A Probabilistic Reformulation of the Quick Medical Reference System

Michael Shwe, Blackford Middleton, David Heckerman, Max Henrion,*
Eric Horvitz, Harold Lehmann, and Gregory Cooper

Section on Medical Informatics, MSOB X215,
Stanford University, 94305-5479
*Rockwell Science Center, Palo Alto Laboratory,
444 High St., Suite 400, Palo Alto, CA 94301

Abstract

We report on the design and implementation of a two-level multiply-connected belief-network representation of the QMR knowledge base. We use probabilities derived from QMR disease profiles and from National Center for Health Statistics hospital-discharge statistics. Using a stochastic simulation algorithm for inference on the belief network, we compare the performance of QMR to that of the probabilistic reformulation on cases abstracted from continuing medical education materials from Scientific American Medicine.

1. Introduction

The Quick Medical Reference (QMR®) program is a decision-support tool for diagnosis in internal medicine that was developed at the University of Pittsburgh as the successor to INTERNIST-1 [1]. Designed to assist a physician in making difficult diagnoses, QMR is built on one of the largest knowledge bases (KBs) in existence. We are developing the foundation for a decision-theoretic version of QMR, which we call QMR-DT for *Quick Medical Reference—Decision Theoretic*. Decision theory is based on probability theory and utility theory. We limit our discussion in this paper to the probabilistic component of QMR-DT.

We believe there are a number of reasons for seeking probabilistic systems in medicine. By using a probabilistic model, we make explicit our assumptions—those used in building both the KB and the algorithms for inference. Moreover, such a KB is built on the well-developed and widely understood language of probability, providing researchers with a theoretical basis for creating diagnostic systems and a common vocabulary for facilitating discussion and collaboration. The use of a common language also makes it possible to share independently developed probabilistic inference algorithms and KBs. We can, for example, incorporate statistics on the local prior probability of disease in various clinical settings. For those diseases with sparse statistical data, we can use subjective estimates of prior probability. We can incrementally update these subjective probabilities as local clinical data are accumulated [2].

The current output from QMR-DT is a differential of leading diagnoses with a posterior marginal probability associated with each disease. We believe that a probabilistic differential is a more meaningful measure of belief than is a differential with heuristic scores. Furthermore, with the development of a utility model, we can use these probabilities for expected-utility decision making, thus building a decision-theoretic system on top of our probabilistic one. We hope eventually to use the results of the expected-utility component of the system for cost-effective test ordering and therapeutic planning.

Our research to date has focused on building a probabilistic foundation and method of inference for diagnosis in internal medicine. The initial goal of the project is to compare the performance of QMR to a probabilistic version of QMR, investigating the computational and representational tractability of a probabilistic approach. Our approach to developing QMR-DT is an incremental one: we build the first probabilistic model using as much of the QMR KB as possible, we test the accuracy of the inferential algorithms on this simple model, then we refine the model and algorithms successively based on the performance of the system.

In this paper, we build on the work of many researchers in general probabilistic inference, probabilistic inference in medicine, and probabilistic inference in medicine using belief networks; see [3, 4] for a review of probabilistic inference on belief networks.

In Section 2 of this paper we describe the probabilistic model of QMR-DT and the algorithms that we use for inference on the model. In Section 3 we describe a preliminary evaluation of the diagnostic performance of QMR-DT.

2. The QMR-DT Model

The QMR-DT model is built on a *belief-network* representation. The belief network is a graphical representation of probabilistic dependencies between variables [5]. More specifically, it is a directed acyclic graph in which each node represents a random variable or uncertain quantity [6]. The arcs in the graph often denote direct causal influences between variables, where the strength of the influence is specified by tables of conditional probabilities. Conversely, the absence of an arc between two nodes denotes an assertion of independence between the corresponding random variables.

We have reformulated the associations between diseases and findings of the QMR disease profiles [1] into a belief-network representation.[1] This reformulation is described in [7, 8]. The QMR-DT KB consists of a two-level belief network of n diseases and m findings, as shown in Figure 1. Each of the n diseases $\{d_1,...,d_n\}$ may be present or absent in a patient, and each of the m findings $\{f_1,...,f_m\}$ may be unobserved or observed to be present or absent. We refer to a disease hypothesis H as an assignment of presence or absence to each disease in $\{d_1,...,d_n\}$, where

---

[1] We are currently using the INTERNIST-1 KB (circa 1986), rather than the more recent QMR KB. These two KBs are quite similar, to the extent that the methods in this paper are applicable to transforming the latter KB as well. Where the distinction is inconsequential, we will refer to the INTERNIST-1 KB as the QMR KB.

$$H = \left\{ d_i = \text{present} \mid d_i \in H^+ \right\} \cup \left\{ d_i = \text{absent} \mid d_i \in H^- \right\}$$

H^+ is set of all diseases asserted to be present, and H^- is the set of all diseases asserted to be absent such that $|H^+| + |H^-| = n$.

An arc of probabilistic dependency between nodes representing a disease d and finding f exists in the QMR-DT KB if and only if there exists a link between d and f in the QMR disease profile of d. Disease-to-disease dependencies are not modeled presently in the QMR-DT KB. The current QMR-DT KB contains $n = 534$ adult diseases and $m = 4040$ findings, with 40,740 arcs depicting disease-to-finding dependencies.
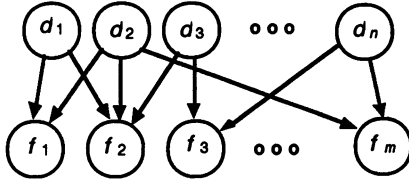

Figure 1 The two-level belief-network representation of the current QMR-DT KB. The disease nodes are labeled $d_1,...,d_n$ and the finding nodes are labeled $f_1,...,f_m$. The probabilistic dependencies between diseases and findings are specified with directed arcs between nodes, where an arc points in the causal direction that we assume; that is, we assume that diseases cause findings.

2.1 Assumptions in the Model

To reduce the representational and computational complexity of QMR-DT, we made several simplifying assumptions. Although we know that the assumptions are invalid in some cases, we are taking an incremental approach to developing the QMR-DT model: We examine the performance of the system under these assumptions with the intention of eventually modifying those that are most critical to accurate diagnostic performance. We do not, however, assume that a patient has only one disease. A primary goal of the current implementation of QMR-DT is to investigate multiple-disease diagnosis using probability as a representation of uncertainty.

Assumptions evident from Figure 1 include marginal independence of diseases, conditional independence of findings given any hypothesis of diseases, and the assumption that findings are manifestations of disease. Also, we assume that diseases and findings are binary valued. We model the influence of multiple diseases on a finding assuming causal independence. In the remainder of Section 2.1, we discuss the three different types of independence that we assume.

2.1.1 Marginal Independence of Diseases The absence of arcs among disease nodes in the belief network of Figure 1 denotes the assumption that diseases are marginally independent. Under this assumption, we can compute the probability of a disease hypothesis H from the prior probabilities of the states of the diseases in H:

$$P(H) = \prod_{d \in H^+} P(d) \prod_{d \in H^-} [1-P(d)] \tag{1}$$

2.1.2 Conditional Independence of Findings The absence of arcs among finding nodes in Figure 1 denotes the assumption that findings are conditionally independent given any disease hypothesis. Let F be a set of findings that are observed for a particular patient, where F^+ is the set of findings observed to be present and F^- is the set of findings

observed to be absent. Note that many findings may be unobserved and thus appear in neither F^+ nor F^-.

The assumption of conditional independence given any disease hypothesis allows us to compute the conditional probability of a set of findings F given a disease hypothesis H as follows:

$$P(F \mid H) = \prod_{f \in F^+} P(f \mid H) \prod_{f \in F^-} [1-P(f \mid H)] \tag{2}$$

2.1.3 Causal Independence We model the effects of multiple diseases on a single finding by assuming that the effects of the diseases on the finding are independent. This assumption, called *causal independence*, has been described by a number of researchers, including by Good [9]. We use the assumption of causal independence in the model of multicausal interactions called the noisy-OR gate [6]. Peng [10], Heckerman [7], and Henrion [8] have described the application of the noisy-OR gate to modeling the effects of diseases on manifestations. Moreover, the developers of QMR implicitly assumed a noisy-OR gate interaction [11]. Under the assumption of a noisy-OR gate, we can avoid representation of the full set of conditional probabilities of the state of a finding given each possible state of the finding's parents. Consider a belief network with binary finding f, where f has binary parents $d_1, d_2,...,d_k$. To construct the complete conditional probability table associated with the arcs from $d_1, d_2,...,d_k$ to f, we would need to acquire a conditional probability for each of the 2^k states of the parents of f. If we assume causal independence, we need to acquire only k conditional probabilities of the form $P(f \mid \text{only } d_i)$,[2] where $1 \le i \le k$.

As its name implies, the causal-independence assumption maintains that the mechanisms by which diseases cause a finding operate independently of one another and independently of any other events, such as other findings. Figure 2 shows a belief network for a noisy-OR interaction between two diseases d_1 and d_2 on a single finding f, as depicted in [7]. Assuming causal independence and binary findings, we can model the influence of multiple diseases on a finding using the noisy-OR gate:

$$P(\bar{f} \mid d_1, d_2) = P(\bar{f} \mid \text{only } d_1) \, P(\bar{f} \mid \text{only } d_2) \tag{3}$$

The intuition behind Equation 3 is that the probability of a finding f not occurring given a hypothesis H (where $H^+ = \{d_1, d_2\}$ in Equation 3) is just the probability that, of the two mechanisms that can cause f to occur, neither succeeds. Because we have modeled the findings as binary variables, we can rewrite Equation 3 as

$$P(f \mid d_1, d_2) = 1 - [1-P(f \mid \text{only } d_1)] \, [1-P(f \mid \text{only } d_2)] \tag{4}$$

In the more general case of a disease hypothesis H, we can compute the probability of f given H as follows:

$$P(f \mid H) = 1 - \prod_{d_i \in H^+} [1-P(f \mid \text{only } d_i)] \tag{5}$$

---

[2] We distinguish $P(f \mid \text{only } d_i)$ from $P(f \mid d_i)$, where the former denotes the probability of the event that f occurs given that only d_i occurs, and that, for all $j \ne i$, d_j is absent. By contrast, we use the notation $P(f \mid d_i)$ to mean the probability of the event that f occurs given that d_i occurs and for all $j \ne i$, each d_j occurs based on its prior probability.
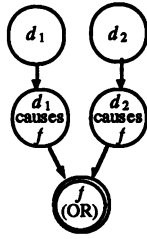
Figure 2 A belief network depicting the causal-independence assumption of the noisy-OR gate. This network depicts the assumption that d_1 and d_2 cause f through intermediate mechanisms that are independent of each other. Our belief that d_1 initiates the mechanism "d_1 causes f" with probability $P(f | \text{only } d_1)$ is represented by the arc from the node labeled "d_1" to the node labeled "d_1 causes f." The absence of an arc between the nodes labeled "d_1 causes f" and "d_2 causes f" represents an assumption of causal independence. That is, the probability that the mechanism "d_1 causes f" is active is not affected by whether the mechanism "d_2 causes f" is active. Also, the absence of the arc from the node "d_2" to the node "d_1 causes f" represents the causal-independence assumption that the probability that the mechanism "d_1 causes f" is active is not affected by the presence or absence of d_2. The same causal-independence assumptions apply to the mechanism by which d_2 causes f. The node with the double boundary is a deterministic node, which represents the belief that, if either of the two intermediate mechanisms occurs, f will be present with certainty.

2.2 Probabilities Used in the Model

The necessary probabilities for our two-level belief network include the prior probabilities of diseases and the conditional probabilities relating diseases to findings. We describe the derivation of each of these probabilities in turn.

2.2.1 Prior Probabilities of Diseases
Our probabilistic model requires that prior probabilities on diseases be made explicit. We derived prior probabilities on diseases in the QMR-DT KB from data compiled by the National Center for Health Statistics (NCHS) on 192,000 inpatients discharged from short-stay nonfederal hospitals in 1984 [12]. The diseases in the NCHS statistics are classified by the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) [13] coding system, but the INTERNIST-1 disease names do not always correspond directly to an ICD-9-CM name; therefore, we developed an approximate mapping between INTERNIST-1 names and ICD-9-CM codes. In about 60 percent of the cases, there was a close match between an INTERNIST-1 disease label and an ICD-9-CM disease label. We manually reviewed the mapping for gross outliers and modified the prior probabilities based on our subjective estimates.

2.2.2 Mapping of QMR Frequencies to Probabilities
The developers of QMR define a *frequency* between a specific disease and finding as a measure of "how often patients with the disease have the finding" [1] (p. 489). We obtained probability estimates of the form $P(f | \text{only } d_i)$ for frequency values of 1 to 5 from R. Miller, one of the primary developers of INTERNIST-1 and QMR. The results of this mapping appear in Table 1.

Table 1 A mapping between QMR frequencies and probabilities

QMR frequency	$P(f | \text{only } d_i)$
1	0.025
2	0.20
3	0.50
4	0.80
5	0.985

2.3 Algorithms for Inference
Given a set of positive and negative findings F and a model of the dependencies between diseases and findings in internal medicine, our goal is to compute $P(d_i | F)$, the posterior marginal probability for each disease $d_i : 1 \leq i \leq n$. We contrast $P(d_i | F)$ with the posterior probability of a single-disease hypothesis, $P(\text{only } d_i | F, \mu)$, where μ is the assertion that diseases are mutually exclusive. This assumption is clearly not applicable to the general problem of diagnosis in internal medicine, where patients often have several diseases simultaneously. The posterior marginal probability, on the other hand, implicitly acknowledges the existence of one or more diseases. In the next subsection, we first discuss the complexity of calculating $P(\text{only } d_i | F, \mu)$, and then describe the complexity of calculating the more general $P(d_i | F)$.

2.3.1 Exact Algorithms
We refer to Bayes' rule under the assumptions of single-disease hypotheses and conditional independence of findings as *tabular Bayes' rule* [3]:

$$P(\text{only } d_i | F, \mu) = \frac{P(F | \text{only } d_i) \, P(\text{only } d_i)}{\sum_{k=1}^{n} P(F | \text{only } d_k) \, P(\text{only } d_k)} \quad (6)$$

where there are n diseases and $P(F | \text{only } d_i)$ is given by Equation 2 when only d_i is present in H. Although the single-disease assumption is very restrictive, the tabular Bayes' formulation is appealing because of its low degree of computational complexity, $\Theta(nm)$.

Consider generalizing to allow the diagnostic hypothesis H to contain any subset of diseases in the KB. This generalization to multiple-disease hypotheses is consistent with the QMR-DT model. Straightforward application of Bayes' theorem to the QMR-DT two-level belief network requires that we sum over 2^n disease hypotheses:

$$P(d_i | F) = \frac{\sum_{H : d_i \in H^+} P(F | H) \, P(H)}{\sum_{H} P(F | H) \, P(H)} \quad (7)$$

The problem of probabilistic inference on two-level belief networks such as that of QMR-DT is known to be NP-hard [14]. Accordingly, we have sought to develop special-case algorithms and approximation algorithms to perform more efficient inference on the QMR-DT belief network.

2.3.2 Approximation Algorithms
Approximation algorithms compute estimates of the posterior marginal probabilities of diseases that converge in the limit to the posterior marginal probabilities implied by the QMR-DT model.

---

[3] The name *tabular Bayes' rule* is derived from the notion that we can compute $P(\text{only } d_i | F, \mu)$ as in Equation 6 from an $n \times |F|$ table of probabilities of the form $P(f_j | \text{only } d_i)$, where $1 \leq i \leq n$ and $1 \leq j \leq |F|$.

We have implemented an approximation algorithm called *likelihood weighting*, which places no a priori restrictions on the connectivity of the belief network. Our goal is to investigate the performance of likelihood weighting on the current QMR-DT belief network and then to use the algorithm on future versions of the network that contain a richer collection of dependencies (for example, dependencies among diseases). Likelihood weighting, a stochastic simulation algorithm, has been described by Fung and Chang [15] and by Shachter and Peot [16]. We have implemented a version of the likelihood-weighting algorithm in LightSpeed Pascal on a Macintosh IIci. For a detailed description of that algorithm, please see [17]. In this paper, we henceforth use *QMR-DT* to refer collectively to the QMR-DT model described in Sections 2.1 and 2.2 and the simulation algorithm that we implemented.

3. A Preliminary Comparison of QMR to QMR-DT

We compared the performance of QMR to that of QMR-DT on cases abstracted from continuing-education materials published by Scientific American Medicine (SAM). The SAM cases are created as clinical scenarios of disease by an expert in the appropriate subspecialty area. We defined the gold-standard (reference) diagnosis of each of these cases to be those diseases listed by SAM as the correct diagnosis for the case. For 23 SAM cases, each containing a single disease in the diagnosis, we recorded the ranks that QMR and QMR-DT assigned to the reference diagnosis. These ranks appear in Table 2; a summary of the ranks appears in Table 3. The QMR ranks in column 2 of Table 2 are taken from QMR's list of "potentially interesting diagnostic hypotheses" after a single iteration of the QMR scoring algorithm. We do not use the probabilities of the QMR-DT differential diagnosis in this comparison, because the QMR differential does not contain probabilities. Neither QMR nor QMR-DT was run on the 23 SAM cases prior to the study reported here.

Note that we used QMR in this study in a manner different from that intended by the system's developers. Specifically, QMR is intended to be used by a physician in an interactive mode [18]. Our use of QMR was limited to applying the QMR diagnostic algorithm once to each set of positive and negative findings. We did not provide the algorithm with additional positive or negative findings based on queries generated by the algorithm. The developers of QMR report that, even after all the positive findings for a case have been entered, the addition of negative findings (to the set of negative findings entered initially) during an interaction with a clinician can increase QMR's diagnostic accuracy [19].

Tables 2 and 3 show that QMR and QMR-DT performed comparably on the SAM cases. We used a two-sided Wilcoxon signed-rank test to investigate the hypothesis that QMR and QMR-DT differed significantly in the ranks that they assigned to the reference diagnoses. The test failed, at the $p = 0.05$ level, to reject this hypothesis.

There are various explanations for the performance of QMR-DT: the SAM test cases may not be sufficiently difficult or diverse to test the multiple-disease methods of QMR-DT, one or more of the QMR-DT modeling assumptions may be poor, or the estimates from the QMR-DT simulation algorithm may not have converged to the posterior probabilities implied by the QMR-DT model. In more detailed studies of the convergence of the algorithm, we found that repeated runs of the simulation on a test case produced similar posterior distributions [17, 20]. Accordingly, we believe that the first two possibilities are more likely.

In particular, consider SAM case 51, for which the reference diagnosis is thyroid papillary carcinoma. The poor performance of QMR-DT on this case appears to result from the generally non-specific findings in the case and the lack of

dependencies between findings in the QMR-DT belief-network model.

Table 2 Ranks assigned to the reference diagnosis of 23 SAM cases

SAM Case Number	Algorithm	
	QMR	QMR-DT
1	2	1
6	2	2
15	1	2
20	1	1
22	1	1
23[a]		103
25	3	1
27	1	1
28	1	1
29	3	9
30	5	7
31	12	24
33	2	2
34	1	4
35	1	1
37	2	2
40	1	1
42	4	2
46	1	1
47	1	1
50	1	1
51	2	57
53	3	1

[a]A blank space appears where QMR did not assign a rank to the reference diagnosis

Table 3 Summary of ranks assigned to the reference diagnosis of 23 SAM cases

Summary statistic	Algorithm			
	QMR	(%)	QMR-DT	(%)
Number in top 1	11	(48)	12	(52)
Number in top 5	21	(91)	18	(78)
Number in top 10	21	(91)	20	(87)
Number in top 20	22	(96)	20	(87)

When presented with the phyiscal and laboratory findings for SAM case 51, QMR-DT ranks subacute thyroiditis at the top of its differential. Note that this disease accounts for the finding in the SAM case of an elevated serum T3. The reference diagnosis, however, was severely penalized by the QMR-DT inference algorithm, since the disease profile of thyroid papillary carcinoma does not contain the finding of an elevated serum T3. We believe that the addition of dependencies between findings would allow QMR-DT to improve its diagnostic performance in this case, allowing the system to recognize the elevated serum T3 as being related to another finding in the case—oral contraceptive use—as the authors of the SAM case intended.

Although the QMR KB does not contain the finding-to-finding link between elevated serum T3 and oral contraceptive use, QMR ranked thyroid papillary carcinoma second in its differential diagnosis. As the majority of the evidence in the case suggests a malignant thyroid disease, the penalty that QMR imposes on thyroid papillary carcinoma for not

explaining the elevated T3 apparently does not markedly affect the diagnostic score that QMR assigns to this disease.

We are continuing to study the performance of QMR-DT on various types of diagnostic cases, attempting to isolate the reasons for QMR-DT's misdiagnoses. We are optimistic that, as we incrementally augment the QMR-DT KB to relax the current assumptions, the performance of the system will improve, especially on more difficult diagnostic cases.

Acknowledgments

References

1. Miller RA, Pople HEJ, Myers JD. Internist-1: An experimental computer-based diagnostic consultant for general internal medicine. New England Journal of Medicine 1982; 307: 468-476.

2. Spiegelhalter DJ, Franklin RCG, Bull K. Assessment, criticism and improvement of imprecise subjective probabilities for a medical expert system. In: Henrion M, ed. *Proceedings of the Fifth Workshop on Uncertainty in Artificial Intelligence.* Windsor, Ontario: 1989: 335-342.

3. Cooper GF. Current research directions in the development of expert systems based on belief networks. Applied Stochastic Models and Data Analysis 1989; 5: 39-52.

4. Horvitz EJ, Breese JS, Henrion M. Decision theory in expert systems and artificial intelligence. Journal of Approximate Reasoning 1988; 2: 247-302.

5. Howard RA, Matheson JE. Influence diagrams. In: Howard RA, Matheson JE, ed. *Readings on the Principles and Applications of Decision Analysis.* Menlo Park, CA: Strategic Decisions Group, 1981: 721-762.

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

7. Heckerman DE. A tractable inference algorithm for diagnosing multiple diseases. In: Henrion M, ed. *Proceedings of the Fifth Workshop on Uncertainty in Artificial Intelligence.* Windsor, Ontario: 1989: 174-181.

8. Henrion M. Towards efficient probabilistic diagnosis in multiply connected networks. In: Oliver RM, Smith JQ, ed. *Influence Diagrams, Belief Nets and Decision Analysis.* Chichester: Wiley, 1990: 385-407.

9. Good IJ. *Good Thinking: The Foundations of Probability and Its Applications,* Minneapolis, MN: University of Minnesota Press, 1983.

10. Peng Y, Reggia JA. A probabilistic causal model for diagnostic problem solving—part I: Integrating symbolic causal inference with numeric probabilistic inference. IEEE Transactions on Systems, Man, and Cybernetics 1987; SMC-17: 146-162.

11. Miller RA. Personal communication, 1986.

12. Lawrence L. Detailed diagnosis and surgical procedures for patients discharged from short-stay hospitals. Vital & Health Statistics 1986; Series 13: No. 82.

13. Health Care Financing Administration. *International Classification of Diseases (Ninth Revision)—Clinical Modification,* Washington, D. C.: U.S. Department of Health and Human Services, Public Health Service, 1980.

14. Cooper GF. The computational complexity of probabilistic inference using Bayesian belief networks. Artificial Intelligence 1990; 42: 393-405.

15. Fung R, Chang KC. Weighting and integrating evidence for stochastic simulation in Bayesian networks. In: Henrion M, ed. *Proceedings of the Fifth Workshop on Uncertainty in Artificial Intelligence.* Windsor, Ontario: 1989: 112-117.

16. Shachter RD, Peot M. Simulation approaches to general probabilistic inference on belief networks. In: Henrion M, ed. *Proceedings of the Fifth Workshop on Uncertainty in Artificial Intelligence.* Windsor, Ontario: 1989: 311-318.

17. Shwe MA, Middleton BF, Heckerman DE, Henrion M, Horvitz EJ, Lehmann HP, Cooper GF. *Probabilistic Diagnosis Using a Reformulation of the INTERNIST-1/QMR Knowledge Base. Knowledge Systems Laboratory Memo no. KSL-90-09.* Stanford CA: Stanford University, 1990.

18. Miller RA, McNeil MA, Challinor SM, Masarie FEJ, Myers JD. The INTERNIST-1/QUICK MEDICAL REFERENCE project—Status report. Western Journal of Medicine 1986; 145: 816-822.

19. Miller RA. Personal communication, 1990.

20. Shwe MA, Cooper GF. An empirical analysis of likelihood weighting on a large, multiply connected belief network. In: Bonisonne P, ed. *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence.* Cambridge, MA: 1990: 498-508.