989.

[56] M. Tribus. *Rational Descriptions, Decisions, and Designs.* Pergamon Press, New York, 1969.

[57] R.A. Howard and J.E. Matheson. Influence diagrams. In R.A. Howard and J.E. Matheson, editors, *Readings on the Principles and Applications of Decision Analysis*, volume II, pages 721–762. Strategic Decisions Group, Menlo Park, CA, 1981.

[58] J. Pearl. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29:241–288, 1986.

[59] R.D. Shachter. Evaluating influence diagrams. *Operations Research*, 34:871–882, 1986.

[60] S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *J. Royal Statistical Society B*, 50:157–224, 1988.

[61] M. Henrion. Propagation of uncertainty by probabilistic logic sampling in Bayes' networks. In J.F. Lemmer and L.N. Kanal, editors, *Uncertainty in Artificial Intelligence 2*, pages 149–164. North Holland, New York, 1988.

[62] D.E. Heckerman. Probabilistic similarity networks. *Networks*, to appear.

[63] D.E. Heckerman. *Probabilistic Similarity Networks.* PhD thesis, Medical Computer Science Group, Section on Medical Informatics, Stanford University, Stanford, CA, in preparation.

[64] B. J. McNeil, S. G. Pauker, H. C. Sox, and A. Tversky. On the elicitation of preferences for alternative therapies. *New England Journal of Medicine*, 306:1259–62, 1982.

[65] M.J. Barry, A.B.Mulley, Jr., F.J. Fowler, and J.W. Wennberg. Watchful waiting versus immediate transurethral resection for symptomatic prostatism: The importance of patients' preferences. *Journal of the American Medical Association*, 259:3010–17, 1988.

[66] H.B. Jimison. Generating explanations of decision models based on an augmented representation of uncertainty. In R.D. Shachter, L. Kanal, T. Levitt, and J. Lemmer, editors, *Uncertainty in Artificial Intelligence 4*. North Holland, New York, 1990. in press.

[67] R.A. Howard. On making life and death decisions. In R.A. Howard and J.E. Matheson, editors, *Readings on the Principles and Applications of Decision Analysis*, volume II, pages 483–506. Strategic Decisions Group, Menlo Park, CA, 1980.

[68] E.J. Horvitz, D.E. Heckerman, K.C. Ng, and B.N. Nathwani. Heuristic abstraction in the decision-theoretic Pathfinder system. In *Proceedings of the Thirteenth Symposium on Computer Applications in Medical Care, Washington, DC*, pages 178–182. IEEE Computer Society Press, Los Angeles, CA, 1989.

[69] E.J. Horvitz, D.E. Heckerman, B.N. Nathwani, and L.M. Fagan. The use of a heuristic problem-solving hierarchy to facilitate the explanation of hypothesis-directed reasoning. In *Proceedings of Medinfo, Washington, DC*, pages 27–31. North Holland, New York, October 1986.

[70] W.J. Clancey. Heuristic classification. *Artificial Intelligence*, 27:289–350, 1985.

[71] R.L. Teach and E.H. Shortliffe. An analysis of physician attitudes regarding computer-based clinical consultation systems. *Computers and Biomedical Research*, 14:542–558, 1981.

[72] E.H. Shortliffe. The computer and medical decision making: Good advice is not enough. *IEEE Engineering in Medicine and Biology Magazine*, 1:16–18, 1982 (guest editorial).

[73] P.E. Politser. Explanations of statistical concepts: Can they penetrate the haze of Bayes? *Methods of Information in Medicine*, 23:99–108, 1984.

[74] M. N. Pollak. Computer-aided information management systems in clinical trials. *Computer Programs and Biomedicine*, 16:243–252, 1983.

[75] F.T. de Dombal, D.J. Leaper, J.C. Horrocks, J.R. Staniland, and A.P. McCain. Human and computer-aided diagnosis of abdominal pain: further report with emphasis on performance. *British Medical Journal*, 1:376–380, 1974.

[76] R.M. Dawes and B. Corrigan. Linear models in decision making. *Psychological Bulletin*, 81:95–106, 1974.

[77] E.H. Shortliffe. *Computer-based Medical Consultations: MYCIN.* North Holland, New York, 1976.

[78] E.H. Shortliffe, A.C. Scott, M.B. Bischoff, A.B. Campbell, W. Van Melle, and C.D. Jacobs. ONCOCIN: An expert system for oncology protocol management. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence, Vancouver, BC*, pages 876–881. International Joint Conference on Artificial Intelligence, August 1981.

[79] D. L. Kent, E. H. Shortliffe, R. W. Carlson, M. B. Bischoff, and C. D. Jacobs. Improvements in data collection through physician use of a computer-based chemotherapy treatment consultant. *Journal of Clinical Oncology*, 3:1409–1417, 1985.

[80] A.H. Murphy. A note on the utility of probabilistic predictions and the probability score in the cost–loss ratio decision situation. *Journal of Applications in Meteorology*, 5:534–537, 1966.

[81] J. Pearl. An economic basis for certain methods of evaluating probabilistic forecasts. *International Journal of Man–Machine Studies*, 10:175–183, 1978.

[82] B. Wise. *An Experimental Comparison of Uncertain Inference Systems.* PhD thesis, The Robotics Institute and Department of Engineering and Public Policy, Carnegie–Mellon University, Pittsburgh, PA, June 1986.

[83] P. Smets, J. Willems, J. Talmon, V. DeMaertelaer, and F. Kornreich. Methodology for the comparison of various diagnostic procedures. *Biometrie–Praximetrie*, 15:89–122, 1975.

[84] B. Asselain, C. Derouesne, R. Salamon, M. Bernadet, and F. Gremy. The concept of utility in a medical decision aid: Example of an application. In *Proceedings of Medinfo, Halifax, Nova Scotia*, pages 123–125. North Holland, New York, October 1977.

[85] J.D.F. Habbema and J. Hilden. The measurement of performance in probabilistic diagnosis IV: Utility considerations in therapeutics and prognostics. *Methods of Information in Medicine*, 20:80–96, 1981.

[86] P. Diaconis and B. Efron. Computer-intensive methods in statistics. *Scientific American*, 248:116–130, 1983.