Toward Normative Expert Systems: Part I
The Pathfinder Project

David E. Heckerman
Departments of Computer Science and Pathology
University of Southern California
HMR 204, 2025 Zonal Ave
Los Angeles, CA 90033
<dheck@sumex-aim.stanford.edu>

Eric J. Horvitz
Palo Alto Laboratory
Rockwell International Science Center
444 High Street
Palo Alto, California 94301
<horvitz@sumex-aim.stanford.edu>

Bharat N. Nathwani
Department of Pathology
University of Southern California
HMR 204, 2025 Zonal Ave
Los Angeles, CA 90033

To appear in *Methods of Information in Medicine,* 1992

Abstract

Pathfinder is an expert system that assists surgical pathologists with the diagnosis of lymph-node diseases. The program is one of a growing number of normative expert systems that use probability and decision theory to acquire, represent, manipulate, and explain uncertain medical knowledge. In this article, we describe Pathfinder and our research in uncertain-reasoning paradigms that was stimulated by the development of the program. We discuss limitations with early decision-theoretic methods for reasoning under uncertainty and our initial attempts to use non-decision-theoretic methods. Then, we describe experimental and theoretical results that directed us to return to reasoning methods based in probability and decision theory.

Keywords: expert systems, decision making, diagnosis, probability theory, decision theory, artificial intelligence, pathology

1 Introduction

Decision-theoretic or *normative* expert systems have the potential to provide better decision support than do traditional expert systems in problem areas or domains where the accurate management of uncertainty is important. This potential for improvement arises because people, including experts, make mistakes when they make decisions under uncertainty. That is, people often deviate from the rules of decision theory, which provides a set of compelling principles or desiderata for how people should behave when reasoning or making decisions under uncertainty. Decision theory includes the rules of probability and the principle that a person should always choose the alternative that maximizes his expected utility.

Traditional expert systems provide decision support by mimicking the recommendations of experts. They do so by managing uncertainty with non-decision-theoretic methods. Such systems are valuable, because they provide important information to a nonexpert who is confronted with a confusing decision, and because they offer reminders to users who may be stressed or fatigued. Nonetheless, they tend to duplicate the errors made by experts.

In contrast, normative expert systems use decision theory to manage uncertainty. The word "normative" comes from decision analysts and cognitive psychologists who emphasize the importance of distinguishing between *normative behavior*, which is what we do when we follow the desiderata of decision theory, and *descriptive behavior*, which is what we do when unaided by these desiderata. By encoding expert knowledge in a decision-theoretic framework, we can reduce errors in reasoning, and thereby build expert systems that offer recommendations of higher quality.

In this article, we describe Pathfinder, a normative expert system that assists surgical pathologists with the diagnosis of lymph-node diseases [1, 2]. The Pathfinder project began in 1983 as a joint project among researchers at Stanford University (David Heckerman, Eric Horvitz, and Larry Fagan) and the University of Southern California (Bharat Nathwani—the primary pathology expert—and Keung-Chi Ng) [3]. Currently, a commercial derivative of Pathfinder, called Intellipath, is being used by practicing pathologists and by pathologists in training as an educational tool [4].

Also in this article, we discuss the importance of the proper management of uncertainty for diagnosis of lymph-node diseases; and we discuss our research in uncertain-reasoning paradigms that was stimulated by the development of Pathfinder. In particular, we examine practical limitations with early decision-theoretic methods for reasoning under uncertainty and our initial attempts to overcome these limitations through the use of non-decision-theoretic reasoning paradigms. Then, we describe experiments with these non-decision-theoretic approaches as well as theoretical analyses that directed us to return to a methodology based in probability and decision theory. In the companion to this article, we describe the decision-theoretic representations that we developed to make practical the construction of a normative version of Pathfinder.

2 Diagnosis in Surgical Pathology and Pathfinder

Surgical pathologists perform diagnosis primarily by examining sections of tissue microscopically. Sometimes, pathologists also incorporate clinical, radiology, and laboratory informa-

tion, and examine tissue with expensive tests derived from immunology, microbiology, and cell-kinetics research. Based on this information, the pathologist provides a diagnosis to the surgeons and oncologists who participate in the patient's treatment. That is, the pathologist tells these physicians, "the patient has disease x."

The well-being of patients depends strongly on the accuracy of the pathologist's diagnosis. In the case of lymph-node diagnosis, for example, let us suppose that the patient has Hodgkin's disease, a malignant disease, but that the pathologist makes a diagnosis of mononucleosis, a benign disease that can resemble Hodgkin's disease. In this situation, the patient's chance of death is significantly greater than it would have been had the diagnosis been correct, because he does not receive immediate treatment for his malignancy. In contrast, let us suppose that the patient has mononucleosis, and that the pathologist makes a diagnosis of Hodgkin's disease. In this case, the patient likely will undergo expensive, painful, and debilitating treatment, to be "cured," only because he never had the malignant disease in the first place.

A general pathologist performs diagnosis on tissue sections from all parts of the body. When a general pathologist has difficulty with diagnosis, he frequently refers the case to a subspecialist, who has expertise in the diagnosis of a particular tissue type. 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].

The subspecialty of lymph-node diagnosis is one of the most difficult areas in surgical pathology [7, 8, 9, 10]. For example, one multisite oncology study analyzed almost 9000 cases of malignant lymphoma. The study found that although experts show agreement with one another, the diagnoses rendered by general pathologists for certain diseases had to be changed by expert lymph-node pathologists in as many as 65 percent of the cases [10]. Our goal in building Pathfinder is to close the wide gap between the quality of lymph-node diagnoses made by general pathologists and those made by subspecialists. We hope to increase the accuracy of in-house pathology diagnoses, to reduce the frequency of referrals, and to assist pathologists with intraoperative diagnosis when there is insufficient time for expert consultation.

Pathologists have difficulty with diagnosis for two reasons. First, they may misrecognize or fail to recognize microscopic features. Second, they may combine evidence inaccurately to form a diagnosis. The second problem arises because the pathologist must consider an enormous number of features and diseases, and because the relationships among diseases and features are uncertain. Most of Pathfinder research has concentrated on the evidence-combination problem. That is, we have worked to develop an expert system that can help pathologists cope with the many uncertain relationships in the diagnosis of lymph-node pathology. Indeed, Pathfinder reasons about more than 60 diseases that can invade the lymph node (25 benign diseases, 9 Hodgkin's lymphomas, 18 non-Hodgkin's lymphomas, and 10 metastatic diseases), using more than 130 microscopic, clinical, laboratory, immunologic, and molecular-biologic features. Similarly, in this article, we concentrate on the problem of managing uncertainty in large domains. Nevertheless, as we mention in Section 9, we also have addressed the feature-recognition problem.

3 A Pathfinder Dialog

In rendering a diagnosis, a pathologist (1) identifies and quantifies features; (2) constructs a *differential diagnosis*, a set of diseases consistent with the observations; and (3) decides what additional features to evaluate and what costly tests to employ to narrow the differential diagnosis. He repeats these steps until he has observed all useful features. This procedure is called the *hypothetico-deductive* approach [11, 12, 13, 14]. Cognitive psychologists have found that physicians frequently employ this approach in performing clinical diagnosis [12, 14].

Pathfinder uses this same method, summarized in Figure 1, to assist pathologists with their task of diagnosis. Associated with each feature are two or more mutually exclusive and exhaustive *instances*. For example, the feature NECROSIS is associated with the instances ABSENT, PRESENT, and PROMINENT. The Pathfinder system allows a user to report instances for one or more salient features of a lymph-node section. Given these feature–instance 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 that the informational benefits are not worth the costs of further observations or tests.

The operation of the latest version of Pathfinder is illustrated by the set of screen photos in Figure 2. Figure 2(a) 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 feature–instance pairs that will be observed by the pathologist, and the DIFFERENTIAL DIAGNOSIS window displays the list of possible diseases and their probabilities. The probabilities in Figure 2(a) are the *prior probabilities* of disease—the probabilities for disease given only that a patient's lymph node has been removed and is being examined.

If the user selects (double-clicks) the feature category SPHERICAL FEATURES, then Pathfinder displays a list of features for that category. To enter a particular feature, the user double-clicks on that feature, and then selects one of the mutually exclusive and exhaustive instances for that feature. For example, Figure 2(b) 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 instances for this feature: NA (not applicable), 1–10%, 11–50%, 51–75%, 76–90%, and >90%. Figure 2(c) shows the result of selecting the last instance for this feature. In particular, the feature–instance 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 2(d) 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 2(e) 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 2(f). In this case, the four features in the middle column have narrowed the differential diagnosis to a single disease: the early phase of AIDS.

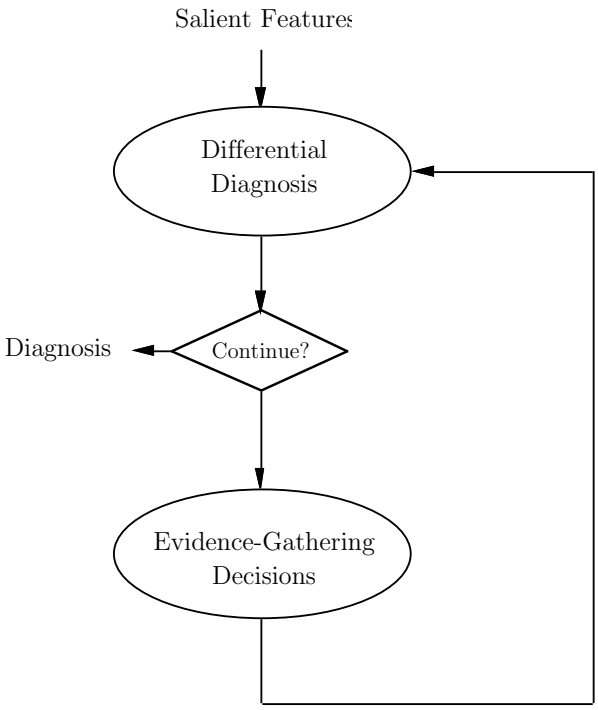

Figure 1: Hypothetico-deductive reasoning in Pathfinder. First, the pathologist reports instances of salient features to the system. The system then constructs a differential diagnosis—a list of hypotheses that are consistent with the observations, and an assignment of likelihood to each such hypothesis. Next, the system analyzes the current differential diagnosis to identify the most useful features for the pathologist to observe. The process cycles until the differential diagnosis is narrowed to a single disease, there are no additional tests or questions, or the pathologist determines that the informational benefits are not worth the costs of further observations or tests.

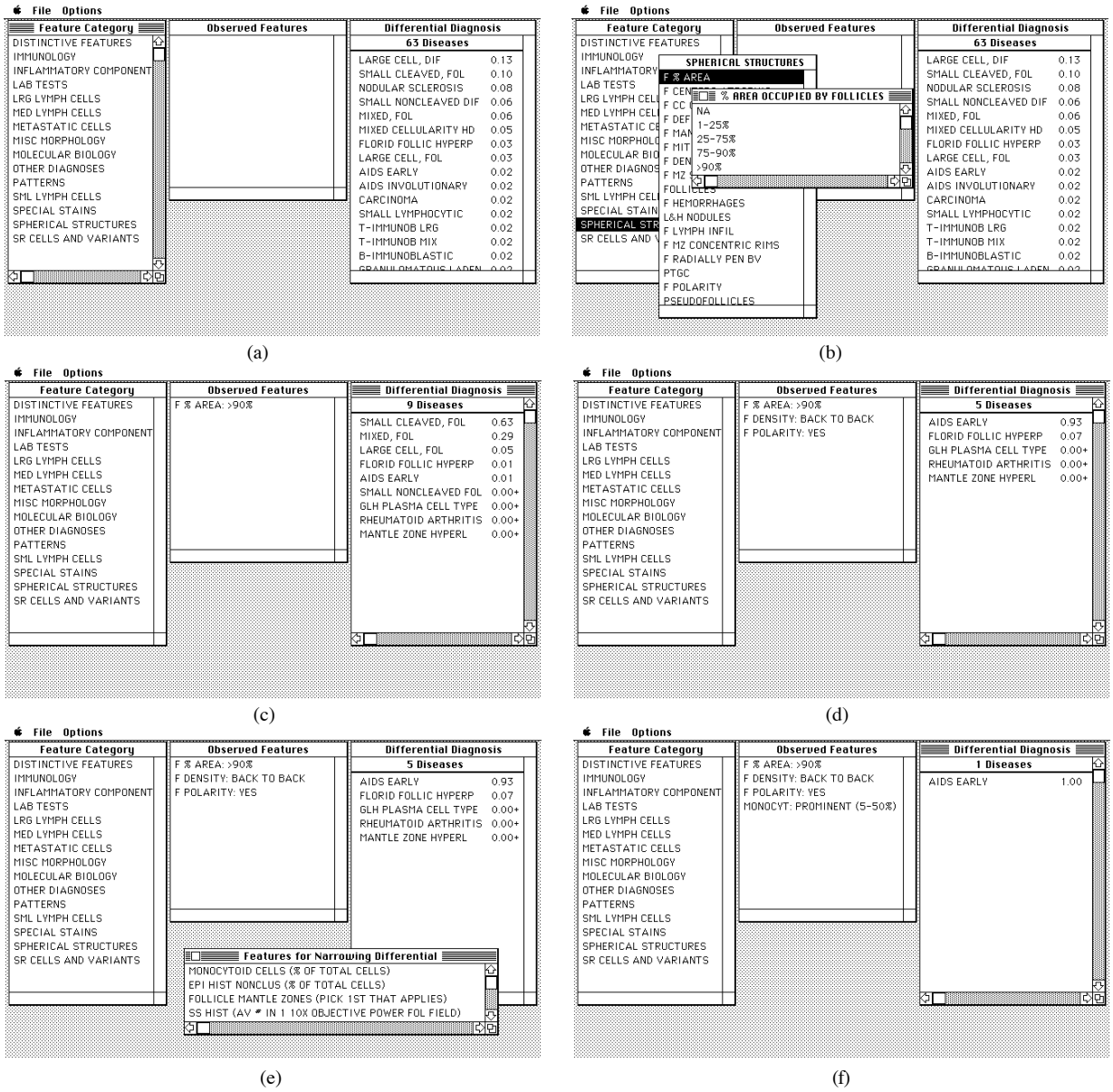

Figure 2: A Pathfinder consultation. (a) Initially, Pathfinder displays (from left to right) the categories of features, an empty window that will contain feature-instance pairs reported by the user, and the prior probabilities of disease. (b) Double-clicking on the category SPHERICAL STRUCTURES and then on the feature F % AREA, the pathologist prepares to report to Pathfinder the percent area occupied by follicles. (c) Double-clicking on the instance >90%, the pathologist reports that more than 90% of the lymph-node is occupied by follicles. In response, the program produces a differential diagnosis in the right-hand window. (d) The pathologist now reports that follicles are in a back-to-back arrangement and show polarity. Pathfinder revises the differential diagnosis. (e) The pathologist has asked Pathfinder to display features that are useful for narrowing the differential diagnosis. The program displays the four most cost-effective features for the user to observe next. The most useful feature is monocytoid cells. (f) The user now reports that monocytoid cells are prominent. Pathfinder determines that only a single disease—AIDS EARLY (the early phase of AIDS)—is consistent with the four observations. (Adapted with permission from D. Heckerman, *Probabilistic Similarity Networks,* MIT Press, Cambridge, MA, 1991.)

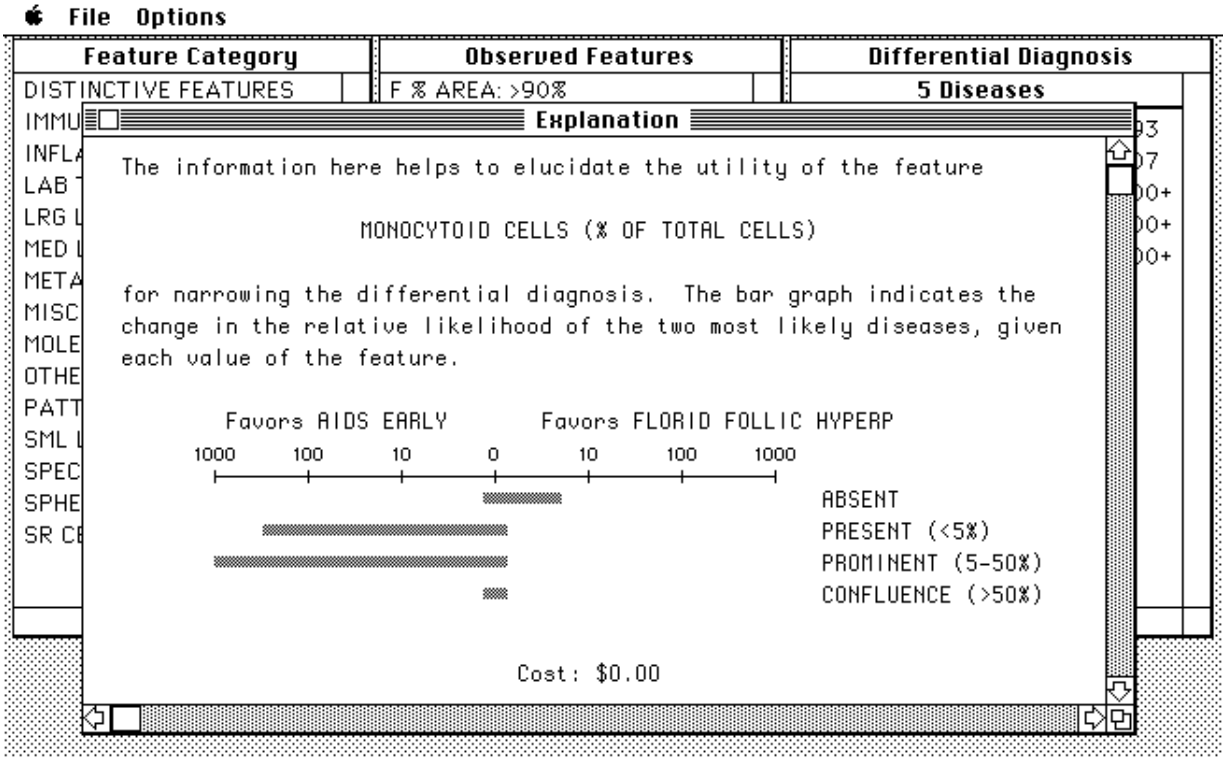

Figure 3: A graphical justification for the recommendation of MONOCYTOID CELLS. For each instance of the feature, the length and direction of a bar reflects the change in the probability of AIDS EARLY relative to that of FLORID FOLLIC HYPERP, given the observation of that feature–instance pair. The justification also includes the monetary cost of observing the feature. (Taken with permission from D. Heckerman, *Probabilistic Similarity Networks,* MIT Press, Cambridge, MA, 1991.)

Pathfinder explains graphically its recommendations for additional observations. A bitmap of Pathfinder's graphical justification of the diagnostic utility of the feature monocytoid cells is displayed in Figure 3. In this explanation, Pathfinder displays the change in probability of the two most likely hypotheses given the observation of each possible instance of the feature. The graph indicates that if monocytoid cells are absent, then the probability of FLORID FOLLIC HYPERP relative to that of AIDS EARLY increases slightly; if monocytoid cells are present or prominent, then the probability of FLORID FOLLIC HYPERP relative to that of AIDS EARLY decreases greatly; and if monocytoid cells show confluence, then the probability of FLORID FOLLIC HYPERP relative to that of AIDS EARLY remains unchanged. By glancing at this graph, we can see that this feature is useful for discriminating these two diseases. The window also displays the monetary cost of evaluating the feature, which is negligible in this case.

4 Decision-Theoretic Computations in Pathfinder

Both an early version and the latest version of Pathfinder employ decision-theoretic computations to assist pathologists with diagnosis. In this section, we examine these computations. In particular, we examine how Pathfinder (1) uses probabilistic inference to generate a differential diagnosis, (2) uses decision theory to recommend a diagnosis, and (3) uses decision theory to recommend features for observation. First, however, let us discuss some fundamentals of probability and decision theory.

4.1 Probability Theory

Probability theory has roots, more than 3 centuries ago, in the work of Bernoulli, Laplace, Fermat, and Pascal [15]. The theory describes how to infer the probability of one event from the probability of related events. The prevalent conception of the probability of some event x is that it is a measure of the frequency with which x occurs, when we repeat many times an experiment that has x as a possible outcome. A more general notion, however, is that the probability of x represents the *degree of belief* held by a person that the event x will occur in a single experiment. If a person assigns a probability of 1 to x, then he believes with certainty that x will occur. If he assigns a probability of 0 to x, then he believes with certainty that x will not happen. If he assigns a probability of between 0 and 1 to x, then he is to some degree unsure about whether or not x will occur.

The interpretation of a probability as a frequency in a series of repeated experiments traditionally is referred to as the *objective* or *frequentist* interpretation. In contrast, the interpretation of a probability as a degree of belief is called the *subjective* or *Bayesian* interpretation, in honor of the Reverend Thomas Bayes, a scientist from the mid-1700s who helped to pioneer the theory of probabilistic inference [16, 15]. Both interpretations follow the same set of mathematical rules.

In the Bayesian interpretation, a probability or belief will always depend on the state of knowledge of the person who provides that probability. For example, if we were to give someone a coin, he would likely assign a probability of 1/2 to the event that the coin would show heads on the next toss. If, however, we convinced that person that the coin was weighted in favor of heads, he would assign a higher probability to the event. Thus, we write the probability of x as $p(x|\xi)$, which is read as the probability of x *given* ξ. The symbol ξ represents the state of knowledge or *background knowledge* of the person who provides the probability.

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. For example, there are more than 14 thousand probabilities in the latest version of Pathfinder; and some of these probabilities are on the order of 10^{-6}. Thus, performing the experiments necessary to determine objective probabilities for Pathfinder would entail much time and great expense. Fortunately, when experimental data is available, the Bayesian approach provides a mechanism for expert systems to update their probabilities, given this data [17, 18, 19].

4.2 Decision Theory and Utility Assessment

Decision theory extends the Bayesian interpretation of probability theory, and prescribes how a decision maker should choose among a set of *alternatives* or actions, given his *utility* or preference for each possible outcome and his belief that each outcome will occur. In particular, decision theory includes the rules of probability theory and the maximum-expected-utility (MEU) principle, which states that a decision maker should always choose the alternative that maximizes his expected utility [20].

Utility assessment is nontrivial and is the subject of many debates. In this article, we mention only a few important points concerning utility assessment for Pathfinder. The interested reader should consult more general discussions by Keeney and Raiffa [21], McNeil et al. [22], and Howard [23].

For each disease pair (d_j, d_k) in Pathfinder, we encode the utility $u(d_j, d_k)$, which summarizes the preferences of the decision maker for the situation in which a patient has disease d_j, but is diagnosed as having disease d_k. 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.

An important consideration in the assessment of these (and any other) utilities is: Who is the decision maker? From our perspective, a pathologist is only a provider of information. Thus, the $u(d_j, d_k)$ 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? As we discussed, 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. The decision about therapy will depend on how the patient feels about the alternative outcomes. Different patients may have dramatically different preferences.

As we discuss in Sections 4.4 and 4.5, differences in patient preferences can in principle affect recommendations made by an expert system for diagnosis. Thus, utility assessment poses a fundamental problem to any researcher who wants to develop such expert systems. Specifically, whenever a patient case is processed by an expert system, the system or a decision analyst should assess the utilities of that patient and provide these utilities to the system. Such utility assessment would be extremely time consuming and expensive. As we see in Sections 4.4 and 4.5, however, only Pathfinder's diagnostic recommendations and not its recommendations for evidence gathering are sensitive to patient utilities. Thus, by allowing Pathfinder to make only evidence-gathering recommendations, we render the program's recommendations insensitive to patient utilities. We can therefore encode in Pathfinder the utilities $u(d_j, d_k)$ from one representative patient.

To construct Pathfinder's utility model, we assessed the utilities of Bharat Nathwani, the primary Pathfinder expert. We found it relatively easy to assess his utilities, because he was familiar with the ramifications of many specific correct and incorrect diagnoses.

Another important consideration in utility assessment is the wide range of severities as-

sociated with outcomes. For example, if a patient has a viral infection and is incorrectly diagnosed as having cat-scratch disease—a disease caused by an organism that is killed with antibiotics—the consequences are not severe. In fact, the only non-negligible consequence is that the patient will take antibiotics unnecessarily for several weeks. If a patient has Hodgkin's disease and is incorrectly diagnosed as having mononucleosis, however, the consequences are often lethal.

It is important for us to measure preferences across such a wide range, because sometimes we must balance a large chance of a small loss with a small chance of a large loss. For example, even though the probability that a patient has syphilis is small—say, 0.001—treatment with antibiotics may be appropriate, because the patient may prefer the harmful effects of antibiotics to the small chance of the harmful effects of untreated disease.

Early attempts to assess preferences for both minor and major outcomes in the same unit of measurement were fraught with paradoxes. For example, in a linear willingness-to-pay approach, a decision maker might be asked, "How much would you have to be paid to accept a one in ten-thousand chance of death?" If the decision maker answered, say, \$1000, then the approach would dictate absurdly that he would be willing to be killed for \$10 million.

Howard (1980) constructed an approach that avoids many of the paradoxes of earlier models. Like several of its predecessors, the model determines what an individual is willing to pay to avoid a given chance of death, and what he is willing to be paid to assume a given chance of death. Also, like many of its predecessors, Howard's model shows that, for small risks of death (typically, $p < 0.001$), the amount someone is willing to pay to avoid, or is willing to be paid to assume, such a risk is linear in p. That is, for small risks of death, an individual acts as would an expected-value decision maker with a finite value of life, called the *small-risk value of 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. Most people find this result to be intuitive.[1]

To use this model, we first determined Bharat Nathwani's small-risk value of life. When asked what dollar amount he would be willing to pay to avoid chances of death ranging from 1 in 20 to 1 in 1000, he was consistent with the linear model to within a factor of 2, with a median small-risk value of life equal to \$20 million (in 1988 dollars). To make the application of the model more convenient, we used Howard's definition of a *micromort*: one–in–1-million chance of death [24]. In these units, the Pathfinder expert's small-risk value of life was \$20 per micromort.[2]

Given this small-risk value of life, we could then measure his preferences for major and minor outcomes in a common unit: *1 minus the probability of immediate, painless death that he was willing to accept to avoid a given outcome and to be once again healthy.* In particular, we assessed his preferences for minor outcomes with willingness-to-pay questions, such as "How much would you be willing to pay to avoid taking antibiotics for two weeks?" We then translated these answers, via the linearity result of Howard's model, to units of probability

[1]The result makes several assumptions, such as the decision maker is not suicidal and is not concerned about how his legacy will affect other people.

[2]In general, the micromort is a useful unit of measurement, because it helps to emphasize that the linear relationship between risk of death and willingness to pay holds for only small probabilities of death.

of death. For example, an answer of $100 translated to a utility of

$$1 - 5 \text{ micromorts} = 1 - 0.000005 = 0.999995$$

We assessed his preferences for major outcomes directly in units of probability of death. For example, he imagined that he had—say—Hodgkin's disease, and that he had been misdiagnosed as having mononucleosis. He then imagined 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. He then provided the value of p that made him indifferent between his current situation and the situation in which he takes the pill. The utility of this outcome is $1 - p$.

4.3 Construction of a Differential Diagnosis

First, let us examine the problem of differential diagnosis in general. Let m and n denote the number of diseases and features in a medical domain, respectively. Also, let $d_1, d_2, \ldots, d_m$ denote the disease entities. For the moment, let us suppose that each disease d_j may be present or absent. Let D_k denote some *instance* of diseases. That is, D_k denotes some assignment of present or absent to each of the diseases $d_1, d_2, \ldots, d_m$. Further, let f_1, f_2, $\ldots$, f_n denote the features in the domain, and let i_j denote the observed instance for the jth feature.

Now imagine that a user of a probabilistic expert system for this domain has observed instances for q features. To simplify the notation, let us renumber the n features so that the user has observed instances for the first q features. Typically, the user will want to know the probability of each disease instance, given the observations $f_1 i_1$, $f_2 i_2$, $\ldots$, $f_q i_q$. This quantity for disease instance D_k is known as the *posterior probability* of D_k, and is denoted

$$p(D_k | f_1 i_1, f_2 i_2, \ldots, f_q i_q, \xi) \tag{1}$$

Thus, the number of probabilities of interest is exponential both in the number of observed features and in the number of diseases.

In principle, an expert could assess directly these posterior probabilities. Aside from the intractable nature of this task, however, 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 $f_1 i_1$, $f_2 i_2$, $\ldots$, $f_q i_q$ will appear given a particular disease instance D_k, denoted

$$p(f_1 i_1, f_2 i_2, \ldots, f_q i_q | D_k, \xi) \tag{2}$$

Using Bayes' theorem, the expert system can compute from these probabilities and the prior probability of disease instances $p(D_k|\xi)$ the desired posterior probabilities

$$p(D_k | f_1 i_1, f_2 i_2, \ldots, f_q i_q, \xi) = \frac{p(f_1 i_1, f_2 i_2, \ldots, f_q i_q | D_k, \xi)\ p(D_k|\xi)}{\sum_{D_l}\ p(f_1 i_1, f_2 i_2, \ldots, f_q i_q | D_l, \xi)\ p(D_l|\xi)} \tag{3}$$

where the sum over D_l runs over all disease instances. Unfortunately, this approach to the problem is also intractable, because the number of probabilities of the form

$p(f_1 i_1, f_2 i_2, \ldots, f_q i_q | D_k, \xi)$ is exponential both in the number of diseases and in the number of features.

To manage the complexity of the general case, researchers who built the first probabilistic expert systems made two assumptions. First, they supposed that *all* findings were *conditionally independent*, given any disease instance. That is, they assumed that, if the true disease state of the patient was known, then the likelihood of seeing any observation $f_k i_k$ did not depend on observations made about any other features. Thus,

$$p(f_j i_j | D_k, f_1 i_1, \ldots, f_{j-1} i_{j-1}, f_{j+1} i_{j+1}, \ldots, f_q i_q, \xi) = p(f_j i_j | D_k, \xi) \tag{4}$$

Given this assumption, it follows from the rules of probability [25] that

$$p(f_1 i_1, f_2 i_2, \ldots, f_q i_q | D_k, \xi) = p(f_1 i_1 | D_k, \xi)\ p(f_2 i_2 | D_k, \xi)\ \cdots\ p(f_q i_q | D_k, \xi) \tag{5}$$

Second, these researchers 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 expert system can compute the posterior probabilities of disease from the tractable computation

$$p(d_k | f_1 i_1, f_2 i_2, \ldots, f_q i_q, \xi) = \frac{p(f_1 i_1 | d_j, \xi)\ p(f_2 i_2 | d_k, \xi)\ \cdots\ p(f_q i_q | d_k, \xi)\ p(d_k | \xi)}{\sum_{d_l}\ p(f_1 i_1 | d_l, \xi)\ p(f_2 i_2 | d_l, \xi)\ \cdots\ p(f_q i_q | d_l, \xi)\ p(d_l | \xi)} \tag{6}$$

where d_k represents the disease instance in which only disease d_k is present. Thus, only the conditional probabilities $p(f_j i_j | d_k, \xi)$ and the prior probabilities $p(d_k | \xi)$ are required for the computation. We call any model that employs these two assumptions a *simple-Bayes model*. Ledley and Lusted proposed this model for medical diagnosis in 1959 [26].

In the domain of lymph-node pathology, 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. 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 inaccurate. For example, given certain diseases, 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. Nonetheless, to simplify our task, we used the simple-Bayes model to construct the first probabilistic version of Pathfinder. Later, after developing several graphical representation languages that we describe in the companion to this article, we encoded successfully the conditional nonindependence or *conditional dependence* among the features in the domain. We shall return to this discussion in Section 8.

4.4 Recommendation of a Diagnosis

As we mentioned, a diagnosis is a statement of the form: "The patient has disease x." Sometimes, as we saw in the patient case in Section 3, the posterior probability of one disease will equal 1 and the posterior probability of all other diseases will equal 0. In this case, making a diagnosis is not a decision. Rather, the diagnosis is a consequence of the rules of logic. In most cases, however, observations usually do not narrow the differential

diagnosis to a single disease. In these situations, making a diagnosis is a decision: an irrevocable allocation of resources under uncertainty.

Using the MEU principle, the system can determine a diagnosis from the probabilities of disease and the utilities $u(d_j, d_k)$. Let ϕ denote the set of feature–instance pairs $f_1 i_1$, $f_2 i_2$, ..., $f_q i_q$ that we have observed thus far. First, for each diagnosis d_k, the system computes $eu(d_k|\phi)$, the expected utility of that diagnosis given observations ϕ, using the formula

$$eu(d_k|\phi) = \sum_{d_j} p(d_j|\phi)\ u(d_j, d_k) \tag{7}$$

To complete the determination, the system selects the optimal diagnosis, denoted $dx(\phi)$, using the equation

$$dx(\phi) = \text{argmax}_{d_k}\ [eu(d_k|\phi)] \tag{8}$$

where the function $\text{argmax}_{d_k}\ [\cdot]$ returns the disease that maximizes its argument.

We do not allow Pathfinder to recommend diagnoses, because we have observed that such recommendations are somewhat sensitive to the utility model. That is, when we change the utilities in the model from values appropriate for one patient to values appropriate for another patient, the program's recommendations can change significantly. By preventing Pathfinder from recommending diagnoses, we hope to encourage a change in the way pathologists and care-providing physicians communicate. In the short term, we hope that pathologists will begin to express clearly—in the language of probability—uncertainty associated with their observations. In the long term, we hope that each physician who is associated with the care of a patient—including the primary physician, the pathologist, the radiologist, the surgeon, the oncologist, and the radiotherapist—and the patient himself will communicate in decision-theoretic terms to determine the best treatment for that patient. Such communication could take place via a shared decision model embodied in an expanded version of Pathfinder.

4.5 Recommendation of Features to Narrow a Differential Diagnosis

Let us now consider how an expert system can use decision theory to recommend features for observation to narrow a differential diagnosis. First, the system enumerates all possible *observation strategies.* An example of an observation strategy is

> Observe f_3. If f_3 is present, then observe f_2; otherwise, make no further observations and make the diagnosis. If f_3 and f_2 are present, then observe f_7, and make the diagnosis. If f_3 is present and f_2 is absent, then make the diagnosis.

Next, the system computes the decision maker's expected utility of all strategies, including the strategy in which the user observes no additional features. Finally, the system chooses the strategy that maximizes the decision maker's expected utility.

In practice, this approach is unfeasible, because there are more than 2^n strategies for n unobserved features. To make computations tractable, both the old and new versions of Pathfinder employ the *myopic* approximation, introduced by Gorry and Barnett in 1968 [27]. In this approximation, a system identifies the best single feature to observe, by maximizing

the expected utility of the decision maker under the assumption that a diagnosis will be made after the user observes only one feature. Once the user observes the feature, the system repeats the myopic analysis, and may recommend additional features for observation.

Let us examine formally the computation in Pathfinder. First, the system computes $eu(dx(\phi)|\phi)$, the expected utility of the optimal diagnosis when the user observes no additional features. From Equations 7 and 8, we have

$$eu(dx(\phi)|\phi) = \sum_{d_j} p(d_j|\phi)\ u(d_j, dx(\phi)) \tag{9}$$

Now the system imagines that the user observes an additional feature f_{new}. Let ϕ' denote the union of the original set of observations and the observation for f_{new}. Pathfinder now identifies the optimal diagnosis, given the new set of observations:

$$dx(\phi') = \text{argmax}_{d_k}\left[\sum_{d_j} p(d_j|\phi')\ u(d_j, d_k)\right] \tag{10}$$

The expected utility of this diagnosis, denoted $eu(dx(\phi')|\phi')$, is given by

$$eu(dx(\phi')|\phi') = \sum_{d_j} p(d_j|\phi')\ u(d_j, dx(\phi')) \tag{11}$$

In contrast, the expected utility of the original diagnosis, given observations ϕ', is given by

$$eu(dx(\phi)|\phi') = \sum_{d_j} p(d_j|\phi')\ u(d_j, dx(\phi)) \tag{12}$$

The quantity $eu(dx(\phi)|\phi')$ is never greater than the measure $eu(dx(\phi')|\phi')$, because, by definition, the diagnosis $dx(\phi')$ maximizes expected utility, given the observations ϕ'. The system now computes the *value of information* of observing f_{new}, denoted $vi(f_{\text{new}}|\phi)$, which is the difference between $eu(dx(\phi')|\phi')$ and $eu(dx(\phi)|\phi')$ averaged over the instances i_{new} of the feature f_{new}.[3] That is,

$$vi(f_{\text{new}}|\phi) = \sum_{i_{\text{new}}} p(\phi'|\phi)\ \left[eu(dx(\phi')|\phi') - eu(dx(\phi)|\phi')\right] \tag{13}$$

The value of information of observing f_{new} represents the largest amount that the decision maker would be willing to pay to observe f_{new}. This quantity is always greater than or equal to 0. Next, the system computes the *net value of information* of observing f_{new}, denoted $nvi(f_{\text{new}}|\phi)$, by subtracting the cost[4] of observing f_{new} from the value of information of observing f_{new}. That is,

$$nvi(f_{\text{new}}|\phi) = vi(f_{\text{new}}|\phi) - cost(f_{\text{new}}). \tag{14}$$

Finally, if there is at least one feature that has a net value of information greater than 0, Pathfinder recommends the feature for observation that has the highest net value of

[3]This definition and the definition of net value of information are appropriate for expected-value decision makers. Howard discusses the general case [28, 29].

[4]We measured costs in dollars and then converted these costs to units of probability via Howard's model.

information. Otherwise, the system suggests that the user should gather no additional evidence and make a diagnosis.

In principle, the myopic approximation could affect the diagnostic accuracy of an expert system. For example, suppose that two features remain unobserved. In this case, the net value of information for the feature pair could exceed 0, and thus the user should observe the features. A value-of-information analysis on each feature alone, however, may indicate that neither feature is cost effective for observation. Consequently, the user would fail to observe these features, and thereby possibly make an incorrect diagnosis. Nonetheless, there is evidence that the myopic approximation does not often cause this problem in practice. For example, Gorry and Barnett have demonstrated that the approximation does not diminish significantly the diagnostic accuracy of their program that assists physicians with the diagnosis of congenital heart disease [27]. In addition, although we have not yet conducted a similar experiment with Pathfinder, our expert almost always has been impressed by the questions generated by the myopic approximation.

As we mentioned in previous sections, Pathfinder's diagnostic recommendations are sensitive to the utility model, and we therefore do not allow Pathfinder to make such recommendations. Fortunately, however, we have found in an informal study that Pathfinder's recommendations for evidence gathering are insensitive to the model. In particular, we have found that Pathfinder's recommendations often are similar to those made by a second version of the program in which $u(d_j, d_k)$ is equal to 1 when both d_j and d_k are benign diseases, $u(d_j, d_k)$ is equal to 1 when both d_j and d_k are malignant diseases, and $u(d_j, d_k)$ is equal to 0 otherwise. Consequently, we allow Pathfinder to make recommendations for evidence gathering.

The observation that recommendations for evidence gathering are less sensitive to the utility model than are diagnostic recommendations may be due to the fact that more factors contribute to the computation of net value of information than to the computation of the diagnosis. That is, only the probabilities of diseases and the utilities $u(d_j, d_k)$ contribute to the computation of the diagnosis, whereas these factors, the information content of a feature, and the cost of a feature contribute to the computation of net value of information.

5 Alternative Reasoning Methodologies

The first medical expert systems employed computations based in probability theory. In particular, throughout the 1960s and early 1970s, medical expert systems used the simple-Bayes model to construct differential diagnoses. These systems included Warner's system for the diagnosis of heart disease [30], Gorry's program for the management of acute renal failure, and deDombal's system for the diagnosis of acute abdominal pain [31].

Evaluations of most of these early systems showed that the programs performed well. In fact, the diagnoses rendered by several of them were more accurate than were those made by experienced physicians [31]. Nonetheless, in the early 1970s, researchers began to criticize these systems. They noted that the domains of these programs were small and did not reflect realistic clinical situations. Furthermore, researchers argued that errors due to the erroneous assumptions of the simple-Bayes model would become unacceptable as the domains of these systems were expanded [32, 33]. One group of investigators showed that

the diagnostic accuracy of an expert system based on the simple-Bayes model deteriorated significantly as the number of features in the system increased. These investigators traced the degradation in performance to violations of the conditional-independence assumptions in the simple-Bayes model [34]. Another group of researchers showed that the assumption of global conditional independence could be unrealistic in small domains as well [35].

In the early 1970s, perceptions of the inadequacy of the early decision-theoretic systems led to the development of alternative methods for reasoning under uncertainty [36, 37, 32]. Many of these developments occurred in the field of Artificial Intelligence in Medicine. Some of the alternative methods were *ad hoc* mechanisms, designed as custom-tailored techniques for particular domains or systems. These approaches included the MYCIN certainty-factor model and the QMR scoring scheme. Other methods were developed as alternative theoretical formalisms, such as the Dempster–Shafer theory of evidence and fuzzy decision theory.

In the first year of Pathfinder research, we appreciated the limitations of the simple-Bayes model, and believed that the use of this model would significantly impair the performance of Pathfinder. Consequently, we examined several alternative reasoning methodologies. In this section, we introduce these approaches. In the following two sections, we describe our empirical and theoretical analyses of these methodologies in the context of Pathfinder.

A well-known ad hoc method for managing uncertainty is the certainty-factor (CF) model [38]. Shortliffe and Buchanan designed the model to augment the rule-based approach to reasoning for MYCIN, a program for the diagnosis and treatment of bacteremias and meningitis [33]. In using the model, an expert attaches a certainty factor to each if–then rule. The certainty factor represents the expert's *change* of the belief in the consequent of the rule, given the antecedent of the rule. In particular, a CF between 0 and 1 means that the expert's belief in a consequent increases if the antecedent is true, whereas a CF between -1 and 0 means that the expert's belief decreases. In a rule base, the consequent of one rule may serve as the antecedent of another rule. In addition, two or more rules may share the same antecedent or consequent. As a result, a rule base forms an inference network: a directed graph in which an arc from proposition a to proposition b corresponds to the rule "if a then b." The CF model prescribes a method for propagating certainty factors through such a network. That is, given an observation of an antecedent in the network, we can use CF-model formulas to compute the effective certainty factor for any consequent in the network that is a descendent of that antecedent. Although the CF model was designed for MYCIN, the model has found many applications in other domains. Today, the model is the most popular method for managing uncertainty in rule-based systems.

Quick Medical Reference or QMR (formerly Internist-1) uses another ad hoc method for managing uncertainty [39, 40]. The QMR project, now in its eighteenth year, assists internists with the diagnosis of more than 600 diseases, through the consideration of approximately 4000 manifestations or features of disease. In QMR, each feature has two instances; in particular, a feature is either absent or present. More important, each disease can be either absent or present. Thus, QMR can address cases in which more than one disease is present. The ad hoc scoring scheme employs two measures to represent the degree of association between a feature and a disease: an evoking strength and a frequency. The *evoking strength* for a given feature–disease pair represents the degree to which the presence of the disease causes that feature to be present. The *frequency* for a given feature–disease pair represents how often that feature is present in patients who have the disease. In addition, the scheme

represents the *import* of each feature, which is inversely proportional to the likelihood that an insignificant disease (such as the common cold) can cause the feature to be present. Given an assignment of present and absent to a subset of features, QMR uses evoking strengths, frequencies, and imports to assign a score to each disease. QMR then displays diseases in order of descending score. Like early decision-theoretic systems, QMR uses a hypothetico-deductive approach to reasoning. In particular, the system contains ad hoc algorithms for generating useful recommendations for additional evidence gathering based on the current differential diagnosis and on evoking strengths, frequencies, and imports.

A more theoretical alternative to probabilistic reasoning was developed by Dempster and extended by Shafer [41, 42]. The approach, now called the Dempster–Shafer theory of evidence, was motivated by theoretical objections to the decision-theoretic approach [43]. Nonetheless, many artificial-intelligence (AI) researchers adopted special cases of the approach to avoid the perceived computational intractability of decision theory [44, 45]. Currently, the theory has many interpretations [42, 46, 47, 48]. One of the most popular interpretations is that given in Shafer's original text. In this interpretation, an expert assesses the degree of support that a piece of evidence lends to hypotheses in the *frame of discernment*: a set of mutually exclusive and exhaustive hypotheses. He does so for every piece of evidence that may be observed. A combination rule can then be used to compute the degree of support that multiple pieces of evidence lend to hypotheses in the frame of discernment. In the interpretation, an expert assesses degrees of support for a single piece of evidence by constructing a *basic probability assignment* over the frame. That is, he assigns a *mass*, ranging from 0 to 1, to each subset of hypotheses in the frame. The mass for a particular piece of evidence and a subset of hypotheses represents the degree of support that the evidence lends to the subset. Like the CF model and the QMR scoring scheme, the Dempster–Shafer theory manipulates measures of change in belief. In Section 7.1, we examine the relationship among these methodologies.

Another theoretical approach for managing uncertainty is fuzzy decision theory [49]. The theory addresses the presence of ambiguous terms such as "large" and "tall" in the specification of decision problems. Fuzzy decision theorists do not object to the use of probability theory or decision theory when events are defined precisely. They argue, however, that it is desirable to reason in situations where there is imprecision in the definition of events in addition to uncertainty about their occurrence. An example of a fuzzy decision problem is as follows:

> An urn contains many balls of various sizes, of which several are large. To draw a ball, you must pay a small sum of money. If you draw a large ball, however, you will win a valuable prize. Should you draw the ball?

6 Empirical Study of Reasoning Methods

During the first year of Pathfinder research, we experimented with the reasoning methodologies described in the previous section. In this section, we examine the results of those experiments.

6.1 Rule-Based Reasoning

The first version of Pathfinder was a rule-based system that employed propositional logic for reasoning. After informally evaluating the system, we discovered two related problems. First, the program did not take into account the uncertainty associated with the relationships between observations and diseases. This deficiency of the program became apparent to us almost immediately. Indeed, as we have mentioned, proper management of uncertainty is crucial to accurate diagnosis in the domain of lymph-node pathology. We might have addressed this concern with the use of the CF model. We found, however, another problem with the rule-based approach, which forced us to abandon the methodology. In particular, our expert was frustrated by the system, because it asked many questions that were irrelevant to discriminating the diseases on the current differential diagnosis. This behavior was a result of the fact that the rule-based methodology generated recommendations for additional observations based on a fixed traversal through the rule base.

As a result of our informal evaluations, we searched for a more flexible approach to the overall control of diagnostic reasoning. We discovered literature describing the hypothetico-deductive approach and systems such as QMR and Gorry's diagnostic program that implemented the approach. We decided to construct a new version of Pathfinder modeled after QMR. Nonetheless, we were not satisfied with the scoring scheme of QMR because it had no theoretical foundation; and we searched for a more principled method for managing uncertainty.

6.2 Fuzzy Reasoning

We considered fuzzy decision theory as a possible reasoning methodology for Pathfinder, but quickly rejected its use. We did so, because we found that neither general pathologists nor experts in hematopathology agreed on the meanings of fuzzy descriptions of feature–instance pairs such as "mild capsule thickening," "rare Sternberg-Reed cells," or "prominent necrosis." For example, one expert stated that Sternberg-Reed cells were "rare" when there were one to five of these cells in any 4-square-centimeter section of a lymph node. Another expert stated that these cells were "rare" when there were one to ten of these cells in any 4-square-centimeter section. We did not believe that fuzzy decision theory—a scheme devised by researchers who were unfamiliar with the domain of lymph-node pathology—nor any other mechanism would provide meaningful inferences, if we continued to employ these fuzzy feature-instance descriptions.

Instead, we asked the four hematopathology experts—Drs. Costan Berard, Jerome Burke, Ronald Dorfman, and Bharat Nathwani—to clarify the meanings of the descriptions that they were using. Although, as we have just discussed, the experts' interpretations did not coincide initially, the experts did not find it difficult to construct unambiguous interpretations for each feature instance. The experts handled disagreements in a manner similar to that used by coauthors of a manuscript who are faced with a disagreement. That is, when their initial interpretations of a feature instance did not coincide, each pathologist put forth an argument for the merits of his interpretation of that feature instance. Then, in most cases, the four experts accepted unanimously one interpretation. When the experts could not agree, the primary author of the system (Bharat Nathwani) selected the interpretation.

6.3 The Dempster–Shafer Theory of Evidence

Next, we examined the Dempster–Shafer theory of evidence. We discovered that the theory could be interpreted as a methodology for combining measures of change in belief, as could the CF model and QMR scoring scheme. We were attracted to the methodology, however, because it appeared to be a more principled approach to uncertainty management. Consequently, we constructed the second version of Pathfinder, using the Dempster–Shafer theory. In particular, we implemented a special case of the theory described by Barnett [44]. In this approach, the expert assigned nonzero masses only to (1) singleton subsets of the frame of discernment and (2) the entire frame of discernment. We refer to this simplified approach as the Dempster–Shafer–Barnett model.

We then evaluated informally this version of Pathfinder by allowing the expert to exercise the system with real and imaginary cases. The expert was satisfied with the diagnostic accuracy of the system.

At this time, probability theory was low on our list as a method for combining evidence to build a differential diagnosis, because of the limitations of the simple-Bayes model. Nevertheless, we were interested in experimenting with probabilistic reasoning, given the pioneering work of Ledley, Lusted, Gorry, Barnett, and others. We re-examined the measures of uncertainty that we had assessed from our expert, and realized that these measures could be interpreted in probabilistic terms. We implemented the simple-Bayes model in Pathfinder, without assessing additional measures of uncertainty.[5]

We then compared the performance of the Dempster–Shafer–Barnett and simple-Bayes versions of Pathfinder. Without informing the expert, we switched the scoring scheme of Pathfinder from the Dempster–Shafer–Barnett approach to the simple-Bayes model. To our surprise, after running several cases with the probabilistic scheme, the expert exclaimed excitedly that the diagnostic accuracy of the program had improved significantly.

Several years later, in a formal study, we compared the diagnostic accuracy of the Dempster–Shafer–Barnett, simple-Bayes, and CF models in the domain of lymph-node pathology. We verified our informal observation that the simple-Bayes model provided greater diagnostic accuracy (i.e., greater agreement with the expert) than did the Dempster–Shafer–Barnett model. We also found that the simple-Bayes model provided greater diagnostic accuracy than did the CF model [50].

7 Theoretical Study of Reasoning Methods

Surprised with the dominance of the simple-Bayes model over the alternative methods, we investigated the relationship between probability theory and alternative reasoning strategies over the next two years. We also studied the theoretical justifications for probabilistic and decision-theoretic reasoning.

[5] We assumed that the prior probability of each disease was equal.

7.1 Probabilistic Interpretations of Alternative Reasoning Methods

Heckerman and other researchers examined the relationship of the CF, QMR, and Dempster–Shafer–Barnett models with a simple probabilistic model for manipulating measures of change in belief called the *odds–likelihood updating scheme*. To understand the probabilistic model, let us suppose that we have a single disease d that can be true (d_+) or false (d_-). Further, suppose that we have n features $f_1, \ldots, f_n$, where each feature can be present or absent. Let us apply the simple-Bayes model to this situation. That is, let us assume that all features are conditionally independent, given d_+ and given d_-. Thus, as in Section 4.3, we can use Bayes' theorem to compute the posterior probability of d_+; we obtain

$$p(d_+|f_1, \ldots, f_n, \xi) = \frac{p(f_1|d_+, \xi) \cdots p(f_n|d_+, \xi)\ p(d_+|\xi)}{p(f_1|d_+, \xi) \cdots p(f_n|d_+, \xi)\ p(d_+|\xi)\ +\ p(f_1|d_-, \xi) \cdots p(f_n|d_-, \xi)\ p(d_-|\xi)} \tag{15}$$

where any f_i can be present or absent. In addition, we can apply Bayes' theorem to compute the posterior probability of d_-; we get

$$p(d_-|f_1, \ldots, f_n, \xi) = \frac{p(f_1|d_-, \xi) \cdots p(f_n|d_-, \xi)\ p(d_-|\xi)}{p(f_1|d_+, \xi) \cdots p(f_n|d_+, \xi)\ p(d_+|\xi)\ +\ p(f_1|d_-, \xi) \cdots p(f_n|d_-, \xi)\ p(d_-|\xi)} \tag{16}$$

When we divide Equation 15 by Equation 16, we obtain

$$\frac{p(d_+|f_1, \ldots, f_n, \xi)}{p(d_-|f_1, \ldots, f_n, \xi)} = \frac{p(f_1|d_+, \xi)}{p(f_1|d_-, \xi)} \cdots \frac{p(f_n|d_+, \xi)}{p(f_n|d_-, \xi)}\ \frac{p(d_+|\xi)}{p(d_-|\xi)} \tag{17}$$

We can rewrite Equation 17 as

$$O(d_+|f_1, \ldots, f_n, \xi) = \lambda(f_1, d_+|\xi) \cdots \lambda(f_n, d_+|\xi)\ O(d_+|\xi) \tag{18}$$

where

$$O(d_+|\xi) = \tfrac{p(d_+|\xi)}{p(d_-|\xi)} \qquad \text{and} \qquad O(d_+|f_1, \ldots, f_n, \xi) = \tfrac{p(d_+|f_1, \ldots, f_n, \xi)}{p(d_-|f_1, \ldots, f_n, \xi)} \tag{19}$$

are the prior and posterior *odds* of d_+, respectively, and

$$\lambda(f_i, d_+|\xi) = \frac{p(f_1|d_+, \xi)}{p(f_1|d_-, \xi)} \tag{20}$$

is the *likelihood ratio* for d_+, given f_i. Equation 18 is the odds–likelihood updating scheme.

Heckerman showed that we can interpret the certainty factor for the rule "if f_i then d_+," denoted $CF(f_i \rightarrow d_+|\xi)$ as a monotonically increasing function of the likelihood ratio $\lambda(f_i, d_+|\xi)$. In particular, he showed that, if we make the identification

$$CF(f_i \rightarrow d_+|\xi) = \begin{cases} \frac{\lambda(f_i, d_+|\xi) - 1}{\lambda(f_i, d_+|\xi)} & \lambda(f_i, d_+|\xi) \geq 1 \\ \lambda(f_i, d_+|\xi) - 1 & \lambda(f_i, d_+|\xi) < 1 \end{cases} \tag{21}$$

then the odds–likelihood updating scheme is identical to the formula in the CF model for combining certainty factors that is applied when a set of rules share the same consequent.

In addition, Heckerman showed that with the identification in Equation 21, the remaining formulas in the CF model are a close approximation to the rules of probability [51].

Grosof then showed that the Dempster–Shafer–Barnett model was isomorphic to the odds–likelihood updating scheme [52] via a different transformation of the likelihood ratio. In addition, Heckerman and Miller demonstrated that QMR's ad hoc scoring scheme was isomorphic to the odds–likelihood updating scheme [53].[6]

These theoretical results helped us to understand the dominance of the simple-Bayes model over the nonprobabilistic alternatives. In particular, the other approaches did not avoid the assumptions of conditional independence of the simple-Bayes model; they merely obscured the assumptions. In fact, these approaches assumed that evidence was conditionally independent, given each disease and given the *negation* of each disease. When there are more than two mutually exclusive and exhaustive diseases in a domain, theses conditional independence assumptions are stronger than are the assumptions in the simple-Bayes model [58, 51].

We can understand the limitations of the alternative scoring schemes at a more intuitive level. Let us consider rule-based inference, in particular. The first rule-based inference schemes used the rules of logic. As a result, these schemes enjoyed a property known as *modularity.* That is, given the logical rule "if a then b," and given that a is true, we can assert that b is true no matter how we established that a is true, and no matter what else we know to be true. For example, given the rule

$$\text{if } l_1 \text{ and } l_2 \text{ are parallel lines then } l_1 \text{ and } l_2 \text{ do not intersect}$$

we can assert that l_1 and l_2 do not intersect once we know that l_1 and l_2 are parallel lines. This assertion satisfies the property of modularity: The assertion depends on neither how we came to know that l_1 and l_2 are parallel, nor what else we know.

The CF model is an extension of the rules of logic that imposes this same principle of modularity on inferences. For example, given the rule

$$\text{if PERITONITIS then APPENDICITIS, } CF = 0.7$$

and given that a patient has peritonitis, the CF model allows us to increase the likelihood that the patient has appendicitis by the amount corresponding to a CF of 0.7, no matter how we establish that peritonitis is present. Given the correspondences described in the first part of this section, we see that the odds–likelihood updating scheme, the QMR scoring scheme, and the Dempster–Shafer–Barnett model also incorporate the property of modularity. We shall refer to these methods collectively as *modular belief updating schemes.*

Unfortunately, these schemes in reality do not satisfy the property of modularity. Continuing our example, suppose the patient has vaginal bleeding. This fact increases the likelihood that she has a ruptured ectopic pregnancy, and thus increases the likelihood that she has peritonitis. The chances that the patient has an appendicitis decreases, however, because the presence of a ruptured ectopic pregnancy can account for the presence of peritonitis. Overall, we have that the likelihood of peritonitis increases, whereas the likelihood of appendicitis decreases. The modular rule linking peritonitis with appendicitis is inconsistent with these relationships.

[6]This work led to the construction of a probabilistic version of QMR, called QMR-DT [54, 55, 56, 57].

In general, logical relationships represent complete models of interaction. In contrast, uncertain relationships encode invisible interactions. We summarize these hidden interactions with numerical measures, such as a certainty factor or likelihood ratio. In the process of such a summarization, we lose information about the detailed categorical interaction. Therefore, when we try to combine uncertain information, unexpected (nonmodular) interactions may occur. We should not expect that any modular belief updating scheme will be able to handle such subtle interactions.

7.2 A Practical Problem with Nonprobabilistic Methods

In continuing to explore the difference between probabilistic and nonprobabilistic alternatives, we encountered a practical limitation associated with the use of modular belief updating schemes. Specifically, these schemes require that we assess the strength of an uncertain relationship in the direction in which it is used. That is, we must specify the change in belief of an unobservable hypothesis, given an observable piece of evidence. Unfortunately, experts often are more comfortable quantifying uncertain relationships in the direction opposite to that in which they are used [59].

In particular, Kahneman and Tversky have shown that people usually are more comfortable when they assess the likelihood of an effect given a cause rather than when they assess the likelihood of a cause given an effect. For example, expert physicians prefer to assess the likelihood of a finding, given a disease, rather than the likelihood (or belief update) of a disease, given a finding [60]. Henrion attributes this phenomenon to the nature of causality. In particular, he notes that a predictive probability (the likelihood of a finding, given a disease) reflects a stable property of that disease. In contrast, a diagnostic probability (the likelihood of a disease, given a finding) depends on the incidence rates of that disease and of other diseases that may cause the finding. Thus, predictive probabilities are a more useful and parsimonious way to represent uncertain relationships—at least in medical domains (see [61], pages 252–3). The developers of QMR make a similar observation [39].

Unfortunately, effects are usually the observable pieces of evidence, and causes are the sought-after hypotheses. Thus, in using a modular belief updating scheme, we force experts to provide judgments of uncertainty in a direction that is more cognitively challenging. We thereby promote errors in assessment. In contrast, when we use probability theory to manage uncertainty, we can assess the strength of an uncertain relationship in one direction, and then reverse the relationship using Bayes' theorem, when the need arises.

7.3 Compelling Principles for Uncertainty Management and Decision Making

After identifying limitations of non-decision-theoretic approaches for uncertain reasoning, we explored theoretical advantages of the decision-theoretic approach. Perhaps the most significant advantage we discovered was the fact that the rules of probability and the MEU principle follow from compelling principles, and that people often violate these principles when unaided by decision-theoretic systems.

For example, Ramsey and deFinetti showed that anyone who does not follow the rules of probability theory would be willing to accept a "Dutch book": a combination of bets

leading to a guaranteed loss of money under any circumstances [62, 63, 64]. In contrast, Cox developed a set of desiderata about fundamental properties of a measure of belief unrelated to betting behavior that also imply the rules of probability [65, 25]. In addition, von Neumann and Morgenstern constructed a remarkable proof of the MEU principle [20]. They developed five compelling principles or desiderata that every decision maker should follow, and demonstrated that these desiderata imply the MEU principle. One desideratum is the principle of transitivity, which states that if a decision maker prefers outcome A to outcome B, and prefers outcome B to outcome C, then he must prefer outcome A to outcome C. To see that this desideratum is compelling, let us suppose that a decision maker's preferences are not transitive. In particular, suppose he prefers A to B, B to C, and C to A. Because he prefers C to A, he should be willing to exchange A for C and a small payment. Similarly, this person should be willing to exchange C for B, and B for A. Thus, we can extract payments from him, and yet leave him with the same outcome. Repeating this procedure, called a *money pump*, we can extract an arbitrarily large payment from this person. Decision theorists and cognitive psychologists have devised justifications for each of von Neumann and Morgenstern's desiderata [20, 66, 67].

Psychological studies have demonstrated that, in the real world, human decision makers exhibit a set of stereotypical deviations or *biases* from the desiderata of decision theory [68, 69]. From the perspective of decision theory, we can view these deviations as mistakes. For example, people sometimes have nontransitive preferences, and thus are vulnerable to a money pump. Also, physicians often forget to consider the prior probability of disease when making a diagnosis. As we mentioned in the introduction, we use the adjectives "normative" and "descriptive" to emphasize the differences between decision making consistent with the rules of decision theory and decision making in the real world.

In medicine, descriptive errors in decision making are particularly dangerous, given the high stakes associated with some decisions. Such errors easily may lead to needless expenditures, pain and suffering, or loss of life. Fortunately, normative expert systems can help physicians to avoid such errors. For example, systems that encode explicitly the prior probabilities of diseases will likely improve the decisions made by physicians who would otherwise forget to consider this important information.

Nonetheless, several researchers have argued that we should employ descriptive rather than normative methods for decision making in expert systems. In particular, investigators have stated that decision-theoretic methods lack the *expressiveness* needed to encode expertise or to describe intelligent behavior [70, 37, 71, 43]. Indeed, several of these investigators have argued that the CF model, fuzzy decision theory, and the Dempster–Shafer theory of evidence may be more appropriate than is decision theory for uncertainty management, because these non-decision-theoretic methods are possibly more descriptive than is decision theory [71, 43, 72]. To our knowledge, however, there are no psychological studies that support these assertions. In fact, our experiment comparing the diagnostic accuracy of the simple-Bayes, CF, and Dempster–Shafer–Barnett models (see Section 6.3) demonstrated that, among these approaches, the simple-Bayes model is the most *descriptive* method for managing uncertainty.

8 A Return to Decision Theory: Practical Considerations

Once we understood the limitations of non-decision-theoretic methods and the theoretical benefits of decision theory for managing uncertainty, we still faced the practical limitations of probability and decision theory. In particular, we were concerned that it would be unfeasible to relax the conditional independence assumptions of the simple-Bayes model for Pathfinder. Nonetheless, we speculated that adding the most salient conditional dependencies would not lead to a combinatorial explosion. Indeed, we conjectured that experts themselves could not appreciate all the subtle conditional dependencies that may exist among the features in large medical domains. To manage the complexity of their domain, these experts must be reasoning under many assumptions of conditional independence. We believed that, if we could capture these assumptions made by the experts, then we could produce normative expert systems that perform at least as well as do experts.

Our work to achieve these goals in the domain of lymph-node pathology was successful. Over the next three years, we developed graphical knowledge representations that allowed us to capture the important conditional dependencies in this domain in a reasonable amount of time. In a formal evaluation, we demonstrated that the diagnostic accuracy of the new version of Pathfinder was at least as good as that of the Pathfinder expert [73]. In the companion to this article, we describe these new knowledge representations in detail.

9 Pathfinder in Clinical Practice

Several years after the Pathfinder project was initiated, we re-engineered the program on MS-DOS computers, and made the system available commercially. The system, named Intellipath, consists of a normative expert system for lymph-node diagnosis and a set of supportive informational tools including an analog videodisc library of images, text information on diseases and microscopic features, references to the literature, and an automated report-generator.

In 1988, the American Society of Clinical Pathologists began selling the Intellipath program to practicing pathologists and pathologists in training in North America. The program was a commercial success, and we constructed similar systems for other types of human tissue. Currently, approximately 200 pathologists are using the program, and systems for breast, bone, larynx, skin, small intestine, stomach, thymus, and urinary bladder pathology and for lung and thyroid cytology are available.

In Section 2, we mentioned that pathologists have difficulty with diagnosis because they are unable to combine evidence accurately and because they misrecognize or fail to recognize microscopic features. Most of Pathfinder research addresses the first problem; the videodisc component of Intellipath, however, addresses the second problem. In particular, when a user has trouble recognizing a feature, he can ask Intellipath to display images of that feature. These images illustrate both typical and atypical presentations of the feature.

Recently, we have integrated the latest version of the Pathfinder expert system with the Intellipath platform, including the videodisc for lymph-node pathology. We will evaluate this program, called Pathfinder II, in clinical trials funded by the National Cancer Institute.

In these studies, we shall determine whether pathologists who use the system perform better than do those pathologists who do not have access to the system. We shall quantify separately improvements in pathologists' ability to combine evidence and improvements in their ability to recognize features. In addition, we shall determine whether pathologists can recognize situations in which they need assistance from an expert system or human expert.

10 Conclusions

We are not alone in the investigation of probabilistic and decision-theoretic methods for uncertain reasoning in medical expert systems. Several other recent research projects have explored normative reasoning in medical expert systems, including the Nestor system for diagnosis of endocrinology disorders [74], the Glasgow Dyspepsia expert system for assisting in gastroenterology diagnosis [75], the Neurex system for diagnosis of neurological disorders [76], the Medas system for assisting physicians in emergency medicine [77], the Munin system for diagnosis of muscular disorders [78], and the Sleep Consultant system for diagnosis of sleep disorders [79]. Furthermore, over the last five years, a dedicated community of researchers addressing the management of uncertainty in reasoning systems has evolved. The proceedings of the main conference of this group of researchers, the *Conference on Uncertainty in Artificial Intelligence,* is a collection of the latest theoretical and empirical research on this topic [80, 81, 82, 83, 84].

We believe that the development of normative expert systems will lead to improvements in the capture and delivery of expert knowledge. This view is supported by our successes with Pathfinder. 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, Keung-Chi Ng, 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. Henri Suermondt, Mark Fischinger, Marty Chavez, and especially Keung-Chi Ng assisted with programming and data management. Lyn Dupré, Greg Cooper, and Keung-Chi Ng provided useful comments on this manuscript.

Research on the Pathfinder project has been supported by the National Cancer Institute under Grant RO1CA51729-01A, and by the National Library of Medicine under Grant RO1LM04529. Computational support has been provided in part by the SUMEX-AIM resource under National Institutes of Health Grant LM05208.

References

[1] 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, Silver Spring, MD, November 1989.

[2] D.E. Heckerman, E.J. Horvitz, and B.N. Nathwani. The Pathfinder project. Technical Report KSL-90-08, Medical Computer Science Group, Section on Medical Informatics, Stanford University, Stanford, CA, February 1990.

[3] 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.

[4] B.N. Nathwani, D.E. Heckerman, E.J. Horvitz, and T.L. Lincoln. Integrated expert systems and videodisc in surgical pathology: An overview. *Human Pathology*, 21:11–27, 1990.

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

[8] 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, 1980.

[9] 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.

[10] 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 lymphoma clinical studies: A comprehensive analysis of cases accumulated since its inception. *Journal of the National Cancer Institute*, 68:43–67, 1982.

[11] F.D. Bartlett. *Thinking*. Basic Books, New York, 1958.

[12] 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.

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

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

[15] I. Hacking. *The Emergence of Probability*. Cambridge University Press, New York, 1975.

[16] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Biometrika*, 46:293–298, 1958. Reprint of original work of 1763.

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

[18] J. Pearl. How to do with probabilities what people say you can't. In *Proceedings of the Second Conference on Artificial Intelligence Applications,* Miami Beach, FL, pages 6–12. IEEE Computer Society Press, Silver Spring, MD, December 1985.

[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, New York, 1986. North-Holland.

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

[21] R.L. Keeney and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Wiley and Sons, New York, 1976.

[22] 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.

[23] R.A. Howard. On making life and death decisions. In R.C. Schwing and W.A. Albers, Jr., editors, *Societal Risk Assessment*, pages 89–113. Plenum Publishing, New York, 1980.

[24] R.A. Howard. Microrisks for medical decision analysis. *International Journal of Technology Assessment in Health Care*, 5:357–370, 1989.

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

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

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

[28] R.A. Howard. Value of information lotteries. *IEEE Transactions of Systems Science and Cybernetics*, SSC-3(1):54–60, 1967.

[29] R.A. Howard. Decision Analysis, EES231 Course Notes. Department of Engineering–Economic Systems, Stanford University, Stanford, CA, 1985.

[30] 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.

[31] 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.

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

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

[34] D.G. Fryback. Bayes' theorem and conditional nonindependence of data in medical diagnosis. *Computers and Biomedical Research*, 11:423–434, 1978.

[35] M.J. Norusis and J.A. Jacquez. Diagnosis 1: Symptom nonindependence in mathematical models for diagnosis. *Computers and Biomedical Research*, 8:156–172, 1975.

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

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

[38] E.H. Shortliffe and B.G. Buchanan. A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23:351–379, 1975.

[39] 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.

[40] 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.

[41] A.P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statistics*, 38:325–339, 1967.

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

[43] G. Shafer. Savage revisited. *Statistical Science*, 1:463–501, 1986.

[44] J. A. Barnett. Computational methods for a mathematical theory of evidence. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence,* Vancouver, BC, pages 868–875. International Joint Conference on Artificial Intelligence, August 1981.

[45] J. Gordon and E.H. Shortliffe. A method for managing evidential reasoning in a hierarchical hypothesis space. *Artificial Intelligence*, 26:323–357, 1985.

[46] G. Shafer. Probability judgment in artificial intelligence. In *Proceedings of the Workshop on Uncertainty and Probability in Artificial Intelligence,* Los Angeles, CA, pages 91–98. Association for Uncertainty in Artificial Intelligence, Mountain View, CA, August 1985. Also in Kanal, L. and Lemmer, J., editors, *Uncertainty in Artificial Intelligence,* pages 127–136. North-Holland, New York, 1986.

[47] R. Hummel and M. Landy. Evidence as opinions of experts. In *Proceedings of the Second Workshop on Uncertainty in Artificial Intelligence,* Philadelphia, PA, pages 135–143. Association for Uncertainty in Artificial Intelligence, Mountain View, CA, August 1986. Also in Kanal, L. and Lemmer, J., editors, *Uncertainty in Artificial Intelligence 2,* pages 43–53. North-Holland, New York, 1988.

[48] P. Smets. Constructing the pignistic probability function in a context of uncertainty. In *Proceedings of the Fifth Workshop on Uncertainty in Artificial Intelligence,* Windsor, ON, pages 319–326. Association for Uncertainty in Artificial Intelligence, Mountain View, CA, August 1989. Also in Henrion, M., Shachter, R., Kanal, L., and Lemmer, J., editors, *Uncertainty in Artificial Intelligence 5,* pages 29–39. North-Holland, New York, 1990.

[49] L. Zadeh. Is probability theory sufficient for dealing with uncertainty in AI: A negative view. In L.N Kanal. and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 103–116. North-Holland, New York, 1986.

[50] D.E. Heckerman. An empirical comparison of three inference methods. In *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence,* Minneapolis, MN, pages 158–169. Association for Uncertainty in Artificial Intelligence, Mountain View, CA, August 1988. Also in Shachter, R., Levitt, T., Kanal, L., and Lemmer, J., editors, *Uncertainty in Artificial Intelligence 4,* pages 283–302. North-Holland, New York, 1990.

[51] D.E. Heckerman. Probabilistic interpretations for MYCIN's certainty factors. In *Proceedings of the Workshop on Uncertainty and Probability in Artificial Intelligence,* Los Angeles, CA, pages 9–20. Association for Uncertainty in Artificial Intelligence, Mountain View, CA, August 1985. Also in Kanal, L. and Lemmer, J., editors, *Uncertainty in Artificial Intelligence,* pages 167–196. North-Holland, New York, 1986.

[52] B. Grosof. Evidential confirmation as transformed probability. In *Proceedings of the Workshop on Uncertainty and Probability in Artificial Intelligence,* Los Angeles, CA, pages 185–192. Association for Uncertainty in Artificial Intelligence, Mountain View, CA, August 1985. Also in Kanal, L. and Lemmer, J., editors, *Uncertainty in Artificial Intelligence,* pages 153–166. North-Holland, New York, 1986.

[53] 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.

[54] M. Henrion. Towards efficient probabilistic inference in multiply connected belief networks. In R.M. Oliver and J.Q. Smith, editors, *Influence Diagrams, Belief Nets, and Decision Analysis*, chapter 17. Wiley and Sons, New York, 1990.

[55] D.E. Heckerman. A tractable algorithm for diagnosing multiple diseases. In *Proceedings of the Fifth Workshop on Uncertainty in Artificial Intelligence,* Windsor, ON, pages 174–181. Association for Uncertainty in Artificial Intelligence, Mountain View, CA, August 1989. Also in Henrion, M., Shachter, R., Kanal, L., and Lemmer, J., editors, *Uncertainty in Artificial Intelligence 5,* pages 163–171. North-Holland, New York, 1990.

[56] M. Shwe, B. Middleton, D. Heckerman, M. Henrion, E. Horvitz, H. Lehmann, and G. Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base: Part I. the probabilistic model and inference algorithms. *Methods in Information and Medicine*, 30:241–250, 1991.

[57] B. Middleton, M. Shwe, D. Heckerman, M. Henrion, E. Horvitz, H. Lehmann, and G. Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base: Part II. evaluation of diagnostic performance. *Methods in Information and Medicine*, 30:256–267, 1991.

[58] R. Johnson. Independence and Bayesian updating methods. In *Proceedings of the Workshop on Uncertainty and Probability in Artificial Intelligence,* Los Angeles, CA, pages 28–30. Association for Uncertainty in Artificial Intelligence, Mountain View, CA, August 1985. Also in Kanal, L. and Lemmer J., editors, *Uncertainty in Artificial Intelligence,* pages 197–201. North-Holland, New York, 1986.

[59] R.D. Shachter and D.E. Heckerman. Thinking backward for knowledge acquisition. *AI Magazine*, 8:55–63, 1987.

[60] A. Tversky and D. Kahneman. Causal schemata in judgments under uncertainty. In D. Kahneman, P. Slovic, and A. Tversky, editors, *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York, 1982.

[61] 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.

[62] F.P. Ramsey. Truth and probability. In R.B. Braithwaite, editor, *The Foundations of Mathematics and other Logical Essays*. Humanities Press, London, 1931. Reprinted in Kyburg and Smokler, 1964.

[63] B. de Finetti. La prévision: See lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1–68, 1937. Translated in Kyburg and Smokler, 1964.

[64] H.E. Kyburg and H.E. Smokler. *Studies in Subjective Probability*. Wiley and Sons, New York, 1964.

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

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

[67] R.A. Howard. Risk preference. In R.A. Howard and J.E. Matheson, editors, *Readings on the Principles and Applications of Decision Analysis*, volume II, pages 629–663. Strategic Decisions Group, Menlo Park, CA, 1970.

[68] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185:1124–1131, 1974.

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

[70] G.A. Gorry, J.P. Kassirer, A. Essig, and W.B. Schwartz. Decision analysis as the basis for computer-aided management of acute renal failure. *American Journal of Medicine*, 55:473–484, 1973.

[71] 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.

[72] L. Zadeh. Personal communication. 1986.

[73] D. Heckerman and B. Nathwani. An evaluation of the diagnostic accuracy of Pathfinder. *Computers and Biomedical Research*, In press.

[74] G.F. Cooper. *NESTOR: A Computer-based Medical Diagnostic Aid that Integrates Causal and Probabilistic Knowledge*. PhD thesis, Medical Computer Science Group, Stanford University, Stanford, CA, November 1984. Report HPP-84-48.

[75] 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.

[76] 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.

[77] 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.

[78] 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 Kaufmann, San Mateo, CA, August 1987.

[79] G. Nino-Murcia and M. Shwe. An expert system for diagnosis of sleep disorders. In M. Chase, R. Lydic, and C. O'Connor, editors, *Sleep Research*, volume 20, page 433. Brain Information Service, Los Angeles, CA, 1991.

[80] L. Kanal and J. Lemmer, editors. *Uncertainty in Artificial Intelligence*. North-Holland, New York, 1986.

[81] J. Lemmer and L. Kanal, editors. *Uncertainty in Artificial Intelligence 2*. North-Holland, New York, 1988.

[82] L. Kanal, T. Levitt, and J. Lemmer, editors. *Uncertainty in Artificial Intelligence 3*. North-Holland, New York, 1989.

[83] R. Shachter, T. Levitt, L. Kanal, and J. Lemmer, editors. *Uncertainty in Artificial Intelligence 4*. North-Holland, New York, 1990.

[84] M. Henrion, R. Shachter, L. Kanal, and J. Lemmer, editors. *Uncertainty in Artificial Intelligence 5*. North-Holland, New York, 1990.