ELSEVIER

Artificial Intelligence 85 (1996) 363–397

Artificial Intelligence

The sensitivity of belief networks to imprecise probabilities: an experimental investigation

Malcolm Pradhan [a,b,*], Max Henrion [a], Gregory Provan [a], Brendan Del Favero [a,c], Kurt Huang [a,b]

[a] *Institute for Decision Systems Research, 4984 El Camino Real, Suite 110, Los Altos, CA 94022, USA*

[b] *Section on Medical Informatics, Stanford University, Stanford, CA 94305, USA*

[c] *Engineering-Economic Systems, Stanford University, Stanford, CA 94305, USA*

Received January 1995; revised October 1995

Abstract

Bayesian belief networks are being increasingly used as a knowledge representation for reasoning under uncertainty. Some researchers have questioned the practicality of obtaining the numerical probabilities with sufficient precision to create belief networks for large-scale applications. In this work, we investigate how precise the probabilities need to be by measuring how imprecision in the probabilities affects diagnostic performance. We conducted a series of experiments on a set of real-world belief networks for medical diagnosis in liver and bile disease. We examined the effects on diagnostic performance of (1) varying the mappings from qualitative frequency weights into numerical probabilities, (2) adding random noise to the numerical probabilities, (3) simplifying from quaternary domains for diseases and findings—absent, mild, moderate, and severe—to binary domains—absent and present, and (4) using test cases that contain diseases outside the network. We found that even extreme differences in the probability mappings and large amounts of noise lead to only modest reductions in diagnostic performance. We found no significant effect of the simplification from quaternary to binary representation. We also found that outside diseases degraded performance modestly. Overall, these findings indicate that even highly imprecise input probabilities may not impair diagnostic performance significantly, and that simple binary representations may often be adequate. These findings of robustness suggest that belief networks are a practical representation without requiring undue precision.

Keywords: Probabilistic reasoning; Bayesian networks

* Corresponding author. E-mail: pradhan@camis.stanford.edu.

PII S0004-3702(96)00002-1

1. The tradeoff between accuracy and cost

Each knowledge representation or model is, by definition, a simplification of reality. When the representation is derived from a human expert, it is a simplification even of the expert's perception of reality. The question in choosing a representation is not whether the representation is completely accurate—it cannot be—but whether the model is *sufficiently accurate* for the purposes for which it is designed. This question is the one that drives our research.

The choice of a representation is a balancing act. On the one hand, a richer representation should improve the accuracy with which the model represents the real-world system. Greater accuracy in representation should lead to improved accuracy in inferences—for example, in a medical application, diagnoses that would be more likely to be correct, and treatment recommendations that would be most effective. On the other hand, a richer representation will require more computational resources for inference and for storage, and will require more effort to construct, verify, and maintain. The success of knowledge-based systems depends critically on the knowledge engineer's ability to find an effective tradeoff between accuracy and cost.

Experienced knowledge engineers generally develop useful intuitions about how to make such choices; however, there is little theoretical or experimental research currently available to guide them. Of course, the exploration of the generality, limitations, and computational complexity of alternative knowledge representations has been a major topic of AI research. However, once we have chosen a particular type of representation—such as rules, a nonmonotonic logic scheme, or Bayesian belief networks—there is little research available to guide the knowledge engineer in deciding how the complexity or richness of the model is likely to affect its performance for a given application. Theoretical analysis can be valuable here. But, due to the analytic complexity of the relationships between representation and performance, experimental work must play an important role.

1.1. Experiments on belief networks

In recent years, there has been substantial growth in interest in Bayesian belief networks (BNs) as a knowledge representation [32]. There has been work on the development of effective knowledge engineering techniques, efficient inference algorithms, and increasing numbers of real-world applications of BNs [14, 17]. The primary goal of the work described here is to investigate how the precision of representation of BNs affects the quality of diagnosis based on the network. We view this research as a contribution towards the eventual goal of developing an empirical and theoretical basis for guidelines for knowledge engineers to help them choose the level and complexity of representation that provides the most appropriate tradeoff between accuracy and cost.

Our investigation is based on a series of real-world BNs, rather than on the randomly generated, abstract knowledge bases (KBs) used in much of the experimental research to compare knowledge representations. Although it is easy to generate BNs with a wide range of different characteristics—such as ratio of arcs to nodes, ratio of source nodes to internal nodes, or frequency of undirected cycles—we wanted to focus on BNs that have

the characteristics of real application domains. We believe such BNs are more likely to be relevant to other real application domains than artificially generated networks. The problem domain that we use in this study is medical diagnosis for hepatobiliary disorders (liver and bile diseases).

We derived the experimental BNs from an early quasi-probabilistic KB, named computer-based patient case simulation (CPCS) [30] that uses a representation derived the Internist-1 [25] and quick medical reference (QMR) [24] expert systems. In these knowledge bases, causal links, such as the relationship between disease and finding are quantified as *frequency weights*, specifying the chance that one disease will give rise to a finding or other variable, on a five-point qualitative scale. In previous work, our group developed a method to convert from the Internist-1/QMR representation to a belief network representation, with specific independence assumptions—conditional independence of findings given diseases, noisy OR influences of diseases on findings, and marginal independence of diseases [40]. Empirical comparison of QMR with the probabilistic reformulation, QMR-BN, demonstrated comparable diagnostic performance [23], even though some information (e.g. linkages between diseases) was not employed in QMR-BN.

Our first task in the current work was to convert the CPCS knowledge base into a coherent BN, mapping frequency weights into *link probabilities*, which are the conditional probabilities of each finding given each disease. We also had to assess additional *leak probabilities*, to quantify the chance that each finding, or other variable, will be present but not caused by one of the diseases or other variable in the knowledge base, and *prior probabilities* to quantify the prevalence rate of each disease or predisposing factor.

Bayesian representations in general, and BNs in particular, have been criticized by certain AI researchers because they require large numbers of numerical probabilities to quantify uncertain relationships. Whether these probabilities are estimated directly from data, or assessed as subjective probabilities by domain experts, or some combination of the two, there is no denying the fact that a conventional BN representation has a voracious appetite for such numbers.

The first question we examined is how precise such numbers need to be. The literature on the expert assessment of subjective probabilities makes clear that subjective probabilities are liable to consistent biases and imprecision. If it turns out that BNs, to achieve adequate diagnostic performance, require numerical probabilities with greater precision than experts can provide, BNs will be of little practical value. But, if a BN's performance turns out to be insensitive to probable errors, we can allay concerns about the reliability of subjective probability assessments.

We performed two experiments to examine the sensitivity of BNs to the expert probabilities. In each experiment, we assessed the effect of the manipulations in terms of their effect on diagnostic performance, measured as the probability assigned to the correct diagnosis averaged over a large number of diagnostic test cases, for three different BNs.

First, we compared the standard, empirically derived mapping [15] from frequency weights into probabilities to two alternative mappings, the *curvilinear mapping* that treats frequency weights as order-of-magnitude probabilities, and the *uniform mapping*, that ignores differences between the numbers by treating all links as having equal strength.

Second, we added random noise to the probabilities derived from the standard mapping. In this case, we added noise separately to the *link probabilities*, *leak probabilities*, and the *prior probabilities*. By examining the effect of noise separately on each of these three types of probability, we were able to differentiate among them in terms of their effect on diagnostic performance.

In our third experiment, we examined the effect of the *domain size* of variables, such as diseases and findings—that is, the number of values each variable can take. We compared the performance of networks containing quaternary domains {*absent*, *mild*, *moderate*, *severe*} with simplified networks containing binary domains {*absent*, *present*}. Enriching the representation from binary to quaternary domains entails much extra effort because more probabilities must be quantified. It also substantially increases the computational effort required for diagnosis. We examined the change in diagnostic performance to discover whether the additional work is likely to be worthwhile.

In our fourth experiment, we examined the effect of including *outside diseases* in the test cases, that is diseases that are not explicit in the network being tested. A major benefit of BNs is that they represent uncertainty explicitly, including uncertainty due to incompleteness of a model. The leak probability for a finding (or other variable) represents the probability that the finding will be present for a reason that is not modeled explicitly in the network—perhaps a false positive or a disease or fault not modeled. Because the BN represents leak events explicitly, we can infer the probability that the true explanation of a finding is a cause not represented explicitly in the model. In this way, the BN supports reasoning about scope and limitations of the representation. For our experiments, we extracted three smaller subnetworks from a large BN. We used some test cases that included diseases in the large network, but not always in the subnetworks. We were able to test performance when there are diseases present that are not in the scope of the diagnostic network.

1.2. Overview

The paper is organized as follows. In Section 2, we review previous work on the sensitivity of probabilistic knowledge representations to variations in the numbers and representation. In Section 3, we describe how we converted the qualitative CPCS knowledge base into a quantitative BN, and how we generalized the noisy OR into the noisy MAX relationship which we used to model the influence of multiple independent cause variables on each effect variable using quaternary domains. In Section 4, we present our experimental approach, including the selection of networks, generation of test cases, and the measures of diagnostic performance. In the following three sections, 5, 6, and 7, we present the experimental designs, results, and discussions for each of the three experiments, changing the probability mappings, random noise in the probabilities, and the domain size, respectively. We summarize our conclusions in Section 9.

2. Previous research on belief network sensitivity

Considering the degree of controversy about the relative merits of schemes for reasoning under uncertainty, there have been relatively few previous studies comparing

performance of alternative schemes. We will group these studies into comparisons of BNs with rule-based schemes, analysis of the sensitivity of BNs to numerical probabilities, effects on BNs of structural independence assumptions, and effects of other simplifications. We will conclude this section by a discussion of the reasons for the apparent variety of findings.

2.1. Comparison of probabilistic to rule-based and symbolic representations

Many comparisons of rule-based or symbolic knowledge representations with probabilistic BNs have found little or no significant difference in performance. For example, studies comparing the diagnostic accuracy of rule-based schemes with independent Bayes found no statistical difference in the diagnosis of gastrointestinal disease [11], acute abdominal pain [4], and acute abdominal pain due to gynaecological origin [42,43]. Chard et al. showed that an ad hoc qualitative scheme can perform as well as a Bayesian model in the diagnosis of gynaecological disorders [3]. O'Neil and Glowinski [29], in the diagnosis of chest pain, found that a Bayesian approach and a linear decision rule with uniform weights produced indistinguishable ROC curves. As noted earlier, a reformulation of a large heuristic knowledge base, Internist-1/QMR into a belief network version called QMR-BN [40], demonstrated comparable diagnostic performance to the original [23].

Heckerman and colleagues, with the Pathfinder project [14], found that a Bayesian belief network model and rule-based system both had comparable diagnostic ability to a human expert physician for lymph node pathology. However, they found that the belief network performed better overall, according to the human experts, perhaps because it had more parameters and was better tuned to the domain expert's subjective assessments.

Wise and Henrion [47] conducted an experimental comparison of the performance of six different schemes for representing uncertainty, including certainty factors, possibilistic logic, and BN schemes. They found that, in some cases, the differences among schemes were insignificant, but that in other cases the differences were substantial, particularly with weak or conflicting evidence. Indeed, some schemes could produce results that were qualitatively incorrect under these circumstances.

2.2. Sensitivity of uncertain reasoning to numerical inputs

Ng and Abramson [27] showed substantial robustness of diagnostic accuracy to noise added to the prior and conditional probabilities for a BN model in the domain of medical pathology. A recent study [18] found that diagnostic performance in a simple belief network for troubleshooting an automobile was barely affected by substituting order-of-magnitude probabilities, based on the kappa calculus [13].

2.3. Sensitivity of probabilistic reasoning to independence assumptions

There have been several empirical studies of the effect of assumptions about the conditional independence of findings in the independence Bayes model introduced by de Dombal [7]. One might expect BN models to provide more accurate diagnoses

than independence Bayes models, assuming all findings are conditionally independent on the disease state, because the BN models can capture conditional dependencies [12, 28]. The experimental evidence here is mixed. Seroussi [39], in the domain of acute abdominal pain, found a 4% increase in diagnostic accuracy (from 63.7% to 67.7%) by accounting for pairwise interactions using a Lancaster model. Todd and Stamper found no statistically significant difference between the two approaches [42, 43], and some have even found independence Bayes to be better [6, 9].

Fryback [12] showed empirically that a large model with many inappropriate independence assumptions tends to overweight positive evidence due to ignoring the dependencies among findings. He found that smaller models with appropriate independence assumptions can outperform larger models with inappropriate assumptions. In our analysis of QMR-BN [23], we also found unrealistically large posterior probabilities due to inappropriate assumption of conditional independence of findings, in examples with many findings (typically 20 to 50 findings per case). These results suggest that appropriate modeling of dependence can significantly affect in large networks.

2.4. Simplification of belief networks

Several researchers have explored schemes to simplify BNs, and examined their effects on reasoning performance. Jensen and Andersen [20] convert BNs to their equivalent clique trees [1] and then simplify the network by setting to 0 the k smallest probabilities in each clique, thereby taking advantage of the smaller probability tables in computing marginal probabilities. Kjaerulff [22] explored a complementary technique that deletes the least important edges from the clique tree, as measured by the Kullback–Leibler metric. Both studies found that useful simplifications could be obtained with little additional error. Sarkar [38] proposed methods for optimal approximation of a general BN with tree structure, in order to reduce computational complexity.

Other researchers have used domain-dependent simplification methods to study tradeoffs between diagnostic accuracy against the richness and size of BN models. Provan and colleagues [36, 37] developed methods to simplify temporal BN models for the medical management of acute abdominal pain. They found that diagnostic accuracy improved as a function of network complexity, but that, with an appropriate penalty for computational effort, a simplified representation could be optimal. Breese and Horvitz [2] and Breese [1] describe approaches to construct belief networks and influence diagrams dynamically from a database in response to the specifics of a problem, also supporting tradeoffs between complexity and accuracy.

2.5. Understanding results on sensitivities

The findings of low sensitivities of diagnostic performance to errors in numerical inputs are consistent with the widely observed robustness of simple linear models for classification under uncertainty. Experimental psychologists, in extensive studies of com-

[1] A clique is a fully connected graph, containing a set of directly dependent nodes. Any BN can be converted into a tree of cliques.

plex, configural judgments by human experts, have found that simple linear models with approximate and even uniform weights often do as well, and sometimes even better than human experts [5]. These results apply in domains where there are several noisy cues or features, so that even optimal performance is limited. The underlying explanation is based on the inherent robustness of linear models for a wide range of classification tasks [45]. Note that diagnostic Bayesian reasoning with conditional independence can be formulated as a weighted linear sum of evidence weights (log-likelihoods plus log-odds prior).

Von Winterfeld and Edwards [44] have shown that the optimal decisions and hence expected utility in a decision analysis are still less sensitive to errors in input probabilities than are the posterior probabilities. They show that this robustness is due to the necessary concavity of expected loss as a function of probability in the region around an optimal decision. Pierce [34] and Fishburn [10] have shown related results.

On the other hand, von Winterfeld and Edwards [44] also showed that errors in model structure can create arbitrarily large losses in utility. We should therefore be concerned about missing findings, missing diseases, or missing relationships, which would change the qualitative structure of a model and so could substantially affect results.

The findings of Wise and Henrion [47] suggest that if there is little or no evidence, the quality of the representation and inference engine makes little difference, because no scheme can compute an accurate diagnosis. On the other hand, if there is strong, consistent evidence, any reasonable scheme will perform well. In either case, there will be little sensitivity to representation or small numerical errors. The largest differences between schemes, and largest sensitivities to errors in inputs, occur when there is moderate or conflicting evidence. Accordingly, in the experiments described below, we vary the quantity of evidence systematically to ensure coverage of the intermediate situation.

3. Knowledge base conversion to belief network

The BN that we used for our experimental analysis supports medical diagnosis for liver and bile (hepatobiliary) diseases. We derived the network from a rich knowledge base, the CPCS system, developed by Parker and Miller [30] in the mid-1980s as an experimental extension of the Internist-1 knowledge base [25].

In this section, we describe the CPCS knowledge base, its relationship to Internist-1 and QMR, and how we mapped it into a BN representation, CPCS-BN. We describe the qualitative methods for identifying variables, their domains, and influences. In Section 3.2 we describe how we mapped from the frequency weights, which express the strength of relationships in Internist-1 and CPCS, to conditional probabilities. We also assessed prior probabilities and leak probabilities, which were additional quantities not derivable from CPCS in its original form. In Section 3.3 we define and explain the generalized noisy OR, or noisy MAX, which is the prototypical influence that we use to represent the probabilistic effects of multiple predecessor or causal variables on each effect variable.

3.1. Internist-1, QMR, and CPCS

The CPCS KB was developed as an experimental extension of the Internist-1 KB to support patient simulation and computer aided instruction. The developers felt that these tasks required a KB with a much richer representation than that of Internist-1. CPCS is restricted to the hepatobiliary medical domain because the developers regarded the knowledge engineering task too great to convert all of Internist-1 to a richer representation based on their experience with the CPCS system.

Internist-1, and more recently QMR, contain only diseases and findings—a two-level representation. The CPCS KB has a multilevel representation that includes diseases and findings as well as predisposing factors to diseases (PFDs) that influence disease prevalence rates, and intermediate pathophysiological states (IPSs) that mediate between diseases and findings. For example, consider a disease *acute gastritis* which involves blood loss and two findings suggestive of blood loss: *pale skin* and *low red blood cell count*. Internist-1 has direct links between this disease and the two findings. In contrast, the CPCS KB includes *anemia* (an IPS) between the disease and the two findings.

Whereas Internist-1 and QMR use binary domains {*absent*, *present*} for diseases and findings, CPCS uses quaternary domains {*absent*, *mild*, *moderate*, *severe*} for certain variables, such as diseases and IPSs. Some findings are also represented with multiple states. CPCS contains directed links between the variables of types predisposing to disease, disease to IPS and findings, and IPS to findings. CPCS, like its predecessors, represents the strength of the relationship between cause and effect variables by a *frequency* (an integer between 0 and 5) that represents a graded degree of likelihood.

3.2. Translation of the qualitative CPCS into a belief network

The original CPCS system was developed in FranzLisp. Diseases and IPSs were represented as Lisp frames. To construct the BN we converted the original CPCS KB to Common Lisp, and then parsed the frames to create nodes. We represented diseases and IPSs as four levels of severity in the CPCS-BN. Predisposing factors of a disease or IPS node were represented as that node's predecessors, and findings and symptoms of a disease or IPS node were represented as the successors for that node. In addition to the findings, CPCS contained causal links between disease and IPS frames; we converted these links into arcs in the BN.

In the conversion of CPCS to the BN representation, we checked for consistency using the domain knowledge of medical doctors associated with this project. Because the original CPCS knowledge base was not designed with probabilistic interpretations in mind, we had to make numerous minor corrections to remove artifactual nodes, to make node values consistent, and to confirm that only mutually exclusive values were contained within a node. For example the node *edema* (swelling due to fluid accumulation) was automatically created as one node containing the states {*none*, *legs*, *scrotum*}. Since edema may occur at more than one site simultaneously the node was broken into two nodes: *edema-legs* and *edema-scrotum*.

The resultant network has 448 nodes and over 900 arcs. Fig. 1 is a snapshot of part of the network that demonstrates the complexity of the CPCS-BN. Because inference in the

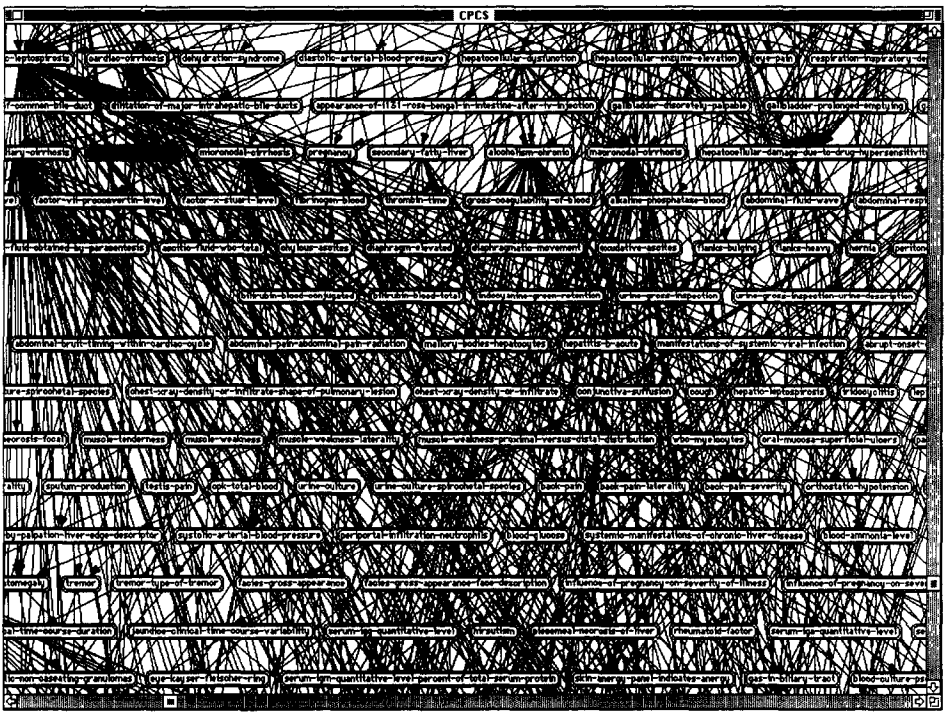

Fig. 1. Approximately one quarter of the full 450-node CPCS-BN.

complete CPCS-BN is extremely time consuming, and we ran approximately 300,000 experiments, we used subsets of the full network comprising 42 nodes (2 diseases), 146 nodes (3 diseases) and 245 nodes (4 diseases). These subsets are described in more detail in Section 4.

To complete the conversion to CPCS-BN, we assessed three sets of probabilities: prior probabilities, leak probabilities, and link probabilities. We assessed over 560 probabilities to specify the network fully, as described in the following subsections.

3.2.1. Assessment of the prior probabilities

The prior probabilities of the predisposing factors in CPCS-BN were assessed by physicians. However, because we are using the posterior probabilities of *disease* nodes as a measure of diagnostic accuracy in these experiments, we removed the predisposing factors and assessed the prior probabilities of the diseases. Predisposing factors often play an important role in medical diagnosis by effectively defining subpopulations with different disease rates (prior probabilities). For example, a population who has high intravenous drug use will have a much higher rate of viral hepatitis compared to the general population. The removal of predisposing factors from the experimental version of the CPCS network reduces it's diagnostic power and medical realism, however this ensures that the prior probabilities of the diseases would be consistent for each network, and not subject to uncontrolled variations due to noise or frequency to probability mapping.

We derived disease prior probabilities from the National Center for Health Statistics data, as had been done for the original QMR-BN.

3.2.2. Assessment of the leak probabilities

Experts assessed the leak probabilities by observing the predecessors of each node that required a leak, and deriving a probability that the node could be true given that each of the predecessors represented in the network was absent (leak probabilities are introduced formally in Section 3.3.2). For example, if a network includes the node *anemia* with predecessors *acute gastritis* and *pregnancy*, then the leak for *anemia* is the probability that a person could present with anemia *not* caused by *acute gastritis* or *pregnancy*.

The assessed leak probability of a node is specific to the particular set of causal factors (parents) of that node; if a parent is added or removed, then the leak probability must be modified. For example, if we include the condition *peptic ulcer* in our model as a new parent of the node *anemia*, the leak probability will decrease because the new condition is a relatively common cause of anemia and it is now explicitly modeled in the network.

3.2.3. Assessment of the link probabilities

The original CPCS system contained causal links from diseases to IPSs, and from diseases and IPSs to findings. The strengths of these links were indicated by an integer from 0 to 5 called a *frequency weight*. The frequency weight of a link was very roughly equivalent to the conditional probability that the successor node would be present given the predecessor node is present.

In constructing the CPCS-BN, we converted these frequency weights to probabilities by mapping the integers into the real interval $[0, 1]$, using the same mapping as was used in [40]. We also tested the sensitivity of the performance of the CPCS-BN to the particular probability mapping used, as described in Section 5.

The link probabilities represent only one-to-one relationships between the nodes, e.g. between a disease and a finding. To combine the effect of more than one disease on a finding, we use the devices of the noisy OR and the noisy MAX, which are described in the next section.

3.3. Modeling causal influences

The link probabilities described in the previous section model one-to-one relationships between diseases and findings. To combine the effect of multiple diseases on a single finding, we use the leaky, noisy OR, and the noisy MAX. These are simplified representations for probabilistic influence that require far fewer parameters than the full conditional probability matrix. For binary variables, the leak noisy OR requires a single parameter, a *link probability* to represent the strength of each link from one variable to another, e.g. from disease to finding. Researchers and practitioners using belief networks have found that noisy OR relationships are sufficient to represent a large majority of actual relationships between binary variables as judged by experts for diagnostic applications in many domains. They represent the situation

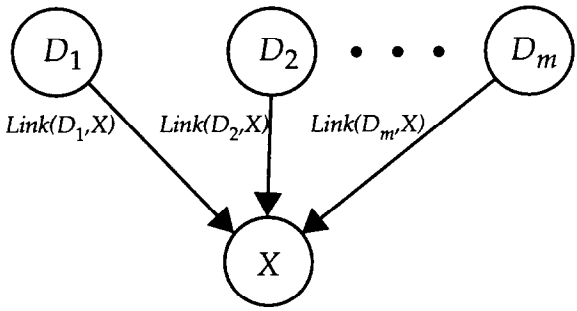

Fig. 2. A noisy OR network.

in which there are multiple diseases or faults that can cause a given finding or observable test-outcome, and where they are *causally independent*; the presence of one disease does not affect the tendency of each other disease to produce their common finding. There are some situations in which synergies and gating can occur among causes, which can be represented by other probabilistic relations, but these are a small minority. In the CPCS KB the presence of IPS nodes between diseases and findings make that the noisy OR and noisy MAX seem the most appropriate probabilistic interpretation.

In the following discussion, we denote variables using upper-case letters (e.g., X) and instantiations of variables using lower-case letters (e.g., x). Let the values (states) that a variable can take be $x[0], x[1], \ldots, x[n-1]$. The value $x[0]$ denotes the the absent state. Any state, $x[j]$, where $j > 0$ means X is present.

3.3.1. Noisy OR

The noisy OR is a model of probabilistic causal influence between a binary effect variable and a set of binary variables that represent its causes. This representation was originally proposed by Pearl [31] and independently by Peng and Reggia [33].

Consider a variable X that has m predecessors $D_1, \ldots, D_m$. The noisy OR can be used when (1) each D_i has a probability $Link(D_i, X)$ of being sufficient to produce the effect in the absence of all other causes, and (2) the probability of each cause D_i being sufficient is independent of the presence of other causes [16]. If these conditions hold, we can model the noisy OR relationship as a belief network, as in Fig. 2.

For the noisy OR network in Fig. 2, let $\Pi(X)$ be the set of explicitly modeled predecessor variables $\{D_1, \ldots, D_m\}$. Let $V_X \subseteq \Pi(X)$ be the subset of predecessors of X that are present and $\overline{V_X} \subseteq \Pi(X)$ be the subset of predecessors of X that are absent. We assume that all predecessors are instantiated (thus, $V_X \cup \overline{V_X} = \Pi(X)$). An instantiation of $\Pi(X)$ is denoted $\pi(X)$, and we define $\pi(X) = v_i$ as a specific instantiation of $\Pi(X)$ in which D_i is present and all other D_j $(j \neq i)$ are absent, or:

$$\pi(X) = v_i \equiv \{D_i = d_i[1] \text{ and } D_j = d_j[0] \text{ for all } j \neq i\}.$$

Let $Link(D_i, X)$, on the arc from D_i to X, represent the *link probability*, the probability that X is present given that D_i is present and all other predecessors are absent:

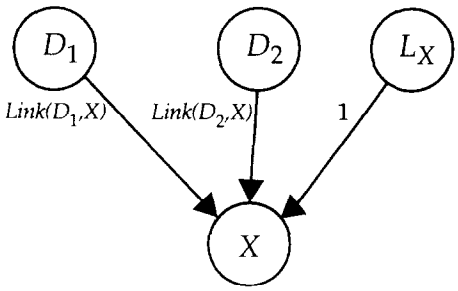

Fig. 3. Explicit representation of the leak probability as a cause of X.

$$Link(D_i, X) = P(X = x[1] \mid v_i). \tag{1}$$

Since the $D_i \in \Pi(X)$ are assumed to be causally independent, X is absent only when all D_i fail to cause X to be present:

$$\begin{aligned} P(X = x[0] \mid \Pi(X)) &= \prod_{i:D_i \in V_X} P(X = x[0] \mid v_i) \\ &= \prod_{i:D_i \subset V_X} (1 - Link(D_i, X)), \end{aligned} \tag{2}$$

from which it follows that the complement is given by:

$$P(X = x[1] \mid \Pi(X)) = 1 - \prod_{i:D_i \in V_X} (1 - Link(D_i, X)). \tag{3}$$

3.3.2. Leaky-noisy OR

Like any model, a BN is an incomplete representation of reality. We can use *leak events* to represent the missing variables that influence a finding. Each variable with predecessors has a corresponding leak event that represents all the possible events that could cause that finding to be present, other than those predecessor variables that are represented *explicitly* in the model. Fig. 3 shows a finding X, with two *explicit* predecessor variables, D_1 and D_2, and a leak event, L_X. Recall the set of explicitly modeled (non-leak) predecessors of X is $\Pi(X) = \{D_1, D_2, \ldots, D_m\}$.

The probability that the leak event is present is the *leak probability*. The leak probability for X, $Leak(X)$, is equal to the probability that X is present when all its m explicitly modeled predecessors $\Pi(X)$ are absent:

$$Leak(X) = P(L_X) = P(X = x[1] \mid D_i = d_i[0],\ i = 1, 2, \ldots, m). \tag{4}$$

Thus, we can model the leak event like any other explicit cause D_i of X. The only difference is that the link probability from L_X to X is exactly 1. Note that if L_X is present, then X is present. Note also that for the leaky-noisy OR, the link probability $Link(D_i, X)$ represents the probability that X is present given that D_i is present and all other predecessors *including the leak node* are absent.

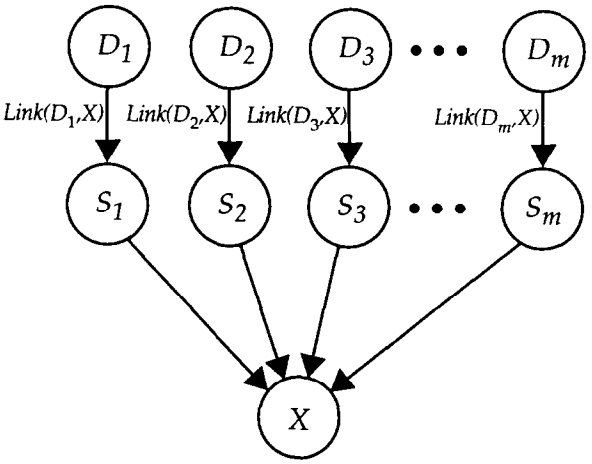

Fig. 4. A node X with predecessors $D_1, \ldots, D_m$ and shadows $S_1, \ldots, S_m$.

By incorporating a leak event into Eq. (3), we arrive at a formula for the leaky-noisy OR:

$$P(X = x[1] \mid \Pi(X)) = 1 - (1 - Leak(X)) \prod_{i:D_i \in V_X} (1 - Link(D_i, X)). \tag{5}$$

3.3.3. Noisy MAX

The binary noisy OR is insufficient for the CPCS-BN application, because we need to accommodate quaternary variables. In this section, we outline how the binary version can be generalized to an n-ary version, termed the noisy MAX.

The generalization of the noisy OR was first proposed by Henrion [16]. The derivation and implementation described here and in [35] follow Henrion's work. Two related generalizations are described in [41] and [8]. The generalization of the noisy OR by Srinivas is different from the formulation described here and is used to model circuits (or other such devices) that can be functional or non-functional. In domains such as medicine, variables may take on more than two values, in which case the binary generalization of Srinivas is insufficient. The noisy MAX generalization in [8] is virtually identical to the one described here, but was derived independently. Also, the formulation in [8] is described within the context of learning models for OR gates, and its application to inference in Bayesian networks is not apparent.

Consider a generalization of the noisy OR model in which each variable domain is a finite discrete (or n-ary) state space in which the states are ordered. For example, the variables in CPCS-BN have states that are ordered by severity: absent, mild, moderate, and severe.

In a noisy OR, each predecessor D_i may be seen as having a *shadow* S_i (Fig. 4). If D_i is present, its shadow is present with probability equal to the link probability. Variable X is simply the standard, noiseless OR of the shadows. The probability of X may be computed directly from this fact. Similarly, for a noisy MAX, each predecessor

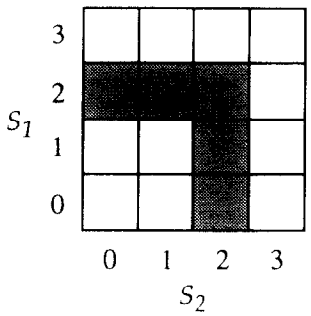

Fig. 5. A node X with predecessors D_1, D_2 and shadow variables S_1, S_2 having ordered states $\{0,1,2,3\}$. The shaded area represents the probabilities required to calculate $P(X = x[2] \mid D_1, D_2)$.

has a shadow. The probability distribution over each shadow S_i is determined by its link probabilities and the probabilities of its predecessor D_i. If the shadow predecessors were known, the value of X would be simply the maximum of the values of the shadows S_i. Hence, the name noisy MAX.

We define $S(k)$ to be the set of instantiations of the shadow variables $(S_1 = s_1, \ldots, S_m = s_m)$ such that $\max_{i=1,\ldots,m}(s_i) = k$.

We can use the noisy MAX to compute the conditional probability

$$P(X = x[k] \mid D_1, \ldots, D_m) = \sum_{\sigma \in S(k)} P(\sigma \mid D_1, \ldots, D_m)$$

$$= \sum_{\sigma \in S(k)} \prod_{i=1,\ldots,m} P(S_i = \sigma_i \mid D_i). \tag{6}$$

For example, consider the case of two predecessors D_1 and D_2, both of which have ordered states $\{0,1,2,3\}$ and corresponding shadow variables S_1 and S_2. If we want to compute $P(X = x[2] \mid D_1, D_2)$ we notice that the *maximum* state of any shadow variable must be 2. In this case the noisy MAX calculation takes into account all combinations of S_1 and S_2 in which the maximum state taken by either variable is 2. This is shown graphically in Fig. 5.

4. Experimental approach

In this section, we describe the experimental approach we used in each of the three experiments, which we will present in the following three sections. Here, we describe the three networks we used in the experiments, how we generated the test cases, and how we analyzed the results, and we provide some sample results.

4.1. Subnetwork selection

Because CPCS-BN is large and multiply connected, it is impractical to perform inference with available inference algorithms using the entire network. If we wish to compute

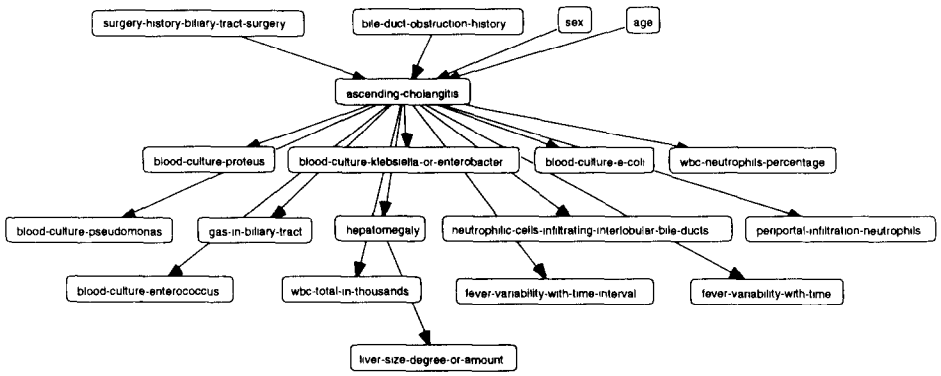

Fig. 6. Subnetwork containing the ancestors and descendants of the disease ascending-cholangitis.

Table 1
Subnetworks used for experiments, including the number of probabilities specified after expansion of the noisy MAX nodes

Network	# nodes	Max # parents	# probabilities	# parameters
BN2	42	2	492	412
BN3	146	4	2420	1480
BN4	245	7	77,988	3180
Full CPCS	365	12	90,813,182	6658

only the posterior probabilities of a small set of diseases, we can perform inference using only the subnetwork of the CPCS network that is relevant. We selected subnetworks from the full CPCS-BN using the BN graphical tool Netview [35]. Netview allows the user to display selected subsets of nodes from a network for simplicity of visualization and editing in a large network. For example, in Fig. 6, Netview displays only the immediate ancestors and descendants of the selected node—ascending-cholangitis, in this case.

We extracted three subnetworks from the full CPCS-BN for the experiments, named BN2, BN3, and BN4, containing, respectively, two, three, and four diseases. We developed versions of BN2, BN3, and BN4 in both quaternary and binary domains. Table 1 summarizes the number of nodes of each type in the three subnetworks in comparison with the full CPCS network. The table also shows the maximum number of parents for a single node in each network, the total number of conditional probabilities needed to fully specify the network (in the quaternary domain), and the number of noisy MAX parameters required to fully specify those conditional probabilities.

Fig. 7 shows BN2, and Fig. 8 shows BN3.

4.2. *Test cases*

We needed far more test cases to estimate reliably the effects of the experimental manipulations on the diagnostic performance than the small number of cases available from real patient data. Accordingly, we generated sample test cases directly from the BNs themselves, generating findings according to the probabilities specified by the network

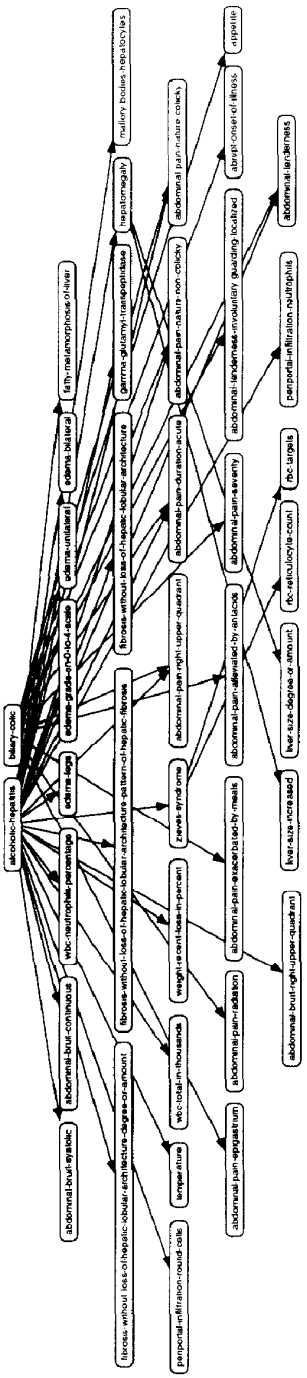

Fig. 7. The 42-node, two-disease subset of the full 450-node CPCS-BN, used as BN2 in our experiments.

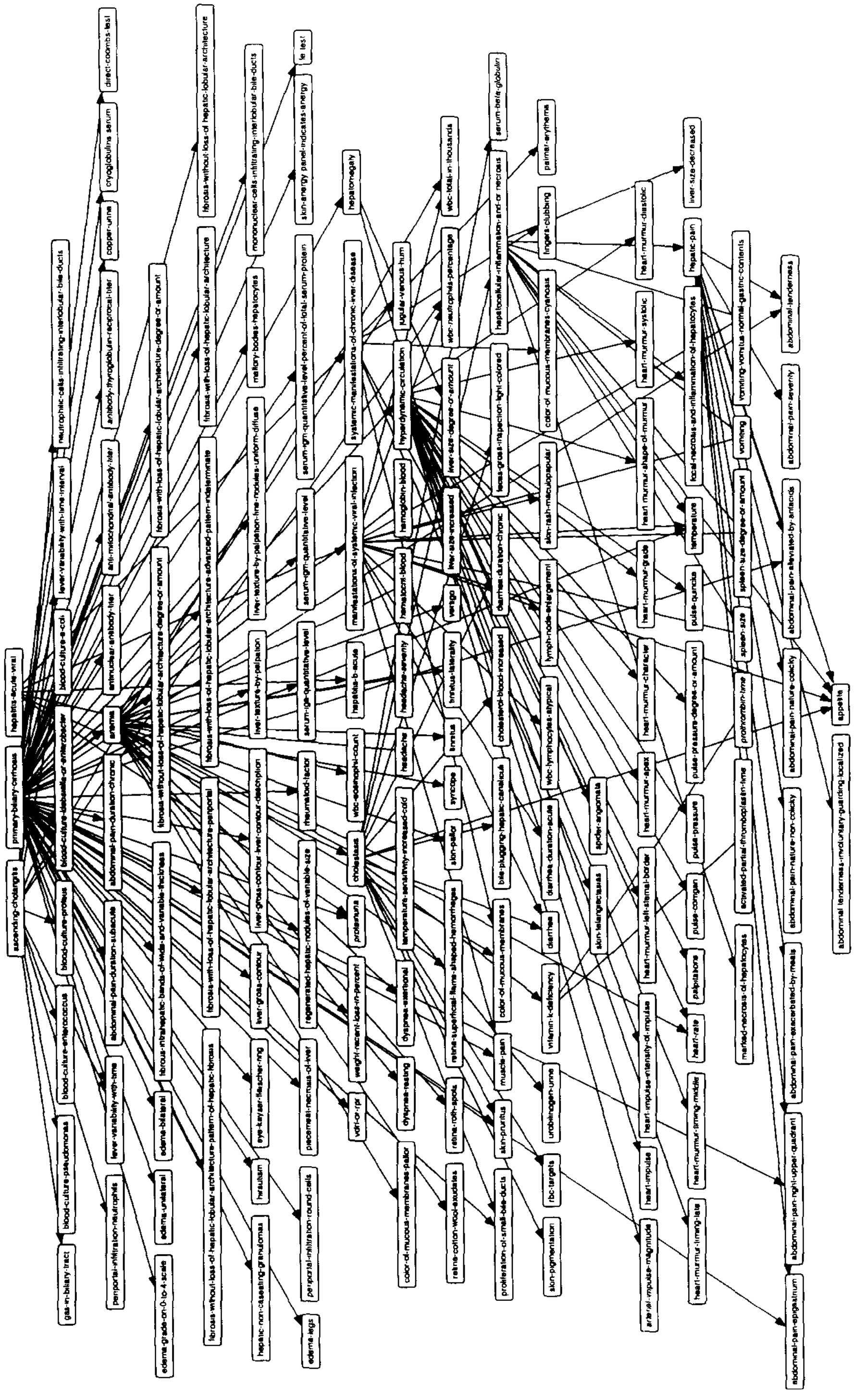

Fig. 8. The 146-node, three-disease subset of the full 450-node CPCS-BN, used as BN3 in our experiments.

using logic sampling [16]. We used the full CPCS network and the standard probability mapping for generating the test cases.

Since we wanted to investigate how the amount of evidence affects sensitivity to the experimental manipulations, we generated cases with varying numbers of findings. The test cases, as initially generated, include values for all findings. To create harder cases with fewer findings, and also for greater medical realism, we created five cases from each initial case, by revealing the findings in five phases, approximating the order in which findings would be revealed in a real medical consultation. We grouped the findings into these five phases corresponding to successive stages in medical diagnosis, as follows:

Phase 1: History, including symptoms and findings volunteered by the patient—e.g., abdominal pain in the epigastrium.

Phase 2: Examination, including objective evidence observed by the physician—e.g., abdominal tenderness.

Phase 3: Inexpensive, laboratory tests, whose results are returned in a few hours.

Phase 4: Expensive tests, non-invasive laboratory tests, whose results are returned in days.

Phase 5: Expenive, invasive laboratory tests, including pathology findings, which are usually obtained through biopsies—e.g., hepatocellular inflammation and/or necrosis.

We generated the test cases with the following five steps:

(1) First, we generated a set of disease combinations for each network in the quaternary representation using the standard probability mapping. For a network with k diseases, each at four severity levels, there are 4^k possible combinations. To reduce the number of possible combinations we restricted ourselves to severity levels "absent" or "severe". We generated combinations of these states to representative coverage of the space. In addition, we generated cases for diseases chosen at random from outside the subnetworks. As an example, for the two-disease network, we generated cases with the following disease settings (using the ordinal representation 0, 1, 2, 3 for severity levels): (0,0), (0,3), (3,0), (3,3). We used the same combination of settings to generate another set of cases with a randomly selected disease from outside the subnetwork set to level 3.

(2) For each disease combination we computed the conditional probability distributions over the findings at four severity levels. We used random sampling from the probability distributions to generate sets of finding levels to comprise each case.

(3) For the cases to be used for binary networks, we reduced the number of levels of each disease and finding from four to two, classifying all levels of severity beyond absent as present.

(4) We categorized the findings into the five phases listed above. From each full case, we generated four partial cases: Phase 1, including only Phase 1 findings; Phase 2, which adds Phase 2 findings to the Phase 1 findings; and so on, up to Phase 4. Each Phase 5 case is the full case including findings from all five phases.

There are, in general, $5n$ phased cases generated from n basic cases. Table 2 shows the number of basic and phased cases for each network.

Table 2
Number of cases used for experiments

Network	Basic cases	Number of phases	Total cases (by phase)
BN2	160	5	800
BN3	160	5	800
BN4	320	5	1600
Total	640	5	3200

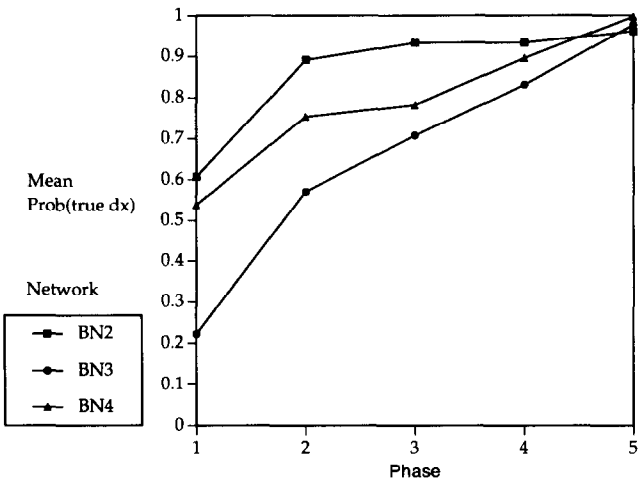

Fig. 9. The average probability of the true diagnosis as a function of the phase (amount of evidence) for each network.

4.3. Measures of diagnostic performance

We quantify diagnostic performance as the probability assigned by each network to each true diagnosis, averaged over the set of test cases. We analyze separately the probabilities assigned to each disease when present—that is, the true positive rate—and the probability assigned to each disease when absent—that is, the false positive rate.

Initially, we aggregate the results by phase so that we can see how performance varies by phase. For example, Fig. 9 plots the average probability assigned to the true diagnosis (true positive and true negative) as a function of the phase for each of the three networks. As expected, diagnostic performance improves consistently with phase—that is, the additional findings available in the later phases lead to a higher average posterior probability of the true diagnosis. Performance starts out relatively poorly for Phase 1, especially for BN3, where the average posterior probability for the true diagnosis is 0.22. But, with the entire set of evidence available in Phase 5, diagnostic performance becomes excellent, averaging 0.95 over the three networks.

For statistical analysis of the results we compared the performance of the modified networks to that of the "gold standard" network—the standard mapping. We expected the standard mapping to perform better than the modified networks because (a) test

cases were sampled from the standard networks, and (b) there is experimental evidence for the basis of the standard mappings [15].

5. Experiment on sensitivity to mappings

In our first experiment, we examined the effect of alternative mappings from the frequency weights used in CPCS (similar to those in Internist-1 and QMR) to the link probabilities used in the BN representation. The frequency weights are qualitative judgments, denoted by integers from 0 to 5, expressing the degree of connection between each cause and effect (e.g. disease and IPS, or IPS and finding) provided by the clinical diagnosticians who created those knowledge bases.

There are two reasons to vary the mappings. First, as we try to develop probabilistic reformulations of CPCS and QMR, we want to know which mapping gives best results. Second, we want to understand how sensitive the network is to changes in the mapping. More generally, we wish to understand how sensitive BNs are to changes in the numerical probabilities. In this experiment we compared the standard mapping, assessed by a principal author of QMR [15], to two extreme mappings, the curvilinear mapping and uniform mappings, as we shall describe. Our hypothesis was that use of these non-standard mappings would degrade the quality of diagnostic performance relative to the standard mapping, since we believed that the standard mapping would provide the best probabilistic interpretation of the frequency weights.

5.1. Design of mapping experiment

We compared three different mappings from frequency weights to link probabilities, as follows:

The *standard mapping* was obtained from an experiment to assess the correspondence between frequency weights and subjective probabilities [15]. For a set of disease-finding pairs, Dr. R. Miller, a principal author of QMR, assessed directly the numerical conditional probability of each finding given the presence of each disease. The experimenters found a simple and consistent relationship between the assessed probabilities and the frequency weights for the corresponding pairs. The standard mapping, as shown in Table 3, is the average probability obtained in this experiment for each frequency weight. This mapping was used in previous experiments reformulating QMR in probabilistic terms [40].

The *curvilinear mapping* provides an order-of-magnitude interpretation of the frequency weights. It interprets frequencies 0 to 3 as the orders of magnitude 0.0001 to 0.1. Frequencies 4 and 5 are orders of magnitude for the complement probability—that is, 0.9 and 0.99.

The *uniform mapping* ignores all distinctions among frequencies, mapping them all to the identical probability of 0.5. We use it to demonstrate the effect of ignoring the differences among the strength of links entirely. With the uniform mapping the strength of evidence in the network depends only on the leak values.

Table 3
Mappings used to represent frequency weight from the original CPCS knowledge base as probabilities in CPCS-BN

Mapping	Frequency					
	0	1	2	3	4	5
standard	0.0025	0.025	0.2	0.5	0.8	0.985
curvilinear	0.0001	0.001	0.01	0.1	0.9	0.99
uniform	0.5	0.5	0.5	0.5	0.5	0.5

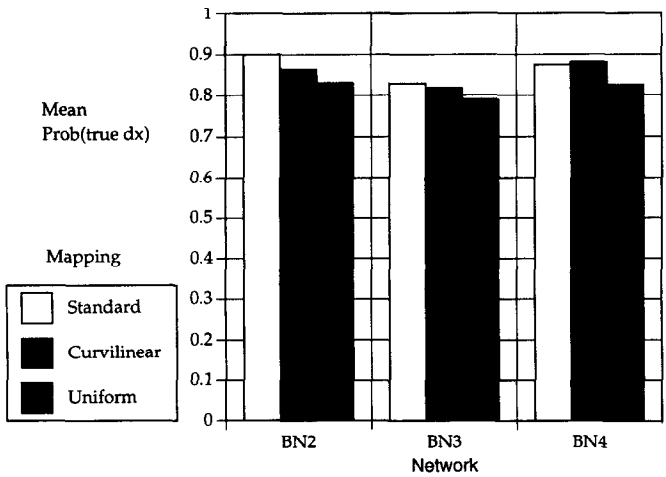

Fig. 10. The effect of alternative mappings from frequency to probability on average diagnostic performance for all phases for each network.

5.2. *Results of mapping experiment*

Fig. 10 compares the three mappings in terms of the diagnostic performance averaged over all cases for each network. As expected, the standard mapping performs best, on average; the curvilinear mapping performs next best, and the uniform mapping performs the worst. This pattern is observed for networks BN2 and BN3. For BN4 performance with standard and curvilinear mapping are almost indistinguishable. In all cases, the performance with uniform mapping is significantly worse than the standard mapping.

We also compared the three mappings with only the Phase 1 findings, to see if the sensitivity was different for cases with less evidence, as shown in Fig. 11. These cases show that the standard mapping is consistently the best, but that the curvilinear mapping can perform worse than the uniform mapping, as in BN2 and BN3.

5.3. *Discussion of mapping experiment*

The finding that the curvilinear and uniform mappings did worse on average than the standard mapping is as expected. According to the experimental calibration by Hecker-

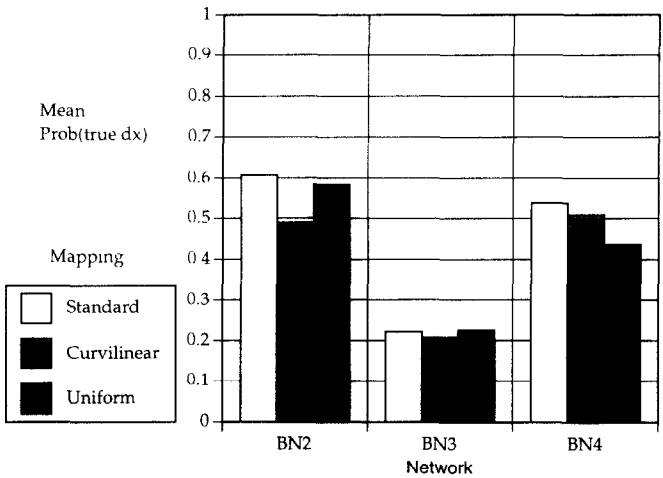

Fig. 11. The effect of alternative mappings from frequency to probability on average diagnostic performance for Phase 1 findings only.

man and Miller, the standard mapping should provide the best probabilistic interpretation of the frequency weights.

What is most interesting, however, is the modest magnitude of the decrement in performance obtained by the two very substantial modifications of the probabilities provided by the curvilinear and uniform mappings. The curvilinear mapping, with its order-of-magnitude interpretation, puts relatively much more weight on the larger frequency weights, 4 and 5, than the smaller weights, 1, 2 and 3. It reduces the importance of the findings with link strengths 0, 1, and 2 by a factor of about twenty. The uniform mapping, on the other hand, totally ignores any differences in link strengths for frequencies 1 to 5. Effectively, it makes the strength of evidence of a finding a function of the leak probability, which is not affected by mapping. In this sense, the uniform mapping reduces the representation to a purely qualitative structure for links—although, the leaks and priors remain quantitative. Despite these substantial changes, the average performance (probability of true diagnosis) is reduced by only 0.05 from (0.87 to 0.82). These results indicate very substantial robustness of diagnostic performance with respect to changes or errors in the link probabilities. They suggest that what matters most is whether there exists a link between a disease and finding. The quantitative strength of the link is of less importance.

6. Experiment on sensitivity to noise

In our second experiment, we examined how random noise in the numerical probabilities affects diagnostic performance. Numerical probabilities for belief networks may be estimated from empirical data or assessed by experts. In either case, the numbers are subject to various sources of inaccuracy and bias. For example, the data may be obtained from a sample that is not truly representative of the application domain, or the expert

may have non-representative experience. Limited sample sizes lead to random error. The process of expert assessment of probabilities is subject to a variety of inaccuracies which have been the subject of extensive study [21,26].

The question we wish to address here is how far these sources of imprecision are likely to matter. Accordingly, we add random noise to the original probabilities to simulate these sources of imprecision. In our first experiment, we examined effects of alternative probability mappings on only the links, but not the leak or prior probabilities. In the second experiment, we compared the effect of noise separately on each of the three types of probability to see whether there are different levels of sensitivity for the three types of probability.

A better understanding of sensitivity to errors or noise in numerical probability can help guide the builder of belief networks in deciding how much effort it is worth putting into probability assessment—whether probabilities are assessed directly by experts, or estimated empirically from collected patient case data. It could also help us understand what levels of precision in diagnosis we can expect given the inevitable imprecision in the input probabilities. A better understanding of the relative sensitivity to links, leaks, and priors could help guide the knowledge engineer in allocating effort in assessing these three classes of probability.

6.1. Design of noise experiment

Perhaps the most obvious way to add noise to a probability is to add a random noise directly to the probability. This approach has two problems. First, a large additive error is likely to produce a probability greater than 1 or less than 0, and so needs to be truncated. Second, an error of plus or minus 0.1 seems a lot more serious in a probability of 0.1, ranging from 0 to 0.2, than it does in a probability of 0.5, ranging from 0.4 to 0.6. Link probabilities near 0 or 1 can have enormous effects in diagnosis for findings that are present or absent (respectively).

A more appealing approach that avoids these problems is to add noise to the log-odds rather than the probability. This approach can be viewed as a version of Fechner's law of psychophysics in which similar just-noticeable differences in quantities such as weight or brightness are approximately constant when measured on a logarithmic scale. Since probability has two bounds, 0 and 1, we wish to have a symmetric effect near each bound. The log-odds transformation provides exactly this behavior.

More specifically, we transformed each probability p into log-odds form, added normal noise with a standard deviation of σ and transform back into probabilities. We define the log-odds transformation as:

$$Lo(p) = \log_{10}[p/(1-p)]. \tag{7}$$

We add log-odds noise to the probability as follows:

$$p' = Lo^{-1}[Lo(p) + e] \quad \text{where } e = Normal(0, \sigma). \tag{8}$$

We start with binary networks using the standard mapping with no noise ($\sigma = 0$), and then add noise, generated independently for each link probability in the network, with $\sigma = 1.0$, $\sigma = 2.0$, and $\sigma = 3.0$. We generated 10 noisy networks independently for

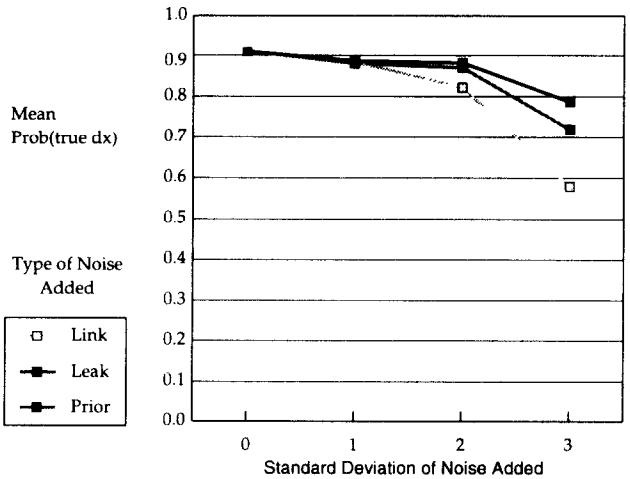

Fig. 12. Effects of noise in the link, leak, and prior probabilities on average diagnostic performance for the BN2 network.

each σ. Similarly, we created noisy networks adding noise only to the leak probabilities, and only to the prior probabilities for each network.

The total number of networks used in this experiment were 273, comprised of 3 levels of noise $\times$ 3 probability types (link, leak, and priors) $\times$ 10 samples $\times$ 3 networks, plus the original 3 standard networks without noise. For each of these networks, we ran the entire set of cases, requiring a total of 291,200 runs. As in the experiments in Section 5, we compared performance using the average probability assigned to the true diagnoses.

6.2. *Results of noise experiment*

Fig. 12 plots the average performance—the probability assigned to the true diagnosis—for the two-disease network against the four levels of noise on the link, leak, and prior probabilities. Fig. 13 and Fig. 14 plot similar measurements for the three-disease and four-disease networks, BN3 and BN4. The results are similar for all three networks. We see that, as expected, increasing noise consistently degrades performance for each type of probability—link, leak, and prior. Performance is relatively more sensitive to noise on links than to noise on priors or leaks. The effect of noise on the leaks and priors is indistinguishable for networks BN3 and BN4.

6.3. *Discussion of noise experiment*

The introduction of noise in the numerical probabilities does degrade performance, as expected. However, the amount of degradation is surprisingly small when one considers the degree of noise. Fig. 15 shows the 10-percentile and 90-percentile values of the probability with noise as a function of the probability without noise for $\sigma = 1.0$ and $\sigma = 3.0$. Even for $\sigma = 1.0$, the noise generates a wide range of probabilities. For $\sigma = 3.0$, the 80% probability interval seems to cover nearly the entire unit square. These graphs

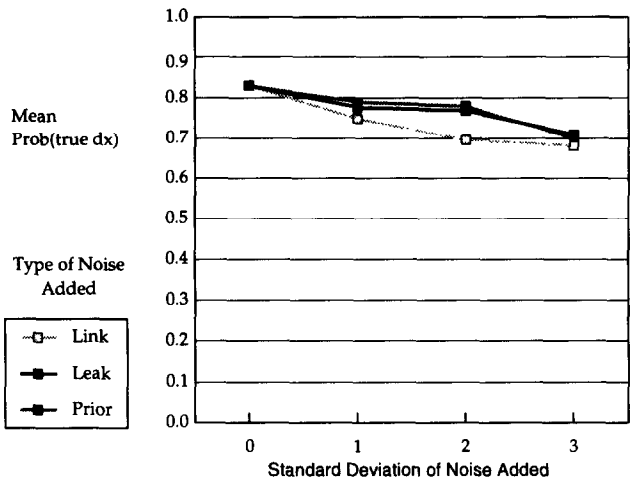

Fig. 13. Effects of noise in the link, leak, and prior probabilities on average diagnostic performance for the BN3 network.

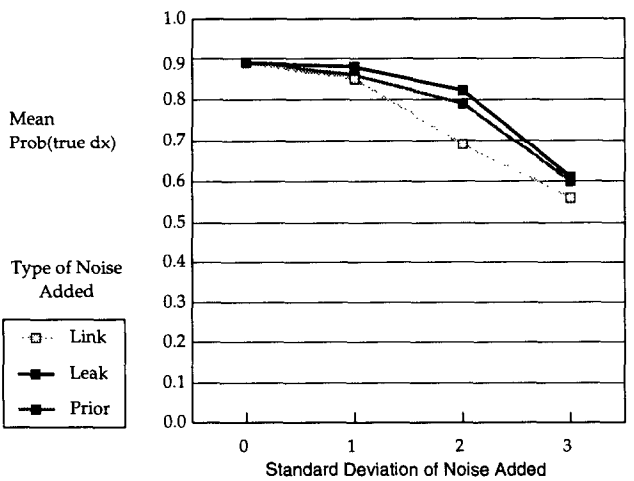

Fig. 14. Effects of noise in the link, leak, and prior probabilities on average diagnostic performance for the BN4 network.

show that noise of $\sigma = 3.0$ and greater can transform any probability into almost any other probability. In spite of this tendency, it appears these vast errors in the probabilities produce only modest degradations in performance.

6.4. *Effects of noise on true positives and negatives*

Hitherto, our analysis has combined the probabilities assigned to true positives (TP)—i.e., the probability of the disease for cases in which the disease is present—and prob-

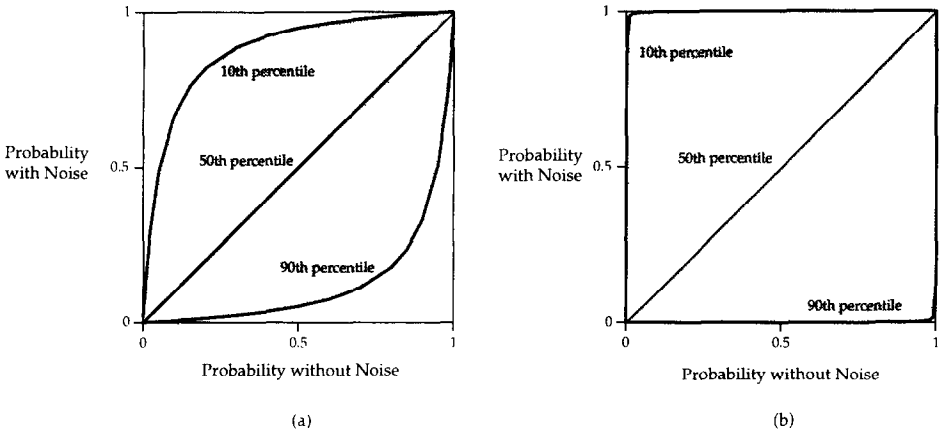

Fig. 15. Effect of adding noise: (a) $\sigma = 1$, (b) $\sigma = 3$.

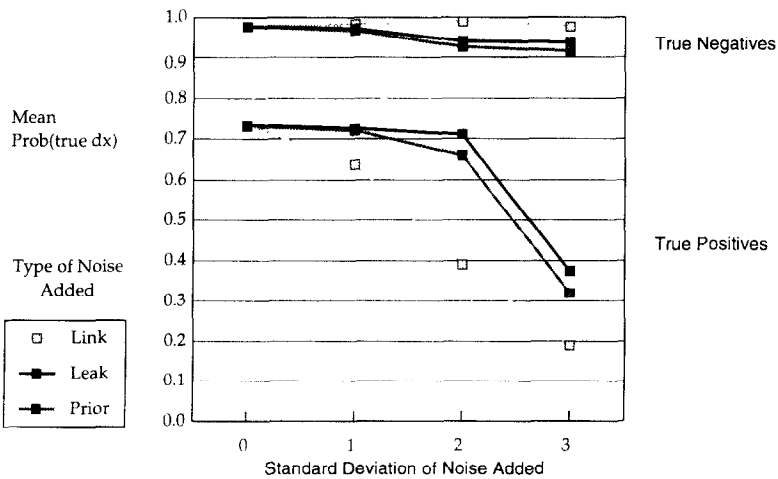

Fig. 16. Effect of noise in the link, leak, and prior probabilities on the true positive and true negative probabilities, averaged over BN2, BN3, and BN4.

abilities assigned to true negatives (TN)—i.e., the probability of no disease for cases in which the disease is absent. We can obtain interesting insights that help explain our results by examining the effects of noise on these two measures separately. Fig. 16 plots the average probability assigned to the true diagnosis separately for TP and TN, as a function of the noise level in the link, leak, and prior probabilities. These results were similar for each of the three networks. Accordingly, for simplicity, Fig. 16 shows results averaged over the three networks.

The first point to note is that, without noise, the average performance for true negatives (TN) at 0.97 is substantially better than for true positives (TP) at 0.73. In other words, the system is more likely to miss a disease that is present than to falsely diagnose a

disease that is not present. This tendency to underdiagnose should be expected because the prevalence of diseases in the test cases is much larger than would be expected according to the prior probabilities on the diseases. Note that we deliberately generated most of the test cases to contain one or more diseases to provide more information on diagnostic performance on interesting cases, even though according to the priors, more cases would have no diseases.

Now let us look at the effect of noise levels on TP and TN. Noise in the link probabilities significantly degrades performance for TP, but has no statistically detectable effect on TN ($\alpha = 0.05$). Conversely, noise in the leak probabilities has no statistically detectable effect on TP at noise levels $\sigma = 1$ and $\sigma = 2$, but link noise significantly degrades TN. Finally, noise in the priors has a similar, slight, but significant, effect in degrading performance on both TP and TN. At the highest noise setting, $\sigma = 3$, the performance of networks with leak noise and prior noise sharply decline because the disruption to the probability values is so extreme (Fig. 15).

Why should link noise and leak noise show these contrary effects on TP and TN? We can explain these results by analyzing the role of the link and leak probabilities in the diagnosis.

For simplicity, let us consider the effect of a single finding F, being present, f, or absent, $\neg f$, on the posterior odds of a single disease D. A standard measure of the strength of diagnostic evidence is the log-likelihood ratio, also known as the evidence weight:

$$EW(f, D) = \log_{10}\left[\frac{P(f \mid d)}{P(f \mid \neg d)}\right], \qquad EW(\neg f, D) = \log_{10}\left[\frac{P(\neg f \mid d)}{P(\neg f \mid \neg d)}\right].$$

$P(f \mid d)$, the probability of the finding when the disease is present is expanded using Eq. (5):

$$\begin{aligned} P(f \mid d) &= 1 - (1 - Leak(F))(1 - Link(D, F)) \\ &= Leak(F) + Link(D, F)(1 - Leak(F)). \end{aligned} \tag{9}$$

$P(f \mid \neg d)$, the probability of the finding when the disease is absent, is the leak probability, $Leak(F)$. We can now rewrite the likelihoods in terms of the link and leak probabilities:

$$EW(f, D) = \log_{10}\left[\frac{Leak(F) + Link(D, F)(1 - Leak(F))}{Leak(F)}\right], \tag{10}$$

$$EW(\neg f, D) = \log_{10}[1 - Link(D, F)]. \tag{11}$$

Notice the leak probability does not play a role in the negative evidence weight (Eq. (11)), because if a finding is absent then the leak must be off by definition.

Fig. 17 plots the evidence weights for positive and negative findings, as a function of the *link* probability, and the mean evidence weight with $\sigma = 2$ noise in the link probability. It demonstrates that, on the average, noise in the link decreases the evidence weight for the finding. This effect arises from the fact that the evidence weights are

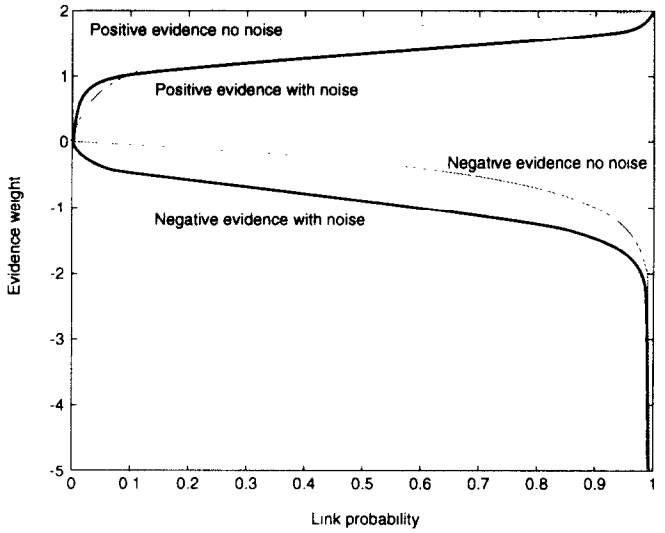

Fig. 17. Evidence weights for a disease from a single positive and negative finding as a function of the link probability, without noise and with $\sigma = 2$ noise. Note that noise tends to decrease the evidence weight for both positive and negative findings.

concave functions of the link probability. Accordingly, the noise in the links will tend to reduce the probability assigned to the true positive, reducing performance as noise increases. Noise in the links, by reducing the evidential strength of findings can only increase the probability assigned to true negative, but this effect is undetectable because the true negative rate is already high.

The impact of noise on the positive evidence (Eq. (10)) is bounded by the value of the leak. The smaller the leak, the greater the possible effect on the positive evidence. In contrast, the negative evidence weight (Eq. (11)) can be significantly decreased if the link probability is close to 1.0, as is the case with sensitive findings.

A related argument demonstrates that noise in leak probabilities will tend to increase the strength of evidence on the average. In consequence, noise in leaks also tends to increase false positives and so degrades performance for true negatives. The effect on true positives is again not detectable.

7. Experiment on sensitivity to domain size

In our third experiment, we examined the effect of the richness of the representation by comparing networks using quaternary domains, with variables at four levels—absent, mild, moderate, and severe—with networks using binary domains—absent, present. Our hypothesis was that the binary representation would degrade the diagnostic performance of the network relative to the quaternary representation. We wanted to quantify the amount of degradation.

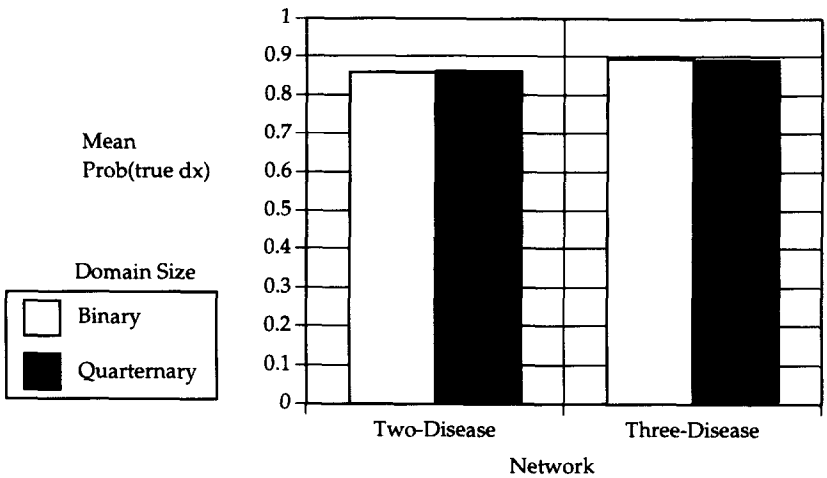

Fig. 18. Binary versus quaternary comparison for two- and three-disease networks.

Creating a quaternary network is significantly more work than a binary representation, since it requires assessment of at least three times as many probabilities. It requires three probabilities instead of one probability for each link, leak, and prior distribution; the remaining (fourth or second) probability, in each distribution is determined by the constraint that they add to unity. The computational effort for inference in quaternary networks is also significantly greater than binary networks. In this case, it was such that we could not perform the experimental runs for the four-disease network in the quaternary representation due to the excessive computation time required. The benefit of knowing the change in diagnostic precision due to changing the domain size is that it would allow knowledge base designers to make more informed decisions about what domain size is likely to be worthwhile in trading off between precision and effort in construction and computation.

7.1. Design of domain experiment

We started with the quaternary representation for the two- and three-disease networks, and reduced them to binary representations. Similarly, we reduced the quaternary test cases to binary test cases for testing the binary networks. In scoring the results, we also converted the posterior disease probabilities from the quaternary to binary representation so that we could compare directly the results from quaternary and binary networks.

7.2. Results of domain experiment

As shown in Fig. 18, we found no statistically significant difference between the diagnostic accuracies of the quaternary and binary networks. We also found no significant difference when we restricted our comparison to Phase 1 cases.

7.3. Discussion of domain experiment

The complete absence of statistically detectable difference in performance between the binary and quaternary domains was unexpected. In general terms, the finding is consistent with the findings from the preceding two experiments. The low sensitivity to changes or noise in the numerical probabilities suggests low sensitivity to changes in the complexity of the representation of the links. These results suggest that there is no reason to invest in the extra work for knowledge engineering and for computation required for the richer representation. A simple binary representation is sufficient—at least, for this domain and class of networks.

8. Experiment on outside diseases

Any real diagnostic system will have to handle cases in which the true disease or fault is not explicitly modeled in the knowledge base. The effect of incomplete knowledge bases on the reliability of diagnostic systems is often discussed, but more seldom studied. Our fourth experiment examines the effect of diseases outside the network on diagnostic performance. As we mentioned in our description of the experimental approach, we generated test cases from the entire network with twelve diseases to analyze the performance of subnetworks with two, three, and four diseases (BN2, BN3, and BN4, respectively). Half of the test cases in all the results reported above include diseases that are outside each subnetwork. In other words, the true diagnosis includes a disease not in the subnetwork. Our goal was to see how having the true disease being outside the network would affect performance. Obviously, the system cannot correctly identify a disease outside the network. The question is whether any findings inside the network which are actually caused by an outside disease will be correctly explained by the findings' leaks—a leak is a proxy for outside diseases—or whether they will be incorrectly explained by invoking a disease inside the network leading to a false positive.

8.1. Design of outside diseases experiment

As described in the section on test case generation, we generated cases using diseases from the entire twelve-disease network, including in half the cases one or more diseases from the diseases outside each subnetwork. For this analysis, we use the standard mapping without noise.

8.2. Results of outside diseases experiment

Fig. 19 shows diagnostic performance as the probability assigned to the true diagnosis, negative or positive, separately for cases which contain no diseases outside the subnetwork and for cases which do contain one or more diseases outside the network. These results are averaged over all five phases and three networks. The results are qualitatively similar for each network separately.

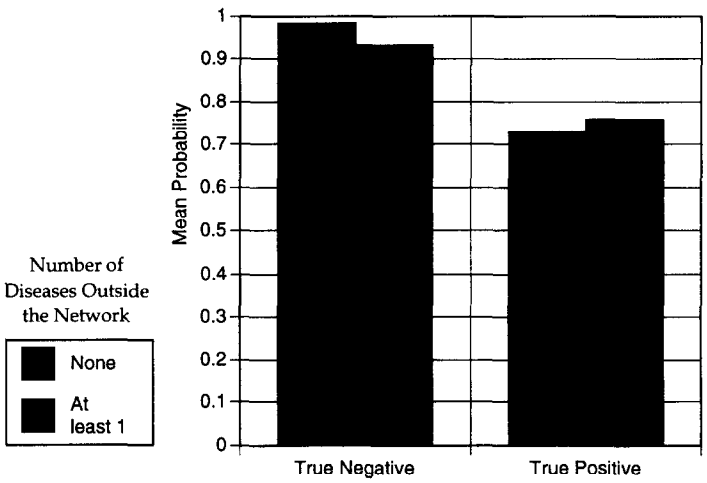

Fig. 19. Effect of diseases outside the network on the average probability assigned to true positives and true negatives.

The results for the true negative cases show almost perfect performance, 0.97, if there are no outside diseases. Performance is significantly reduced, to 0.92, by the presence of outside diseases. The results for the true positives are slightly improved, from 0.73 to 0.76, by the presence of outside diseases.

8.3. *Discussion of outside diseases experiment*

The findings for outside diseases accord with our expectations for both true negative and positive cases. For the true negative cases, outside diseases may cause findings in the network—where two or more diseases have common findings—and lead to false positives, invoked erroneously to explain the findings. For this reason, we observe that outside diseases reduce the true negative rate.

To understand how the outside diseases can improve the true positive rate, consider an outside disease that can cause a finding in the network that is also linked to a disease in the network. For a test case in which both inside and outside diseases are present, the outside disease increases the probability that the common finding will be present, which will then be interpreted as evidence for the inside disease, and so increase the true positive rate. An outside disease cannot reduce the prevalence of findings in the network, and so cannot reduce the probability of any disease in the network. Accordingly, cases with outside diseases can only increase the probability assigned to true positives, as we observe.

Although the outside diseases degrade the true negative rate and improve the true positive rate, the latter effect is significantly larger, so that overall the effect of outside diseases is to degrade performance.

9. Conclusions

In this paper, we have examined the sensitivity of several belief networks on diagnostic performance to imprecision in the representation of the numerical probabilities. Overall, we have found a surprising level of robustness to imprecision in the probabilities. Here we summarize the key findings, explore their implications, and discuss their limitations.

Extreme changes in the probability mapping from the qualitative frequency weights to numerical link probabilities had modest effects on the diagnostic performance. The curvilinear mapping, which interprets the frequencies as order-of-magnitude probabilities, provided performance that was worse than the standard mapping on average. The uniform mapping which ignores all differences in link strength degraded performance further on average, although it performed better than the curvilinear mapping with limited evidence (the Phase 1 cases).

The addition of massive amounts of random noise to the link, leak, and prior probabilities produced only modest decrements in diagnostic performance. Noise in the link probabilities had the largest effect in reducing performance for all three networks. Noise in the leak and prior probabilities had smaller effects, but performance consistently degraded with the level of noise for all three networks.

The surprisingly small effect of large amounts of random noise should be reassuring for those constructing belief networks. It provides empirical evidence that it is much more important to obtain the correct qualitative information, identifying findings, and diseases, IPSs, and their relationships, than to quantify the relations with a high level of precision. Experience suggests that domain experts are much more comfortable providing these kinds of qualitative knowledge than they are providing quantitative probabilities, although use of probability elicitation methods can make the latter more acceptable. Knowledge that high levels of precision are not necessary should greatly improve acceptance of these techniques.

The surprising lack of detectable effect of simplifying from the quaternary to binary representation for each variable, if it turns out to be general, is also good news for the BN knowledge engineer. A binary BN requires, at most, one third of the number of probabilities of a quaternary BN, assuming noisy OR and noisy MAX influences. If the network contains more complex influences, the relative advantage of binary domains increases rapidly. Moreover, small domains require much less computational effort than larger ones. Our results suggest that a binary representation may be adequate for many applications.

In our fourth experiment, we report one of the few studies to examine systematically the effect of one class of incompleteness of the knowledge base. The belief net representation, with leaky, noisy ORs and MAXes, provides a representation as leaks of the potential existence of causes (diseases or faults) that are not explicit in the knowledge base. We found that, as we expected, performance on test cases in which the diseases were missing was degraded, even for the diseases inside the network. However, the effect was moderate. An important question that we did not address was to provide an estimate of the probability that a disease was present outside the network.

Ultimately the purpose of any diagnostic system is to lead to better decisions—more cost-effective treatments of diseases, or repair to complex artifacts. In this paper, we have measured performance by accuracy of diagnosis, not by improved decisions. However, if imprecision in the representation does not degrade the diagnosis, it should not degrade the decision. In general, diagnostic accuracy is more sensitive to imprecision in the model of system being diagnosed than is the quality of the decision. Therefore, where we find that the quality of diagnosis is robust to imprecision, we can be confident that the quality of decisions will be equally or more robust.

While we believe that these results provide intriguing and suggestive evidence, we should caution that they should not be viewed as definitive for all BNs. First, note that these results are for a diagnostic application. There is reason to believe that predictive applications may show greater sensitivity. Second, these networks, like most existing large BNs, use noisy OR influences, or their generalization to noisy MAX influences. In fact, in most diagnostic belief networks constructed hitherto, the large majority of influences *are* noisy OR links. But, BNs that make extensive use of other types of influence may show different sensitivities.

Clearly, there is a need for additional work to explore these possibilities. While we believe that further experimental work is essential, we expect that theoretical analysis will also help to provide a deeper understanding of some of the findings, and suggest profitable avenues for further experimentation.

We are not the first to argue that the conclusions of diagnostic and other expert systems may have low sensitivity the imprecision in the numerical parameters. However, in heuristic representations where both the structural assumptions embody unexplicated simplifications of principles of rationality, it is often hard to separate the question of numerical approximations from structural simplifications. In the context of a probabilistic belief network, it is possible to be clear about both structural simplifications, such as independence assumptions, and the effects of numerical approximation, and so differentiate among these potential sources of error, in a way that is impossible in heuristic representations of uncertainty.

Our results lend support for the value of qualitative probabilistic representations, such as the QPNs [19,46] and infinitesimal probability schemes [13]. Indeed, we have performed some initial experimental comparisons of the performance of a BN for machine diagnosis using a qualitative infinitesimal representation (the κ calculus) with a numerical BN. We found little difference in diagnostic performance between the numerical and infinitesimal representations for cases with small fault priors [18]. The findings we have presented here help to explain the small differences between the qualitative and quantitative representations.

Acknowledgements

This work was supported by NSF grant IRI 91-20330 to the Institute for Decision Systems Research. We would like to thank Joseph Kahn for providing feedback and commentary on earlier drafts, Dr. Blackford Middleton for his help in developing CPCS-BN, and Lyn Dupré for her editorial help.

References

[1] J. Breese, Construction of belief and decision networks, Tech. Rept. 30, Rockwell International Science Center, Palo Alto, CA (1991).

[2] J. Breese and E. Horvitz, Ideal reformulation of belief networks, in: *Proceedings Sixth International Workshop on Uncertainty in AI*, Cambridge, MA (1990) 64–72.

[3] D. Chard, Mathematical methods in medical diagnosis, *Med. Decision Making* **2** (2) (1991) 69–89.

[4] D. Croft and R. Machol, Mathematical methods in medical diagnosis, *Ann. Biomed. Eng.* **2** (1987) 69–89.

[5] R.M. Dawes and B. Corrigan, Linear models in decision-making, *Psych. Bull.* **81** (1974) 95–106.

[6] F.T. de Dombal, The diagnosis of acute abdominal pain with computer assistance: worldwide perspective, *Ann. Chir.* **45** (1991) 273–277.

[7] 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 Med. J.* **2** (1972) 9–13.

[8] F.J. Diez, Parameter adjustment in Bayes networks: the generalized noisy OR-gate, in: *Proceedings Ninth Annual Conference on Uncertainty in Artificial Intelligence*, Washington, DC (1993) 99–105.

[9] F.H. Edwards and R.S. Davies, Use of a bayesian algorithm in the computer-assisted diagnosis of appendicitis, *Surg. Gynecol. Obstet.* **158** (1984) 219–222.

[10] P.C. Fishburn, A.H. Murphy and H.H. Isaacs, Sensitivity of decisions to probability estimation errors: a re-examination, *Oper. Res.* **16** (1968) 254–267.

[11] J. Fox, D. Barber and K.D. Bardhan, A quantitative comparison with rule-based diagnostic inference, *Meth. Inform. Med.* **19** (1980) 210–215.

[12] D.G. Fryback, Bayes' theorem and conditional nonindependence of data in medical diagnosis, *Comput. Biomed. Res.* **11** (1978) 423–434.

[13] M. Goldszmidt and J. Pearl, Reasoning with qualitative probabilities can be tractable, in: *Proceedings Eighth Conference on Uncertainty in AI*, Stanford, CA (1992) 112–120.

[14] D.E. Heckerman, E.J. Horvitz and B.N. Nathwani, Toward normative expert systems: Part I. The Pathfinder project, *Meth. Inform. Med.* **31** (1992) 90–105.

[15] D.E. Heckerman and R.A. Miller, Towards a better understanding of the INTERNIST-1 knowledge base, in: *Proceedings Medinfo*, Washington, DC (North-Holland, New York, 1986) 27–31.

[16] M. Henrion, Propagation of uncertainty by probabilistic logic sampling in Bayes' networks, in: J. Lemmer and L.N. Kanal, eds., *Uncertainty in Artificial Intelligence* **2** (North-Holland, Amsterdam, 1988) 149–163.

[17] M. Henrion, J.S. Breese and E.J. Horvitz, Decision analysis and expert systems, *AI Magazine* **12** (4) (1991) 64–91.

[18] M. Henrion, A. Darwiche, M. Goldszmidt, G. Provan and B. Del Favero, An experimental comparison of infinitesimal and numerical probabilities for diagnostic reasoning, in: G. Provan, ed., *Proceedings Fifth International Workshop on Principles of Diagnosis*, New Paltz, NY (1994) 131–139.

[19] M. Henrion and M.J. Druzdzel, Qualitative and linguistic explanations of probabilistic reasoning in belief networks, in: *Proceedings Sixth International Conference on Uncertainty in AI*, Cambridge, MA (1990) 10–20.

[20] F. Jensen and S.K. Andersen, Approximations in Bayesian belief universes for knowledge based systems, in: *Proceedings Sixth International Conference on Uncertainty in AI*, Cambridge, MA (1990).

[21] D. Kahneman, P. Slovic and A. Tversky, *Judgment under Uncertainty: Heuristics and Biases* (Cambridge University Press, Cambridge, 1982).

[22] U. Kjærulff, Aspects of efficiency improvements in Bayesian networks, Ph.D. Thesis, Department of Mathematics and Computer Science, Aalborg University, Aalborg (1993).

[23] B. Middleton, M. Shwe, D.E. Heckerman, M. Henrion, E.J. Horvitz, H. Lehmann and G.F. Cooper, Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base II: Evaluation of diagnostic performance, *Meth. Inform. Med.* **30** (1991) 256–267.

[24] R.A. Miller, F.E. Masarie and J.D. Myers, Quick medical reference (QMR) for diagnostic assistance, *Med. Comput.* **3** (1986) 34–48.

[25] R.A. Miller, H.E.J. Pople and J.D. Myers, INTERNIST-1: An experimental computer-based diagnostic consultant for general internal medicine, *New England J. Med.* **307** (1982) 468–476.

[26] M.G. Morgan and M. Henrion, *Uncertainty: A Guide to the Treatment of Uncertainty in Quantitative Policy and Risk Analysis* (Cambridge University Press, New York, 1990).
[27] K. Ng and B. Abramson, A sensitivity analysis of Pathfinder: a follow-up study, in: *Proceedings Seventh International Conference on Uncertainty in AI*, Los Angeles, CA (1991) 242–248.
[28] M.J. Norusis and J.A. Jacquez, Diagnosis I: symptom nonindependence in mathematical models for diagnosis, *Comput. Biomed. Res.* **8** (1975) 156–172.
[29] J. O'Neil and R. Glowinski, The ARC and AURC cooperative group: computer-aided diagnosis of acute abdominal pain when taking into account interactions, *Med. Inform.* **25** (1990) 194–198.
[30] R.C. Parker and R.A. Miller, Using causal knowledge to create simulated patient cases: the CPCS project as an extension of INTERNIST-1, in: W.W. Stead, ed., *Proceedings Eleventh Annual Symposium on Computer Applications in Medical Care*, Washington, DC (IEEE, New York, 1987) 473–480.
[31] J. Pearl, Fusion, propagation and structuring in belief networks, *Artif. Intell.* **29** (1986) 241–288.
[32] J. Pearl, *Probabilistic Reasoning in Intelligent Systems* (Morgan Kaufmann, San Mateo, CA, 1988).
[33] Y. Peng and J. Reggia, Plausibility of diagnostic hypotheses, in: *Proceedings AAAI-86*, Philadelphia, PA (1986) 140–145.
[34] D.A. Pierce and J.L. Folks, Sensitivity of Bayes procedures to the prior distribution, *Oper. Res.* **17** (1969) 344–350.
[35] M. Pradhan, G.M. Provan, B. Middleton and M. Henrion, Knowledge engineering for large belief networks, in: *Proceedings Tenth International Conference on Uncertainty in AI*, Seattle, WA (1994) 484–490.
[36] G. Provan, Tradeoffs in constructing and evaluating temporal influence diagrams, in: *Proceedings Ninth Annual Conference on Uncertainty in Artificial Intelligence*, Washington, DC (1993) 40–47.
[37] G.M. Provan, Tradeoffs in knowledge-based construction of probabilistic models, *IEEE Trans. Syst. Man Cybern.* **24** (11) (1994) 287–294.
[38] S. Sarkar, Using tree-decomposable structures to approximate belief networks, in: *Proceedings Ninth Annual Conference on Uncertainty in Artificial Intelligence*, Washington, DC (1993) 376–382.
[39] B. Seroussi, Computer-aided diagnosis of acute abdominal pain when taking into account interactions, *Meth. Inform. Med.* **25** (1986) 194–198.
[40] M. Shwe, B. Middleton, D.E. Heckerman, M. Henrion, E.J. Horvitz, H. Lehmann and G.F. Cooper, Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base I: Probabilistic model and inference algorithms, *Meth. Inform. Med.* **30** (1991) 241–255.
[41] S. Srinivas, A generalization of the noisy-or model, in: *Proceedings Ninth Annual Conference on Uncertainty in Artificial Intelligence*, Washington, DC (1993) 208–215.
[42] B.S. Todd and R. Stamper, The formal design and evaluation of medical diagnostic programs, Tech. Rept., Technical Monograph PRG-109, Oxford University Computing Laboratory, Oxford (1993).
[43] B.S. Todd and R. Stamper, The relative accuracy of a variety of medical diagnostic programs, *Meth. Inform. Med.* **33** (1994) 402–416.
[44] D. von Winterfeldt and W. Edwards, *Decision Analysis and Behavioural Research* (Cambridge University Press, Cambridge, 1986).
[45] H. Wainer, Estimating coefficients in linear models: it don't make no nevermind, *Psych. Bull.* **83** (1976) 213–217.
[46] M.P. Wellman, Graphical inference in qualitative probabilistic networks, *Networks* **20** (1990) 687–701.
[47] B.P. Wise and M. Henrion, A framework for comparing uncertain inference systems to probability, in: L.N. Kanal and J. Lemmer, eds., *Uncertainty in Artificial Intelligence* **4** (North-Holland, Amsterdam, 1986) 169–184.