An Evaluation of the Diagnostic Accuracy of Pathfinder*

David E. Heckerman

Departments of Computer Science and Pathology

Bharat N. Nathwani

Department of Pathology

University of Southern California

HMR 204, 2025 Zonal Ave

Los Angeles, CA 90033

November 27, 1991

*This work is an adaptation of Heckerman (1991). All figures and tables are printed with permission from MIT Press.

Abstract

We present an evaluation of the diagnostic accuracy of Pathfinder, an expert system that assists pathologists with the diagnosis of lymph-node diseases. We evaluate two versions of the system using both informal and decision-theoretic metrics of performance. In one version of Pathfinder, we assume incorrectly that all observations are conditionally independent. In the other version, we use a belief network to represent accurately the probabilistic dependencies among the observations. In both versions, we make the assumption—reasonable for this domain—that diseases are mutually exclusive and exhaustive. The results of the study show that (1) it is cost effective to represent probabilistic dependencies among observations in the lymph-node domain, and (2) the diagnostic accuracy of the more complex version of Pathfinder is at least as good as that of the Pathfinder expert. In addition, the study illustrates how informal and decision-theoretic metrics for performance complement one another.

1 Introduction

Pathfinder is an expert system that assists community pathologists with the diagnosis of lymph-node pathology (Heckerman et al., 1989; Heckerman et al., 1990). The Pathfinder project began in 1983 as a joint project among researchers at Stanford University and the University of Southern California, including David Heckerman, Bharat Nathwani, Eric Horvitz, and Larry Fagan (Heckerman et al., 1985). Currently, a commercial version of Pathfinder, called Intellipath, is being used by practicing pathologists and by pathologists in training as a teaching and learning tool (Nathwani et al., 1990).

The domain of Pathfinder is large by any standard of comparison for expert systems. Over 60 diseases can invade the lymph node (25 benign diseases, 9 Hodgkin's lymphomas, 18 non-Hodgkin's lymphomas, and 10 metastatic diseases). In addition, there are over 100 morphologic distinctions or *features* within lymph nodes that can be easily recognized on microscopic examination. Each feature is associated with two or more mutually exclusive and exhaustive *instances*. Also, Pathfinder contains over 30 features reflecting clinical, laboratory, immunological, and molecular biological information that are relevant to the diagnosis of lymph-node disease.

Pathology diagnosis, like diagnosis in other areas of medicine, involves significant uncertainties. Pathfinder uses subjective or Bayesian probability theory to manage these uncertainties. In particular, expert knowledge is encoded in a belief network (Howard and Matheson, 1981; Pearl, 1988), and system recommendations are derived from inferences applied to this network.

In this article, we compare the diagnostic accuracy of two versions of Pathfinder. In one version, called Pathfinder 1, we make the simplifying assumption that features are conditionally independent, given disease. In another version, called Pathfinder 2, we encode the dependencies among features. We compare the two versions of Pathfinder in a three-phase experiment. In the first phase, we ask the question: Is the diagnostic accuracy of Pathfinder 2 greater than that of Pathfinder 1? In the second phase, we ask: What factors are responsible for improvement, if any? In the third phase, we ask: Are the improvements worth the effort of building the more sophisticated version of Pathfinder? In the first phase, we use a

simple, informal measure of diagnostic accuracy. In the third phase, we use a more complex measure of diagnostic accuracy based on decision theory. Specifically, for each version of Pathfinder, we measure the difference between the expected utility of the patient who is given a gold-standard diagnosis, and the expected utility of the same patient who is given Pathfinder's diagnosis.

The results show that (1) the diagnostic accuracy of Pathfinder 2 is greater than that of Pathfinder 1, (2) the dominant factor leading to improvement is that the Pathfinder expert provided better probability assessments for a feature when he was allowed to condition the assessments on the observation of other features, and (3) the benefits of the improvements outweigh the costs of constructing the more sophisticated knowledge base. Also, in phase 3, we find that the diagnostic accuracy of Pathfinder is at least as good as that of the Pathfinder expert.

The results of this study have general significance. First, there has been much debate as to whether or not probabilistic representations are practical for large, real-world expert systems (Gorry, 1973; Shortliffe and Buchanan, 1975; Patil, 1981; Henrion and Cooley, 1987; Heckerman, 1991). This study shows that probabilistic representations can be cost effective for complex domains.

Second, many researchers have evaluated expert systems using informal measures of performance (e.g., see de Dombal et al., 1972, Buchanan and Shortliffe, 1984, and Cooper, 1984); several researchers have identified systematically factors responsible for improvements in performance (Miller et al., 1982); and several researches have used decision theory to quantify the diagnostic accuracy of a system (Smets et al., 1975; Asselain et al., 1977; Habbema and Hilden, 1981). In this study, we show that the three phases of evaluation are complementary. We suggest that other evaluations of expert systems, including evaluations of the clinical effectiveness of expert systems, follow this three-phase approach.

2 The Systems: Pathfinder 1 and 2

Pathfinder provides a differential diagnosis—a probability distribution over diseases—given a set of set of observations made by the user, and suggests to the user additional features for

observation that are likely to narrow the differential diagnosis. To accomplish these tasks, the expert system uses a hypothetico-deductive approach (Bartlett, 1958; Elstein et al., 1971; Elstein, 1976; Elstein et al., 1978). First, when evaluating sections of a lymph node, a user enters a set of salient morphologic observations, and Pathfinder uses Bayesian inference to compute the posterior probabilities of disease, given these observations. Next, the program computes the *value of perfect information* or the *value of clairvoyance* for each feature that has not yet been reported to the system (Howard, 1967); and displays those features with the highest scores to the user. To avoid a combinatorial explosion, the value-of-information computation employs the assumption that at most one additional feature will be reported by the user. Then, the user observes one or more of these features, and the process iterates. For a detailed description of the operation of Pathfinder, see Heckerman (1991).

When building Pathfinder 1, we employed what is commonly refered to as the *idiot-Bayes model*. Specifically, we assumed that diseases were mutually exclusive and exhaustive, and that all features were conditionally independent, given disease. The assumption that diseases were mutually exclusive was 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 made reasonable the assumption that the set of diseases was exhaustive. The assumption of global conditional independence, however, was 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. Thus, we constructed a second version of the system, called Pathfinder 2, in which we represented accurately the dependencies among the features.

As we mentioned in Section 1, we used a belief network to represent these dependencies. The Pathfinder belief network is shown in Figure 1. In a belief network, nodes represent uncertain variables, and arcs represent probabilistic dependencies among the variables. Each node is associated with a set of mutually exclusive and exhaustive instances. Each node is also associated with a set of probability distributions. An arc from node x to node y means that the probability distributions for y may depend on the instances of x. Conversely, the absence of arcs reflects assertions of conditional independence. Given any belief network, we can construct a joint probability distribution for the variables in the network from the proba-

bility distributions associated with the nodes and the assertions of conditional independence encoded by the network.

In the Pathfinder belief network, there is one node labeled DISEASE. Each instance of this node corresponds to a specific disease. This representation is possible, because we have assumed that the diseases are mutually exclusive and exhaustive. If this assumption were invalid, we may represent each disease as a separate node in the influence diagram. Let us consider the node MITOTIC FIG depicted as the bottom center of Figure 1, which represents the number of mitotic figures per unit area in the nonfollicular areas of the lymph node. The only arc pointing to this node is directed from the node DISEASE. Thus, the node MITOTIC FIG is associated only with the probability distributions $p(\text{MITOTIC FIG}|d_i)$—one distribution for each disease d_i.[1] The absence of other arcs pointing to MITOTIC FIG represents the expert's belief that, given the identity of the patient's disease, MITOTIC FIG is conditionally independent of all other features.

We could not construct directly the belief network for lymph-node diagnosis. Because of the wide scope of the domain, we were uncomfortable assessing conditional independencies among some features. As a result, the first author developed a representation called a *similarity network* (Heckerman, 1990; Heckerman, 1991). Using this representation, a user builds many belief networks for small differential-diagnosis tasks. These small beliefs networks are then combined with a simple algorithm to form a large belief network for the entire domain. We constructed the belief network shown in Figure 1 using a similarity network. As is indicated in Table 1, the construction required approximately 35 hours.

Insert Figure 1 about here.

The first author also developed a representation called a *partition* that facilitates the assessment of probability distributions associated with nodes in a belief network. As is shown in Table 1, the Pathfinder belief network contains 74,854 probabilities. Using the partition representation, however, we computed 61,118 of those probabilities from other probabilities. Therefore, the expert needed to assess only 13,736 probabilities. Overall, the construction

[1] In this article, we use $p(x|y_i)$ to denote the probability distribution over variable x, given an instance i of variable y.

of Pathfinder 1 required approximately 48 hours, whereas the construction of Pathfinder 2 required approximately 85 hours. In phase 3 of this evaluation, we will compare these costs with the benefits derived from the two versions of Pathfinder.

Insert Table 1 about here.

3 Selection of Cases

A set of cases for the study was selected in sequence from a large library of cases that had been referred to the second author (the Pathfinder expert) from community pathologists. Because such cases were referrals, they were likely to be at least as difficult as cases in which nonexperts would seek the help of a computer aid. Cases were rejected only if glass slides were unavailable or if the case did not involve lymph-node tissue. Sections that were poorly stained or improperly sliced were not excluded. Over 100 cases were selected; because of time constraints, however, the experiment was conducted on only the first 53 cases.

4 Entry of Features

A community pathologist entered features observed in each case into both Pathfinder 1 and 2. She entered only morphologic features (i.e., features observed through a microscope); she did not perform tests that were expensive or that would have caused significant delays in the experiment. She was allowed to see the recommendations for additional observations made by both systems if she was unsure about what feature to enter next. Also, if she was unsure about the identification of a feature, she was allowed to access a library of over 4000 video images that illustrates the morphologic features that can be reported to the two systems.

The pathologist chosen for the experiment was recently a fellow in hematopathology with the second author. She was selected because she was familiar with the lymph-node domain and with most of the terminology used by Pathfinder.

5 Phase 1: An Informal Metric

In phase 1 of the experiment, we wanted to determine whether the diagnostic accuracy of Pathfinder 2 was greater than that of Pathfinder 1. For each case, the second author (the expert) was shown the features reported by the nonexpert, as well as the probability distributions produced by the two versions of the system. The expert was blinded as to the identity of the distributions, and the distributions were displayed in random order. For each probability distribution, the second author was asked, "On a scale from zero to ten—zero being unacceptable and ten being perfect—how accurately does the distribution reflect your beliefs?"

The mean and standard deviation of the expert ratings for Pathfinder 1 and Pathfinder 2 are shown in Table 2. Case-by-case results are given in Heckerman (1991), Appendix D. The experiment reveals a significant difference between the two systems. Specifically, a bootstrap permutation test (Diaconis and Efron, 1983) yields an achieved significance level (ASL) of 0.007. The permutation test indicates that there is only a 0.007 chance that a more extreme result would be obtained if data were drawn at random from the set union of the ratings for both Pathfinder 1 and Pathfinder 2.

Insert Table 2 about here.

6 Phase 2: A Case-by-Case Analysis

The experiment described in the previous sections shows that there is a difference between Pathfinder 1 and Pathfinder 2, but it does not identify those aspects of the two knowledge bases that are responsible for these differences. To discern the causes for the observed differences, we examined each patient case where the difference between the expert ratings for Pathfinder 1 and 2 exceeded 1.5.

There were 12 cases in which the expert rating for Pathfinder 2 exceeded that for Pathfinder 1 by this threshold. In nine cases, a single factor was responsible for the increased performance; in two cases, two factors were responsible; and in one case, four factors were responsible. In contrast, there were only three cases in which the expert rating for

Pathfinder 1 exceeded that for Pathfinder 2 by 1.5. In all three cases, a single factor of the knowledge base was responsible for this decrease in diagnostic accuracy. Table 3 summarizes the factors of the Pathfinder 2 knowledge base that increased or decreased its diagnostic accuracy relative to Pathfinder 1. Many of these factors affected performance in more than one case. The tables show the number of times each attribute contributed to a difference in performance.

Insert Table 3 about here.

6.1 Causes of Increased Diagnostic Accuracy

In eight of the 12 cases where Pathfinder 2 outperformed Pathfinder 1, the representation of feature dependencies contributed to the superior performance of Pathfinder 2. In three of the cases, Pathfinder 2's increased accuracy was a direct consequence of the explicit encoding of dependencies. That is, in these three cases, the community pathologist observed features that were dependent. In the remaining five cases, however, the source of the improvement was indirect. In particular, by conditioning probability assessments for a feature on other features, the expert provided probabilities of higher quality. For example, let us consider the assessment of the probability distribution for LLC CYTOPLASM (color of large-lymphoid-cell cytoplasm), given DISEASE. In Pathfinder 1, the expert provided these assessments directly. In Pathfinder 2, however, the expert conditioned these assessments on LLC IDENTITY (identity of large lymphoid cells). That is, he assessed a probability distribution for LLC IDENTITY, given DISEASE, and probability distributions for LLC CYTOPLASM, given DISEASE and LLC IDENTITY. This technique for decomposing the assessment of a probability distribution is called *extending the conversation*. Using this technique, an expert can avoid having to average over a set of distributions in his head, and thereby can produce better assessments. For a detailed discussion of this technique and the conditions under which it is useful, see Tribus (1969, Chapter 3), de Finetti (1977), and Heckerman and Jimison (1987).

Another source of increased accuracy was that the second author's knowledge improved since the construction of Pathfinder 1. For example, in a previous evaluation of Pathfinder 1 (Heckerman, 1988), the system performed poorly in many cases because the probabilities

assessed for the feature EPI HIST CLUS (epithelioid clusters of histiocytes) were contradicted by data. That is, the second author said that these clusters were never seen in most diseases, yet, in the process of evaluating the system, he saw small numbers of these clusters in unexpected settings. During the year since that experiment, the second author paid close attention to these clusters in his daily diagnostic workups. Thus, the probability distributions for this feature that he provided later were significantly more informed than were those he provided for Pathfinder 1.

In one case, the diagnostic accuracy of Pathfinder 2 was superior to that of Pathfinder 1 because we introduced disease subtypes into the latter system. To see how the failure to include disease subtypes can decrease diagnostic accuracy, let us consider the disease necrotizing lymphadenitis, which has subtypes Kikuchi's and nonKikuchi's. In nonKikuchi's necrotizing lymphadenitis, we always see necrosis, and we sometimes see large numbers of plasma cells. On the other hand, in Kikuchi's necrotizing lymphadenitis, we may not see necrosis, and we never see large numbers of plasma cells. Furthermore, these two features are conditionally independent, given disease. Thus, if we fail to observe necrosis in a given lymph node, and if we see a large number of plasma cells in that same node, then both subtypes of necrotizing lymphadenitis should be ruled out. Suppose, however, that we construct an expert system that does not distinguish the two subtypes of disease, and retains the assertion conditional independence. In this case, if we observe no necrosis and abundant plasma cells in a given lymph node, the expert system incorrectly reports that necrotizing lymphadenitis is a possible contender for the diagnosis of that node. Here, when we combine the two subtypes of necrotizing lymphadenitis, necrosis and plasma cells become conditionally dependent, given disease. Consequently, the diagnostic accuracy of such a system is less than that of a system that includes the distinction.[2]

In four of the 12 cases, our use of the partition representation led to increased diagnostic accuracy. Heckerman (1991, Chapter 5) describes this representation and the reasons it led to increased diagnostic accuracy.

[2]In principle, we could avoid the introduction of disease subtypes by representing the feature dependencies that result from such a representation. Usually, however, the number of induced dependencies is large, and this approach is impractical.

Finally, in two of the 12 cases, we traced the improvements to differences between the systems in specific probability assessments. We could not, however, identify the underlying cause of the improvements.

6.2 Causes of Decreased Diagnostic Accuracy

In three cases, Pathfinder 1 outperformed Pathfinder 2. In one of these three cases, the decreased performance of Pathfinder 2 was caused by our failure to correctly maintain the consistency of the Pathfinder 2 knowledge base. To illustrate the problem, let us consider the feature NECROSIS, which conditions the feature KARYORRHEXIS. The probability that NECROSIS is ABSENT, given CAT SCRATCH DISEASE is 0. Thus, we do not need to assess a probability distribution for KARYORRHEXIS, given CAT SCRATCH DISEASE and NECROSIS is ABSENT.

In the current implementation of our knowledge-acquisition software for Pathfinder, called SimNet, the user must make sure that the assessments follow this constraint. Thus, there is the possibility of inconsistency. In particular, when assessing a feature that is dependent on other features, an expert might sometimes believe incorrectly that a conditioning instance was impossible, and fail to assess a needed distribution. Avoiding inconsistencies becomes especially difficult during the phase of knowledge-base development in which the system is tested and the probabilities are modified. When a probability of seeing a feature is changed from 0 to another value, an expert easily can forget to assess the distributions that are conditioned by that feature.

While assessing the probabilities for Pathfinder, we were careful to avoid these errors. Whenever there was a question as to the possibility of a conditioning instance, we checked the probability of that event, using the display facilities of SimNet. Nonetheless, despite our efforts, we made several errors when modifying probability distributions in the later stages of knowledge-map construction. One of these errors surfaced in this evaluation.

In the other two cases where Pathfinder 1 outperformed Pathfinder 2, we traced the decrement in accuracy to differences in specific probability assessments, but we could not determine the source of these differences.

7 Phase 3: A Decision-Theoretic Metric

The two approaches for evaluation that we have examined are easy to apply. Furthermore, they readily expose differences between the diagnostic accuracy of Pathfinder 1 and 2 and the causes of these differences. Unfortunately, it is difficult to infer the *importance* of differences based on these experiments. Specifically, in Section 2, we saw that the construction of Pathfinder 2 required approximately 40 more hours of effort than did the construction of Pathfinder 1. Neither the difference between the average expert ratings of approximately 1.0 on a scale from 0 to 10 nor the identification of system factors responsible for the improvement, however, can tell us whether this additional effort was worth the improvement in diagnostic accuracy.

In this section, we use an evaluation procedure, based on decision theory, that can address this tradeoff. In particular, we measure the value to a patient of each version of Pathfinder. To do this, we compute a quantity called *inferential loss* for both versions of the program and for each of the 53 test cases. The inferential loss associated with a version of Pathfinder and a given case is the difference between the expected utility of a patient who is given a gold-standard diagnosis, and the expected utility of the same patient who is given Pathfinder's diagnosis. This measure does not necessarily reflect the true value of each program to a patient. For example, in deriving this metric, we assume incorrectly that a pathologist who uses Pathfinder will never modify Pathfinder's diagnosis. Also, we assume that a patient always has access to a gold-standard diagnosis (see the following subsection). Nonetheless, the measure provides an approximation of the value of each system.

To compute inferential loss, we require (1) gold-standard probability distributions for each case, and (2) the utility of every possible correct and incorrect diagnosis, given every disease that a patient might have. In Sections 7.1 and 7.2, we examine these components of the computation; in Section 7.3, we discuss the computation in detail.

7.1 Gold-Standard Distributions

It is difficult to produce an adequate gold standard in the domain of pathology. One approach, illustrated in Figure 2(a), is simply to use the *true* disease to construct the gold-

standard distribution. That is, we assign a probability of 1 to the established diagnosis. In pathology, the disease that is manifested in a lymph node is determined (1) by an expert pathologist examining tissue sections under a microscope; (2) by expensive immunology, molecular biology, or cell-kinetics tests; (3) through observations of the time course of a patient's illness; or (4) by a combination of these approaches.

There are two problems with this gold standard. First, its use ignores the distinction between a good *decision* and a good *outcome*. For example, suppose the observations for a case suggest—say, through statistical data—that there is a 0.7 chance of Hodgkin's disease and a 0.3 chance of mononucleosis. Furthermore, suppose that mononucleosis is the true disease (not an unlikely event). In such a situation, an inference method that produces exactly this probability distribution for Hodgkin's disease and mononucleosis receives (unjustly) a lower rating than a distribution that produces a higher chance of mononucleosis. This problem with the approach, however, is not serious. We can attenuate differences between good decisions and good outcomes by considering a large number of cases.

A second, more serious, problem with this construction stems from details of how microscopic observations are made by experts and nonexperts. In our experience, when experts examine such tissue sections, they typically see many features at once and come to a diagnosis immediately. When asked to identify specific features that appear in the tissue, these pathologists report mostly features that confirm their diagnosis. Moreover, it is difficult to train these experts to do otherwise, and essentially impossible to determine whether or not such training is successful. Thus, when experts are used to identify features, both Pathfinder 1 and 2 tend to perform well, and, in practice, it becomes impossible to identify significant differences from an experimental comparison. On the other hand, pathologists who do not specialize in the lymph-node domain misrecognize or fail to recognize some of features associated with diagnosis. It is unreasonable to compare the distributions produced by Pathfinder 1 and 2, derived from one set of observations, with the true disease, derived from a different set of observations. In fact, in a separate study, we showed that errors in diagnosis resulting from the misrecognition and lack of recognition of features by a nonexpert were sufficient to obscure completely the differences between the two versions of Pathfinder, when the true diagnosis was used as the gold standard (Heckerman et al., 1990).

An alternative procedure for constructing a gold standard is shown in Figure 2(b). In this procedure, an expert looks at only a list of observations for a case produced by another pathologist (expert or nonexpert), and assesses directly a probability distribution over the diseases. An associated drawback is that this construction ignores the possibility that one or both versions of Pathfinder might outperform the initial impressions of the expert. That is, if, for each case, our expert were to undergo a detailed decision analysis, the probability distributions determined by these analyses may be closer to the distributions produced by Pathfinder 1 or 2 than to the expert's initial assessments of probability.

Insert Figure 2 about here.

Thus, in the pathology domains, there appears to be no ideal gold standard. For this experiment, however, the construction of the gold-standard distribution using the true disease is unworkable, given the difficulties of feature observation associated with experts and nonexperts. Consequently, we employed the procedure illustrated in Figure 2(b). As mentioned in Section 4, a nonexpert pathologist identified features.

7.2 A Utility Model for Diagnosis

Pathfinder 1 and 2 share the same utility model. For each disease pair (d_i, d_j), the model contains the utility u_{d_i,d_j}, which summarizes the preferences of the decision maker for the situation in which a patient has disease d_i, but is diagnosed as having disease d_j. Factors that influence such preferences include the length of the patient's expected life, the pain associated with treatment and with the disease itself, the psychological trauma to the patient and his family, and the monetary cost associated with treatment and with disability.

The most fundamental question in utility assessment is: Who is the decision maker? From our perspective, a pathologist is only a provider of information. Thus, the u_{d_i,d_j} in the utility model of a computer-based diagnostic system should reflect the patient's preferences. For example, consider the situation where a pathologist believes, after reviewing a case, that the probability of the benign infection mononucleosis is 0.9, and that the probability of Hodgkin's disease is 0.1. Should the patient be treated for Hodgkin's disease now, or should he wait for more definitive diagnostic signs to develop? Delaying treatment of Hodgkin's

disease decreases the chances of long-term survival if the patient has this condition. On the other hand, the treatment for Hodgkin's disease is highly invasive and thus carries significant risk. In addition to suffering the acute trauma of chemotherapy and radiotherapy, a patient is likely to become sterile and is put at increased risk for developing other cancers. The decision about therapy will depend on how the *patient* feels about the alternative outcomes. Different patients may have dramatically different preferences. Consequently, the inferential losses computed in this evaluation are meaningful to an individual only to the degree that the diagnostic utilities used in the evaluation match the diagnostic utilities of the that individual.

For this study, the second author provided the utilities, because he was familiar with many of the ramifications of correct and incorrect diagnosis. The second author, because he is an expert, had biases that made his initial preferences deviate from those of a typical patient. For example, many sets of diseases of the lymph node currently have identical treatments and prognoses. Nonetheless, experts like to distinguish diseases within each of these sets, because doing so allows research in new treatments to progress. That is, experts often consider the value of their efforts to future patients. In addition, experts generally suffer professional embarrassment when their diagnoses are incorrect. Also, experts are concerned about the legal liability associated with misdiagnosis. In an effort to remove these biases, the second author imagined that he himself had a particular disease, and assessed the diagnostic utilities accordingly.

Another important consideration in utility assessment is the wide range of severities associated 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 nonnegligible consequence is that the patient will take antibiotics unnecessarily for several weeks. As we discussed, however, if a patient has Hodgkin's disease and is incorrectly diagnosed as having mononucleosis, 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 that he would be willing to be killed for $10 million. This inference is absurd.

Howard (1980) has 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*. 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.[3]

Using this model, the second author provided his preferences for major and minor outcomes in a common unit: *the probability of immediate, painless death that he was willing to accept to avoid a given outcome and to be once again healthy.* First, we established the second author'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).[4] Next, we assessed his preferences for minor outcomes with willingness-to-pay

[3]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.

[4]Howard (1980) proposes a model by which this value can be computed from other assessments. A simple version of the model requires a decision maker to trade off the amount of resources he consumes during his lifetime and the length of his lifetime, to characterize his ability to turn present cash into future income (summarized, for example, by an interest rate), and to establish his attitude toward risk. Nonetheless, we

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 of death. For example, an answer of $100 translated to a utility of a 5 in 1 million probability of death. Finally, we assessed the second author's 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.

Note that, with this utility model, the inferential losses computed for Pathfinder 1 and 2 will have units "probability of death." In many cases, we shall see that the losses are small in these units (on the order of 0.0001). Consequently, it is useful to define a *micromort*, a one–in–1-million chance of death. In these units, for example, the second author, who has a small-risk value of life of $20 million, should be willing to buy and sell risks of death at the rate of $20 per micromort. This unit of measurement is also useful because it helps to emphasize that the linear relationship between risk of death and willingness to pay holds for only small probabilities of death. Howard (1989) discusses in detail the use of the micromort for medical decision making.

Finally, an important consideration is the complexity of the utility-assessment procedure. There are approximately 60 diseases represented in Pathfinder. The direct measurement of the u_{d_i,d_j} therefore requires about $60^2 = 3600$ assessments. Clearly, the measurement process would be tedious. Thus, several steps were taken to reduce the complexity of the task. For one, the second author established sets of diseases that have identical treatments and prognoses. He identified 36 such equivalence classes, reducing the number of direct utility assessments to $36^2 = 1296$. In addition, we decomposed many of the utilities into independent assessments such as the disutility of a disease when correctly treated, the disutility of delaying the appropriate treatment, and the disutility of the treatment in the absence of disease (e.g., the disutility of taking antibiotics or undergoing surgery). Through such

did not find it difficult to assess the small-risk value of life directly.

decomposition, more than 80 percent of the direct assessments were avoided. In total, the construction of the utility model took approximately 60 hours.

7.3 The Computation of Inferential Loss

The procedure for computing inferential loss is identical for both versions of Pathfinder and is illustrated in Figure 3. First, based on the features reported for a given case, Pathfinder produces a probability distribution over diseases. Next, based on these same features, we determine a gold-standard distribution as described in Section 7.1.

Insert Figure 3 about here.

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

$$dx_{\text{pf}} = \text{argmax}_{d_j} \left[\sum_{d_i} p_{\text{pf}}(d_i)\ u_{d_i,d_j} \right]$$

$$dx_{\text{gs}} = \text{argmax}_{d_j} \left[\sum_{d_i} p_{\text{gs}}(d_i)\ u_{d_i,d_j} \right]$$

where $p_{\text{pf}}(d_i)$ and $p_{\text{gs}}(d_i)$ represent the probability of the ith disease under the Pathfinder and gold-standard probability distributions, respectively.

Next, we compute the expected utility of dx_{pf} and dx_{gs}, denoted $eu(dx_{\text{pf}})$ and $eu(dx_{\text{gs}})$, respectively. When computing expected utility, we use the gold-standard distribution, which reflects the assumed best distribution. That is, we compute

$$eu(dx_{\text{pf}}) = \sum_{d_i} p_{\text{gs}}(d_i)\ u_{d_i,dx_{\text{pf}}}$$

$$eu(dx_{\text{gs}}) = \sum_{d_i} p_{\text{gs}}(d_i)\ u_{d_i,dx_{\text{gs}}}$$

Finally, we determine inferential loss, denoted IL, for the Pathfinder distribution, by subtracting the expected utility of the Pathfinder diagnosis from the expected utility of gold-standard diagnosis. That is,

$$\text{IL} = eu(dx_{\text{gs}}) - eu(dx_{\text{pf}})$$

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

7.4 Results

The mean and standard deviation of inferential loss for the two versions of Pathfinder are shown in Table 4. Case-by-case results are given in Heckerman (1991), Appendix D. Unlike the difference of 0.95 produced by the expert ratings, these results clearly reflect the increase in value of Pathfinder 2 as a result of this system's superior diagnostic accuracy. Specifically, assuming that a patient is willing to convert micromorts to dollars at a rate of \$20 per micromort,[5] as our expert was, the results in this metric show that it is worth approximately \$6000 *per case* to the patient to have the more sophisticated Pathfinder knowledge be used instead of the earlier knowledge base that assumed global independence among features. As was mentioned earlier, it took us approximately 40 hours longer to construct Pathfinder 2 than it did to construct Pathfinder 1. Thus, assuming we should be paid a combined hourly rate of \$400, the additional effort would more than pay for itself after only three cases had been run.

Insert Table 4 about here.

The standard deviations for inferential loss are quite large relative to the means. The reason for such large variances is easily appreciated. For many of the cases, the optimal diagnosis associated with the distributions produced by both versions of Pathfinder are identical to the optimal diagnosis associated with the gold standard. In particular, the optimal diagnoses for Pathfinder 1 agreed with the gold-standard diagnoses in 47 of the 53 cases; those for Pathfinder 2 agreed in 50 of the 53 cases. In these cases, inferential loss is zero. In the remaining cases, the approaches determine a diagnosis that differs from the gold standard. Most of these nonoptimal diagnoses are associated with expected utilities that are

[5]The value of \$20 per micromort applies to the expert when he is healthy. We use this value to approximate his small-risk value of life in situations where he is ill.

significantly lower than is the expected utility associated with the gold-standard diagnosis. Thus, inferential losses fluctuate from zero in most cases to large values in the remainder.

Despite the large standard deviations for inferential loss, a Bootstrap permutation test suggests that the results are not due only to chance. In particular, the test yields an ASL of 0.08. Again, this means that there is only an 8 percent chance that the difference in diagnostic accuracy would be more extreme than what the current results show, if inferential losses were drawn at random for the set union of inferential losses for Pathfinder 1 and 2.

Finally, the small inferential loss for Pathfinder 2—16 micromorts—suggests that this version of Pathfinder performs as well as the expert. A Bootstrap permutation test comparing the inferential losses of Pathfinder 2 with those of the expert (IL = 0 for every case) shows that there is a 12.5 percent chance that the difference in diagnostic accuracy would be no less than 16 micromorts, if inferential losses were drawn at random from the union of the two sets of inferential losses. Given that this study, by its design, cannot identify situations in which the expert system outperforms the expert, we have strong evidence that the diagnostic accuracy of Pathfinder 2 is at least as good as that of the Pathfinder expert.

8 Discussion

All phases of this experiment provided useful results. The expert-rating approach was easy to implement, and it showed that there were differences in diagnostic accuracy between the two versions of Pathfinder. Thereby, it suggested that pursuing each of the second two phases would be worthwhile. The second phase showed that several factors were responsible for the superior performance of Pathfinder 2, the most frequent cause being that improvements in probability assessments resulted from conditioning the assessments on other events. Finally, the third phase showed that the additional work required to construct Pathfinder 2 was well worth the effort.

The informal metric used in phase 1 was more sensitive to differences in the two systems than was the decision-theoretic metric used in phase 3. This observation is not surprising because experts, who have their integrity at stake, tend to be hypersensitive to errors in diagnosis, regardless of the degree to which such errors matter to the patient. Of course, the

decision-theoretic metric can be modified to be more sensitive. Considerations of integrity or liability, for example, can be incorporated into the diagnostic utilities.

This evaluation has concentrated on an analysis of differences in diagnostic accuracy between Pathfinder 1 and 2 that arise from differences in the probabilistic knowledge represented within these systems. This focus is important for evaluating the usefulness of belief networks, similarity networks, and partitions. Nonetheless, we can use the experimental design described in this article to investigate other facets of Pathfinder performance. For example, as we mentioned, we used this methodology to measure the decrease in diagnostic accuracy that arises from the misrecognition and lack of recognition of morphologic features by a nonexpert (Heckerman et al., 1990). Also, using this approach, we are comparing the diagnostic accuracy of community pathologists who have access to Pathfinder to that of pathologists who do not have such access. In addition, we can measure the sensitivity of diagnostic accuracy to changes in the probabilities associated with the Pathfinder belief network. In yet another study, we can quantify the differences in diagnoses rendered by expert pathologists. In general, we can use this approach to evaluate, in clear terms, a wide variety of issues related to the safety and efficacy of expert systems for medicine.

Acknowledgments

Doyen Nguyen served as the community pathologist in the formal evaluation of Pathfinder; Larry Diamond also assisted with the evaluation; and Majorie Bernstein-Singer helped us to build the Pathfinder utility model. Henri Suermondt, Mark Fischinger, Marty Chavez, and especially Keung-Chi Ng assisted with programming and data management. Eric Horvitz, Lyn Dupre, Greg Cooper, and Keung-Chi Ng provided useful comments on this manuscript.

This work was supported by the National Cancer Institute under Grant RO1CA51729-01A, and by the National Library of Medicine under Grant RO1LM04529.

References

Asselain, B., Derouesne, C., Salamon, R., Bernadet, M., and Gremy, F. (1977). The concept

of utility in a medical decision aid: Example of an application. In *Proceedings of Medinfo,* Halifax, Nova Scotia, pages 123–125. North-Holland, New York.

Bartlett, F. (1958). *Thinking.* Basic Books, New York.

Buchanan, B. and Shortliffe, E. (1984). Uncertainty and evidential support. In Buchanan, B. and Shortliffe, E., editors, *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, pages 209–232. Addison–Wesley, Reading, MA.

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

de Dombal, F., Leaper, D., Staniland, J., McCann, A., and Horrocks, J. (1972). Computer-aided diagnosis of acute abdominal pain. *British Medical Journal*, 2:9–13.

de Finetti, B. (1977). Probabilities of probabilities: A real problem or a misunderstanding. In Aykac, A. and Brumet, C., editors, *New Developments in the Applications of Bayesian Methods*, pages 1–10. North-Holland, New York.

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

Elstein, A. (1976). Clinical judgment: Psychological research and medical practice. *Science*, 194:696–700.

Elstein, A., Loupe, M., and Erdman, J. (1971). An experimental study of medical diagnostic thinking. *Journal of Structural Learning*, 2:45–53.

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

Gorry, G. (1973). Computer-assisted clinical decision making. *Methods of Information in Medicine*, 12:45–51.

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

Heckerman, D. (1988). 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. 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.

Heckerman, D. (1990). Probabilistic similarity networks. *Networks*, 20:607–636.

Heckerman, D. (1991). *Probabilistic Similarity Networks*. MIT Press, Cambridge, MA.

Heckerman, D., Horvitz, E., and Nathwani, B. (1985). Pathfinder research directions. Technical Report KSL-89-64, Medical Computer Science Group, Section on Medical Informatics, Stanford University, Stanford, CA.

Heckerman, D., Horvitz, E., and Nathwani, B. (1989). 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.

Heckerman, D., Horvitz, E., and Nathwani, B. (1990). Toward normative expert systems: The Pathfinder project. Technical Report KSL-90-08, Medical Computer Science Group, Section on Medical Informatics, Stanford University, Stanford, CA.

Heckerman, D. and Jimison, H. (1987). A perspective on confidence and its use in focusing attention during knowledge acquisition. In *Proceedings of the Third Workshop on Uncertainty in Artificial Intelligence,* Seattle, WA, pages 123–131. Association for Uncertainty in Artificial Intelligence, Mountain View, CA. Also in Kanal, L., Levitt, T., and Lemmer, J., editors, *Uncertainty in Artificial Intelligence 3,* pages 123–131. North-Holland, New York, 1989.

Henrion, M. and Cooley, D. (1987). An experimental comparison of knowledge engineering for expert systems and for decision analysis. In *Proceedings AAAI-87 Sixth National*

Conference on Artificial Intelligence, Seattle, WA, pages 471–476. AAAI Press, Menlo Park, CA.

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

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

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

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

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

Nathwani, B., Heckerman, D., Horvitz, E., and Lincoln, T. (1990). Integrated expert systems and videodisc in surgical pathology: An overview. *Human Pathology*, 21:11–27.

Patil, R. (1981). *Causal representation of patient illness for electrolyte and acid-base diagnosis.* PhD thesis, Computer Science Department, Massachusetts Institute of Technology.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, CA.

Shortliffe, E. and Buchanan, B. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23:351–379.

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

Tribus, M. (1969). *Rational Descriptions, Decisions, and Designs.* Pergamon Press, New York.

Figures

Figure 1: The Pathfinder belief network. The belief network represents over 100 features that are relevant to diagnosis. The node DISEASE contains over 60 lymph-node diseases. (Taken from Heckerman, 1990.)

Figure 2: Alternative gold standards for a given case. (a) From the morphologic observations of the expert, expensive immunology, molecular biology and cell-kinetics tests, and information about the time course of the patient's illness, we determine the true disease. (b) Given only the list of observations reported by another pathologist, the expert assesses a gold-standard probability distribution, represented by the shaded probability wheel. (Taken from Heckerman, 1991.)

Figure 3: The computation of inferential loss. Based on the features reported, Pathfinder (1 or 2) produces a probability distribution over diseases (represented by the shaded probability wheel). Based on these same features, we also determine a gold-standard distribution. Given these distributions, we identify the optimal diagnoses associated with them, denoted dx_{pf} and dx_{gs}, using the principle of maximum expected utility. We compute the expected utility of the two diagnoses with respect to the gold-standard distribution. We then calculate the inferential loss associated with the Pathfinder distribution by subtracting the expected utility of the Pathfinder distribution from the expected utility of the gold-standard distribution. (Taken from Heckerman, 1991.)

Tables

Table 1: Statistics for the construction of Pathfinder 1 and 2. (Adapted from Heckerman, 1991.)

Statistic	Knowledge Base	
	Pathfinder 1	Pathfinder 2
Time to define features and diseases	8 hours	8 hours
Time to build similarity network	-	35 hours
Probabilities required by system	16,620	74,854
Probabilities derived from partitions	-	61,118
Probabilities provided by expert	16,620	13,736
Time to assess probabilities	40 hours	39 hours

Table 2: Expert ratings for Pathfinder 1 and 2. (Taken from Heckerman, 1991.)

Knowledge Base	Expert Ratings (0-10 scale) mean	sd
Pathfinder 1	7.99	2.32
Pathfinder 2	8.94	1.51

Table 3: Factors of Pathfinder 2 that increased and decreased its diagnostic accuracy relative to Pathfinder 1. (Adapted from Heckerman, 1991.)

Factors that Increased Diagnostic Accuracy

Number of Cases Affected	Factor
5	Conditioning produced better assessments
3	Dependencies existed among observed features
2	Expert's knowledge improved since construction of Pathfinder 1
1	Disease subdistinction introduced
4	Probability assessments improved by partitions
2	Probability assessments improved for unknown reasons

Factors that Decreased Diagnostic Accuracy

Number of Cases Affected	Factor
1	Failure to maintain consistency of knowledge base
2	Probability assessment worsened for unknown reasons

Table 4: Inferential losses for Pathfinder 1 and 2. (Taken from Heckerman, 1991.)

Knowledge Base	Inferential Loss (micromorts) mean	sd
Pathfinder 1	340	1684
Pathfinder 2	16	104