ELSEVIER

Artificial Intelligence in Medicine 10 (1997) 177–200

Artificial Intelligence in Medicine

BANTER: a Bayesian network tutoring shell

Peter Haddawy [a], Joel Jacobson [b], Charles E. Kahn Jr. [c,*]

[a] *Decision Systems and Artificial Intelligence Lab, Department of EE and CS, University of Wisconsin-Milwaukee, Milwaukee, WI 53201, USA*
[b] *Introl, 301 N. Water Street, Milwaukee, WI 53202, USA*
[c] *Section of Information and Decision Sciences, Department of Radiology, Medical College of Wisconsin, Milwaukee, WI 53226, USA*

Accepted 21 October 1996

Abstract

We present an educational tool for bringing the information contained in a Bayesian network to the end user in an easily intelligible form. The BANTER shell is designed to tutor users in evaluation of hypotheses and selection of optimal diagnostic procedures. BANTER can be used with any Bayesian network containing nodes that can be classified into hypotheses, observations, and diagnostic procedures. The system enables one to present various types of queries to the network, to test one's ablity to select optimal diagnostic procedures, and to request explanations. We describe the system's capabilities by illustrating how it functions with two structurally different network models of real-world medical problems. © 1997 Elsevier Science B.V.

Keywords: Bayesian networks; Computer-aided instruction; Explanation

1. Introduction

In recent years Bayesian belief networks have become the representation of choice for building decision-making systems in domains characterized by uncertainty. The popularity of Bayesian networks has lead to a proliferation of domain

* Corresponding author. Tel.: +1 414 2592173; fax: +1 414 2599290; e-mail: ckahn@mcw.edu

PII S0933-3657(96)00374-0

models covering diverse areas [1,4,18,12,11]. While most applications of Bayesian networks are geared toward providing decision support, the large number of models both currently available and under development provides a wealth of detailed knowledge that could be used for educational purposes as well. Unfortunately, the information contained in these models is not easily intelligible to humans. A vehicle is required to make this information accessible for teaching purposes.

We present a generic Bayesian network-based tutoring shell, BANTER, which is designed to tutor users in diagnosis and in selection of optimal diagnostic procedures. BANTER can be used with any Bayesian network containing nodes that can be classified into hypotheses, observations, and diagnostic procedures. The system is designed so that the user need know nothing about Bayesian networks in order to interact with it effectively. In fact, nothing in the way the system interacts with the user would even indicate that the system is using a Bayesian network to perform its reasoning. The user needs only some knowledge of the particular domain and an elementary understanding of probability. BANTER provides the capability to

- compute the posterior probability of a hypothesis,
- determine the best diagnostic procedure to affirm ('rule in') or exclude ('rule out') a hypothesis,
- quiz the user in the selection of optimal diagnostic procedures, and
- explain the system's reasoning.

Since almost all of the system's reasoning is driven by the Bayesian network knowledge base, setting up the system to work with a new network requires only minimal effort.

2. System capabilities

We illustrate the capabilities of BANTER with two Bayesian network models: one for gallbladder disease and one for magnetic resonance imaging (MRI) diagnosis of liver lesions. We chose the gallbladder model because the complexity of its topology makes explanation generation challenging. We chose the MRI diagnosis model because it has a hypothesis node with a large number of states, which represents another challenge for explanation generation.

2.1. Gallbladder disease

When using BANTER in medical domains, the hypotheses are the primary diagnoses, the observations are divided into patient history and physical findings, and the diagnostic procedures are the various available tests. The belief network model shown in Fig. 1 was developed to analyze the effectiveness of various diagnostic imaging procedures for the diagnosis of gallstones and cholecystitis in patients with acute abdominal pain[1]. The sources of data used to construct the model and other details are described in [11].

[1] The network model assumes that all patients present with abdominal pain.

In this model, the two principal diagnoses are gallstones and cholecystitis (inflammation of the gallbladder). Four diagnoses serve as alternative causes of acute abdominal pain: appendicitis, gastroenteritis, small bowel obstruction (SBO), and abdominal pain of unknown cause. The presence of gallstones influences the probability that cholecystitis is present. The remaining nodes represent a patient's history, physical findings, and test results. The patient history consists of age and sex, which are demographic factors influencing the presence of gallstones, and the patient's reports of anorexia, vomiting, diarrhoea, obstipation, and 'similar prior symptoms.' The physical findings are rigidity, guarding, rebound tenderness, rectal tenderness, abnormal bowel sounds, temperature, and white blood cell (WBC) count. The possible states of each node are shown in the lower half of each box in Fig. 1; where not listed, the states are 'present' and 'absent.' The model contains two imaging tests for gallstones: ultrasound and computed tomography (CT). Three imaging tests are included for cholecystitis: the sonographic Murphy sign (maximal tenderness upon gallbladder compression during ultrasound examination), thickened gallbladder wall by ultrasound, and radionuclide hepatobiliary imaging ('HIDA').

Given such a network and a specification of which nodes represent history, physical findings, and diseases, BANTER works in three basic modes: query the knowledge base, quiz the user, and explain reasoning. These functions are provided through a graphical interface as shown in Fig. 2. Only the two principal diagnoses, gallstones and cholecystitis, are listed in the disease menu. The choice of which diseases to display is made when the system is configured for a given network. In

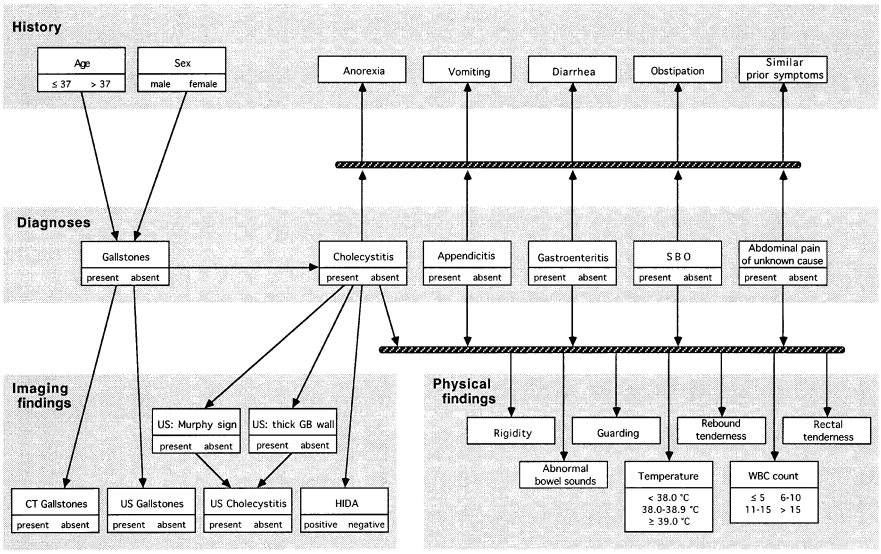

Fig. 1. Bayesian-network model of gallbladder disease. The horizontal bars simplify the figure: all five nodes that lead to the bars influence all of the nodes that lead from the bars.

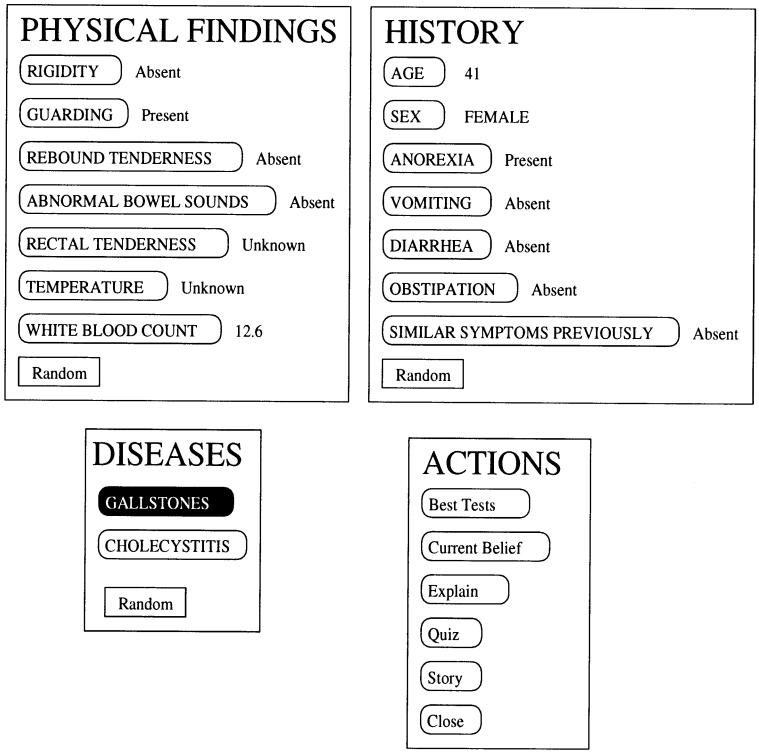

Fig. 2. BANTER graphical interface.

this case, we display only the principal diagnoses since the available tests are not useful for diagnosing the four alternative diagnoses contained in the network.

2.1.1. Querying the knowledge base

The user queries the knowledge base by setting up a scenario. A scenario is created by specifying a set of known values for the history and physical findings, as well as a set of diseases of interest. This is done by clicking on nodes in windows displaying for history, physical findings, and diseases of interest. In Fig. 2 the user has entered the following scenario. A 41-year-old woman presents with anorexia and acute abdominal pain; she denies vomiting, diarrhoea, obstipation or similar previous symptoms. Guarding is present with no rigidity or rebound tenderness. There were no abnormal bowel sounds. Her white blood cell count is 12 600 cells/cm^3. We are interested in diagnosing the presence of gallstones.

The user now can ask the system to compute the posterior probability of the selected diseases or to determine the best tests to rule in and rule out the selected diseases. Requesting the system to compute the best test produces the following response.

The best test to rule in GALLSTONES is CT.
The best test to rule out GALLSTONES is ULTRASOUND FOR GALLSTONES.

2.1.2. Requesting an explanation

The user can obtain an explanation of the reasoning that led the system to select these tests simply by selecting the 'explain' button in the actions menu. The system starts by explaining how the known history and physical findings influence the probability of gallstones.

Before presenting any evidence, the probability of GALLSTONES being present is 0.128.

The following pieces of evidence are considered important (in order of importance):

- Presence of GUARDING results in a posterior probability of 0.175 for GALLSTONES.
- AGE of 41 results in a posterior probability of 0.172 for GALLSTONES.

Their influence flows along the following paths:

- GUARDING is caused by CHOLECYSTITIS, which is caused by GALLSTONES.
- AGE influences GALLSTONES.

Presentation of the evidence results in a posterior probability of 0.227 for the presence of GALLSTONES.

Having explained how the pretest probability of gallstones was arrived at, the system continues by explaining how each possible test further influences the probability of gallstones.

The best tests to rule in GALLSTONES (in order):

- A positive CT test results in a probability of 0.987 for GALLSTONES.
- A positive ULTRASOUND FOR GALLSTONES test results in a probability of 0.601 for GALLSTONES.
- A positive HIDA test results in a probability of 0.406 for GALLSTONES.
- A positive ULTRASOUND FOR CHOLECYSTITIS test results in a probability of 0.344 for GALLSTONES.

Their influence flows along the following paths:

- GALLSTONES are seen by CT.
- GALLSTONES are seen by ULTRASOUND FOR GALLSTONES.
- GALLSTONES cause CHOLECYSTITIS, which is detected by HIDA.

- GALLSTONES cause CHOLECYSTITIS, which causes SONOGRAPHIC MURPHY SIGN, which is detected by ULTRASOUND FOR CHOLECYSTITIS.
- GALLSTONES cause CHOLECYSTITIS, which causes ULTRASOUND THICK GB WALL, which is detected by ULTRASOUND FOR CHOLECYSTITIS.

The best tests to rule out GALLSTONES (in order):

- A negative ULTRASOUND FOR GALLSTONES test results in a probability of 0.016 for GALLSTONES.
- A negative CT test results in a probability of 0.058 for GALLSTONES.
- A negative HIDA test results in a probability of 0.176 for GALLSTONES.
- A negative ULTRASOUND FOR CHOLECYSTITIS test results in a probability of 0.183 for GALLSTONES.

Their influence flows along the following paths:

- GALLSTONES are seen by ULTRASOUND FOR GALLSTONES.
- GALLSTONES are seen by CT.
- GALLSTONES cause CHOLECYSTITIS, which is detected by HIDA.
- GALLSTONES cause CHOLECYSTITIS, which causes SONOGRAPHIC MURPHY SIGN, which is detected by ULTRASOUND FOR CHOLECYSTITIS.
- GALLSTONES cause CHOLECYSTITIS, which causes ULTRASOUND THICK GB WALL, which is detected by ULTRASOUND FOR CHOLECYSTITIS.

If the user requests an explanation after asking the system to compute the posterior probability of selected diseases, only the first part of the explanation above is generated.

2.1.3. Quizzing the user

In addition to asking the system to perform computations, the user can ask to be quizzed in the selection of optimal diagnostic procedures. This can be done in two ways. By clicking on the ‘quiz’ button the user obtains the menu shown in Fig. 3. He can then specify a scenario and choose the test he thinks best to rule in or rule out the selected disease. When the user clicks on ‘test’ the system checks his answers. If an answer is incorrect, the system tells the user which tests are preferable to the one he selected. He can then request an explanation and the same type of explanation as above will be generated.

The second way the user can be quizzed is by selecting the ‘story’ action. In this mode, the system randomly selects a patient history, a set of physical findings, and a disease of interest, and presents the scenario to the user as a case in English. Below is one story the system generated.

Mrs. Jones is 36 years old, and presents with OBSTIPATION, SIMILAR-SX-PREVIOUSLY, and denies ANOREXIA, VOMITING.

```
Her temperature is 38.3. Her white blood cell count is 13800.
Physical examination reveals RIGIDITY, and no evidence of RE-
BOUND-TENDERNESS, ABNORMAL-BOWEL-SOUNDS.

What is the best test to perform to rule out GALLSTONES?
```

At this point, the user can select his answer from the quiz menu and continue as in the other quiz mode described above.

2.2. MRI diagnosis of liver lesions

Abdominal magnetic resonance imaging (MRI) plays an important role in the evaluation of liver abnormalities. The interpretation of MR images requires expert training in a rapidly changing field. DAFODILL (Decision Aid for Diagnosing Liver Lesions) is a decision-support tool designed to aid radiologists in the diagnosis of hepatic lesions seen on MRI [20]. DAFODlLL uses the Bayesian belief network shown in Fig. 4 to model its domain (see also Table 1).

The network contains a single disease node (Lesion) with 16 states, each representing a possible diagnosis. Patient history consists of age, gender, ethnic background, and whether or not the patient has a history of liver disease or another known cancer. The MRI may reveal the number and size of lesions, the intensity of a lesion relative to surrounding tissue on T1- and T2-weighted images, the T2 relaxation time, the pattern of enhancement after contrast administration, presence of a pseudocapsule, and portal vein thrombosis.

There are two major differences between this network and the previous one. The first, and most interesting, is that the disease node in this belief network has many possible states, not simply present and absent. It is with these non-hoolean type nodes that BANTER'S influence calculation is fully exercised, because influence can affect each of the various diagnoses differently. The other major difference is that this network does not contain any 'tests.' While this eliminates the need for a 'best test' scenario, it still allows for the analysis of the various diagnoses and the influences of the various pieces of evidence on those diagnoses.

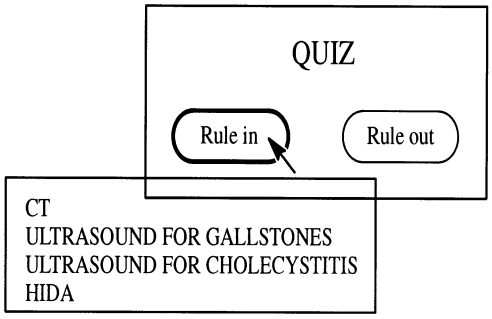

Fig. 3. Quiz menu.

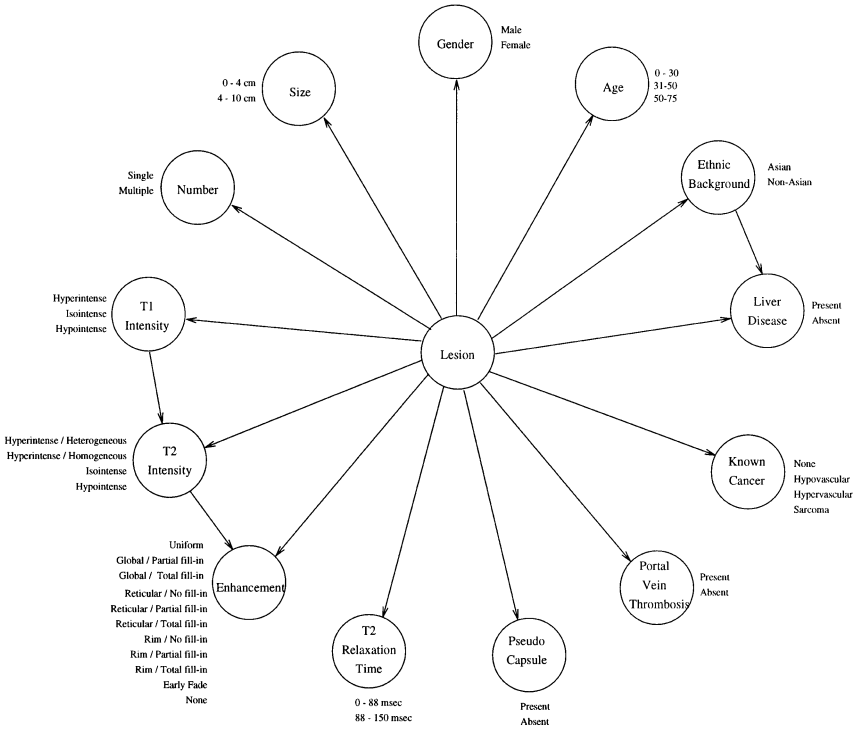

Fig. 4. Bayesian network model for MRI diagnosis of liver lesions. Adapted with permission from Tombropoulos et al.

Table 1
List of states for the liver lesion node in the DAFODILL network

Hemangioma	Benign
Hypovascular metastasis	Malignant
Hypervascular metastasis	Malignant
Sarcomatous metastasis	Malignant
Hepatocellular carcinoma	Malignant
Cholangiocarcinoma	Malignant
Lymphoma	Malignant
Focal nodular hyperplasia	Benign
Scar	Benign
Adenoma	Benign
Regenerative nodule	Benign
Focal fatty infiltration	Benign
Abscess	Infectious
Cyst	Benign
Other	Unknown
Hepatoblastoma	Malignant

As an example, we consider a 45-year-old woman with a history of liver disease. MRI examination reveals one lesion, isointense on T1-weighted images and heterogeneously hyperintense on T2-weighted images, with a T2 relaxation time of 80 ms, portal vein thrombosis, and no pseudocapsule.

Selecting 'Explain' after setting the above evidence generates the following explanation:

The following pieces of evidence have a strong influence on Lesion in the same direction as the posterior:

	Prior	Posterior	Liver disease	Portal vein thrombosis
HEMANGIOMA	0.199	0.000	0.080	0.057
HYPOVASCULAR METASTASIS	0.151	0.001	0.061	0.112
HYPERVASCULAR METASTASIS	0.151	0.001	0.061	0.112
SARCOMATOUS METASTASIS	0.081	0.001	0.032	0.060
HEPATOCELLULAR CARCINOMA	0.130	0.990	0.615	0.532
CHOLANGIOCARCINOMA	0.026	0.006	0.014	0.056
LYMPHOMA	0.017	0.000	0.010	0.005
FOCAL NODULAR HYPERPLASIA	0.017	0.000	0.010	0.005
SCAR	0.017	0.000	0.010	0.005
ADENOMA	0.017	0.000	O.010	0.005
REGENERATIVE NODULE	0.017	0.000	0.010	0.005
FOCAL FATTY INFILTRATION	0.009	0.000	0.005	0.002
ABSCESS	0.017	0.000	0.010	0.005
CYST	0.087	0.000	0.048	0.025
OTHER	0.043	0.000	0.024	0.000
HEPATOBLASTOMA	0.020	0.000	0.001	0.014

Their influence flows along the following paths:

- Liver disease is affected by Ethnic background, which influences Lesion.
- Liver disease influences Lesion.
- Portal vein thrombosis is a characteristic of Lesion.

The following pieces of evidence have a strong influence on Lesion in the direction away from the posterior:

	Prior	Posterior	Pseudo capsule
HEMANGIOMA	0.199	0.000	0.222
HYPOVASCULAR METASTASIS	0.151	0.001	0.146
HYPERVASCULAR METASTASIS	0.151	0.001	0.154
SARCOMATOUS METASTASIS	0.081	0.001	0.082
HEPATOCELLULAR CARCINOMA	0.130	0.990	0.103
CHOLANGIOCARCINOMA	0.026	0.006	0.027
LYMPHOMA	0.017	0.000	0.018
FOCAL NODULAR HYPERPLASIA	0.017	0.000	0.018
SCAR	0.017	0.000	0.018
ADENOMA	0.017	0.000	0.014
REGENERATIVE NODULE	0.017	0.000	0.018
FOCAL FATTY INFILTRATION	0.009	0.000	0.009
ABSCESS	0.017	0.000	0.018
CYST	0.087	0.000	0.093
OTHER	0.043	0.000	0.044
HEPATOBLASTOMA	0.020	0.000	0.018

Their influence flows along the following paths:

- Pseudo capsule is a characteristic of Lesion.

The following pieces of evidence have mixed but strong influence on Lesion:

	Prior	Posterior	No.	Intensity T2	T2 relaxation time	Intensity T1
HEMANGIOMA	0.199	0.000	0.278	0.058	0.062	0.293
HYPOVASCULAR METASTASIS	0.151	0.001	0.063	0.235	0.213	0.092
HYPERVASCULAR METASTASIS	0.151	0.001	0.063	0.201	0.201	0.092
SARCOMATOUS METASTASIS	0.081	0.001	0.034	0.061	0.076	0.123

HEPATO-CELLULAR CARCINOMA	0.130	0.990	0.282	0.240	0.199	0.199
CHOLAN-GIOCAR-CINOMA	0.026	0.006	0.056	0.047	0.040	0.015
LYMPHOMA	0.017	0.000	0.024	0.026	0.027	0.010
FOCAL NODULAR HYPER-PLASIA	0.017	0.000	0.010	0.020	0.027	0.005
SCAR	0.017	0.000	0.005	0.000	0.027	0.005
ADENOMA	0.017	0.000	0.015	0.016	0.027	0.010
REGENERA-TIVE NODULE	0.017	0.000	0.010	0.016	0.027	0.005
FOCAL FATTY INFILTRA-TION	0.009	0.000	0.012	0.001	0.013	0.083
ABSCESS	0.017	0.000	0.024	0.007	0.001	0.010
CYST	0.087	0.000	0.048	0.003	0.001	0.003
OTHER	0.043	0.000	0.048	0.050	0.034	0.026
HEPATOBLAS-TOMA	0.020	0.000	0.028	0.017	0.028	0.030

Their influence flows along the following paths:

- Number is a characteristic of Lesion.
- Intensity T2 is an imaging property of Lesion.
- T2 relaxation time is an imaging property of Lesion.
- Intensity T1 is an imaging property of Lesion.
- Intensity Tl is related to Intensity T2, uhich is an imaging property of Lesion.

The above explanation differs considerably from the first half of the gallstones example due to the multiple states of the hypothesis node (Lesion). The tabular printout lends itself to direct comparison between the prior probability, the posterior probability given all the evidence, and the posterior probability given each individual piece of significant evidence. It is clear that after all the evidence has been instantiated, the most appropriate diagnosis is Hepatocellular Carcinoma. Casual glances through the probabilities of the evidence nodes listed shows that the patient's history of liver disease had the strongest influence in the direction of the posterior, and the presence of pseudocapsule had the strongest contradictory influence. The section with mixed influence shows a set of evidence nodes which

have, in fact, influenced the most probable diagnosis in the correct direction, but have also influenced several other nodes in the opposite direction.

The fact that the lesion is solitary makes hemangioma and hepatocellular carcinoma more likely, and decreases the likelihood of metastases. The T2 intensity is consistent with carcinomatous (hypo- or hypervascular) metastases.

3. Algorithms

3.1. Explanation generation

Our approach to explanation generation is based on Suermondt's [19] INSITE method, according to which explanation generation consists of two procedures. The first one identifies those pieces of evidence that had the most influence on the probability of a given hypothesis. The second takes a set of evidence nodes and a hypothesis node and identifies the strongest and most comprehensible paths linking each evidence node with the hypothesis node. These algorithms are used in two different modes of the system. In the gallbladder example of Section 2.1.2, both algorithms are used to explain the current belief in the presence of gallstones. We first identify those nodes among the specified history and physical findings that were most influential in producing the reported posterior probability of the disease and then find the paths along which that influence flowed. In generating explanations for the selection of the best test, only the second algorithm is used. Here we already know that we are interested in determining the influence of each test outcome on the disease, so we only need to find the paths of influence.

3.1.1. Identifying influential pieces of evidence

In the current work, we focus on determining the impact of individual pieces of evidence on a hypothesis, since determining the impact of all possible subsets of a set evidence on a hypothesis is prohibitively complex due to the combinatorics. Consider a hypothesis H and a set of evidence $E = \{E_1, E_2, \ldots, E_n\}$ We measure the influence of a piece of evidence E_i on the hypothesis H in terms of whether and to what extent the shift from $P(H)$ to $P(H|E_i)$ agrees with the shift from $P(H)$ to $P(H|E)$. If the hypothesis can have more than two states, we can distinguish three types of important evidence nodes: those which have strong influence in the same direction as the posterior, those which have strong influence in the opposite direction of the posterior, and those that have strong but mixed influence. To define a metric for these notions, we use the information-theoretic concept of mutual information.

If A and B are two random variables with states a_i and b_j, respectively, then the information provided about the event $A = a_i$ by the event $B = b_j$ is $I(a_i; b_j) = \log(P(a_i|b_j)/P(a_i))$. A large positive value means that the evidence b_j strongly increases the probability of a_i and a large negative value means that it strongly

decreases the probability. For any state h_i of hypothesis H, we can determine whether the probability shift produced by a single piece of evidence E_i agrees or disagrees with the shift produced by the pool of evidence E by computing $I(h_i; E) \cdot I(h_i; E_i)$. A large positive value indicates that it strongly agrees and a large negative value indicates that it strongly digagrees. To determine the overall effect of a piece of evidence on the hypothesis, we simply sum over all the states:

$$\texttt{influence}(H; E; E_i) = \sum_{h_j \in H} I(h_j; E) I(h_j; E_i) \tag{1}$$

One of the consequences of this definition may not seem quite intuitive, so it stands explaining. Suppose that $P(h_1) = 0.2$, $P(h_1|E) = 0.4$, and $P(H_1|E_1) = 0.9$. According to the above measure, we would conclude that E_1 strongly influences state h_1 in the direction of the posterior $P(h_1|E)$, even though it overshoots the posterior. This is a reasonable conclusion to use in an explanation since E_1 causes the probability of h_1 to increase and the fact that $P(h_1|E)$, is not as high as $P(h_1|E_1)$ means that the presence of some other pieces of evidence tended to pull the probability of h_1 down.

The measure in expression Eq. (1) separates evidence that shifts the probability of the hypothesis toward and away from the posterior given all the evidence. It is also possible for a piece of evidence to have a strong but mixed influence on a hypothesis, shifting some states in the direction of the posterior and some away from the posterior. The influence of a piece of evidence on the hypothesis, without regard to direction can be determined by computing:

$$\texttt{impact}(H; E_i) = \sum_{h_j \in H} \left| I(h_j; E_i) \right| \tag{2}$$

Using these two measures, we identify influential pieces of evidence as follows.

Find_Influential_Evidence (H, E)

1. Remove all evidence from the network.
2. Instantiate each evidence node individually and record the posterior probability of the hypothesis.
3. Compute $\texttt{influence}(H; E; E_i)$ for each evidence node E_i.
4. $\texttt{agree-list} = \{E_i\text{: } \texttt{influence}(H; E; E_i) > 0\}$
   $\texttt{disagree-list} = \{E_i\text{: } \texttt{influence}(H; E; E_i) < \text{O}\}$
5. For each E_i in `agree-list` and `disagree-list` compute the normalized influence $\texttt{norm-infl}(E_i)$ by dividing the influence value of each element by the largest positive and largest negative value in each list, respectively.

   $\texttt{strongly-agree} = \{E_i\text{: } \texttt{norm-infl}(E_i) > \texttt{influence-threshold}\}$
   $\texttt{strongly-disagree} = \{E_i\text{: } \texttt{norm-infl}(E_i) > \texttt{influence-threshold}\}$

$+\,(E)\,+$ $-\,(\neg E)\;-+$	$-\,(E)$ $+\,(\epsilon \wedge \neg E)\;+-$	$+\,(\neg\epsilon \wedge \neg E)\,-$

Fig. 5. Traversal chart for identifying active paths.

6. For each E_i not in `strongly-agree` or `strongly-disagree`, compute `impact`(E_i) and compute the normalized values `norm-impact`(E_i) by dividing by the maximum impact value.
7. `strong-mixed` $= \{E_i$: `norm-impact`$(E_i) >$ `impact-threshold`$\}$
8. Report all elements of `strongly-agree` as strongly influencing the probability of the hypothesis in the direction of the posterior.

Report all elements of `strongly-disagree` as strongly influencing the probability of the hypothesis in the direction away from the posterior.

Report all elements of `strong-mixed` as having a strong but mixed influence on the hypothesis.

3.1.2. Identifying paths of influence

Determining the paths along which an evidence node influences a hypothesis node is a multistep process. We first identify all paths along which evidence can flow based on d-separation [16]. This set will often be too large to serve as a meaningful explanation, so we limit the explanation to five paths, first selected according to the strength of the path and then according to the length of the path. The rationale behind this scheme is that our foremost objective is to tell the user how the evidence influences the hypothesis. For the explanation to be accurate, it needs to provide the strongest paths. But we may have many paths that provide an equally strong link between evidence and hypothesis. Since a good explanation should be concise and understandable, we choose the shortest paths among those that are equally strong.

The generation of paths of influence starts with a backward search through the network to mark nodes as being predecessors of the hypothesis node or an evidence node, or both.

`MarkNodes`(*Network*)

Mark the hypothesis node and all evidence nodes, as well as all direct and indirect predecessors of the hypothesis node and the evidence nodes as *possibly related*. Mark all direct and indirect predecessors of evidence nodes as being *epsilon nodes*. These two markings can be performed in one pass through the network.

Next we use these markings to identify all paths of influence between the evidence nodes identified as important and the hypothesis node. To do this we use the chart

shown in Fig. 5. The chart uses the definition of d-separation to indicate the paths along which evidence can flow through a given node based on whether the node is an evidence node (E) or has an evidence node below it (ϵ). For example, the first entry says that if a node is an evidence node and a path enters the node along an incoming edge then it can exit the node only along another incoming edge. An incoming edge is represented by a $+$ and an outgoing edge is represented by a $-$.

We find the paths by doing a depth-first search from each important evidence node. This is initiated by calling FindlnfluentialPaths with the network, the hypothesis node, and the set of important evidence nodes.

```
FindlnfluentialPaths(Network, H, E)

For each evidence node e∈E

Let N be the set of possibly related direct parents and children of e

For each node n∈N

  If n is a parent node of e, FindPaths(Network, n, H, −, (e,n))
  Else FindPaths(Network, n, H, +, (e,n))

End

FindPaths(Network, CurrentNode, DestinationNode, Direction, Path)

If CurrentNode = DestinationNode
  Add Path to the list of paths

Else

  Given Direction, the epsilon value of CurrentNode, and whether CurrentNode is
  an evidence node, determine the set N of direct parents and children of
  CurrentNode that are linked by an allowed edge by inspecting the traversal
  chart.

  For each possibly related node n∈N

    If n is not in Path

      Add n to the end of Path

      If n is a parent of CurrentNode,
        FindPaths(Network, n, DestinationNode, −, Path)

      Else FindPaths(Network, n, DestinationNode, +, Path)

End
```

We now have a list of all active paths from the important evidence nodes to the hypothesis node. But in any sizable network, the number of active paths can be enormous. Displaying all this information would overwhelm the user. To solve this problem, we use a two-stage scheme. First, we generate only those paths of length less than a prespecified value. This value is selectable by the user and defaults to

seven. Limiting the maximum path length speeds up the chain calculation processes considerably. If we still have more than five paths linking an evidence node to the hypothesis node then we sort the paths according to strength.

We define the strength of a path using the notion that a chain is only as strong as its weakest link. Specifically, for each node N along an active path between a evidence node E_i and the hypothesis we compute `impact`(N; E_i). (The probabilities for this were obtained during the procedure that identifies influential pieces of evidence.) The strength of the chain is the minimum of the impact values. We take the minimum value because for a path to represent a strong connection between evidence and hypothesis, the impact of the evidence must be strongly transmitted along each link in the path. Since paths may share nodes, it is possible for every node but one in a path to have a large impact value. The one small impact value renders the path a weak transmission channel. This is illustrated in Fig. 6. Path (E_i, A, C, D, H) strongly connects E_i with H, while path (E_i, A, B, D, H) does not.

Using this information, we select the paths to display as follows.

1. Take the four strongest paths.
2. Starting from the fifth, take all paths with strength equal to that of the fifth.
3. Of the four paths selected in step 1, keep aside those (N) with strength different than that of the fifth.
4. For all paths that have strength equal to that of the fifth (this may include some of the four selected in step 1), sort them by length.
5. Keep shortest 5-N.

3.2. Determining the best test

Given some known physical findings and patient history data, the best test to rule in or rule out a hypothesis is determined by positively and negatively instantiating each test outcome and determining the posterior probability of the hypothesis given the test outcome (posttest probability). The best test to rule in the hypothesis is the one that results in the highest postest probability and the best test to rule out the

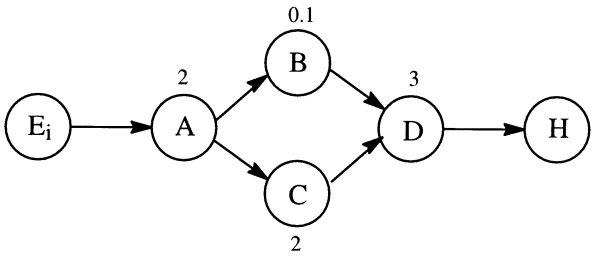

Fig. 6. Two paths of influence between evidence and hypothesis. The values next to the nodes designate the impact values.

hypothesis is the one that results in the lowest posttest probability. If the user selects more than one hypothesis of interest, this procedure is performed for each of the selected hypotheses.

4. Implementation

BANTER is designed to be both portable and easily configured for new networks. Setting up the system for a new network requires providing only the network definition file, a BANTER defnition file, and a story template file, all described below.

BANTER is implemented in C and runs on top of the HUGIN Bayesian network inference system [3]. HUGIN is used to perform all probability computations using a belief network specified in HUGIN'S network definition format. The HUGIN interface consists of a set of functions from the HUGIN libraries [9] which are used to load and compile a belief network, instantiate and uninstantiate nodes, propagate changes in individual nodes throughout the network, and obtain probability values for nodes.

4.1. The configurable GUI

BANTER'S graphical interface is written using the Xaw graphics toolkit that comes with the X11 public-domain windowing package. Configuring the interface for a new network requires providing only a BANTER definition file, which in the case of a medical domain model contains a list of nodes grouped by history, physical findings, diseases, and diagnostic procedures. Each node must be followed by a value type which can be one of 'FLOAT', 'INTEGER', 'STRING', or 'BOOLEAN'. These nodes and their corresponding classes will determine which menus each node will appear in, and what type of method will be required to set the state of a given node (BOOLEAN and STRING type nodes use a pull-down menu, and FLOAT and INTEGER type nodes use a pop-up dialog box).

4.2. Explanation generation

To generate pseudo-English language explanations, BANTER uses a links definition file that provides annotations to the links in a network, indicating the type of influence that exists between the two connected nodes. The syntax of such the file is as follows:

{node$_1$...node$_n$} phrase1:phrase2: {node$_{n1}$...node$_m$}

where phrase$_1$ indicates the type of influence from any node in the set on the left to any node in the set on the right and phrase$_2$ indicates the type of influence in the other direction. For example, the following is an excerpt from the file used in the explanations produced for the gallbladder network:

```
{AGE SEX} ''influences'':''is influenced by'' {GALLSTONES}
```

The above could generate any one of the following:

```
AGE influences GALLSTONES
SEX influences GALLSTONES
GALLSTONES is influenced by AGE
GALLSTONES is influenced by SEX
```

4.3. Story problem generation

The story template file is used to create the text for randomly generated story problems. The system generates a story problem by randomly choosing a set of values for the patient history and physical findings, randomly choosing a disease of interest, and expressing these choices by instantiating the story template. The template contains the following types of directives.

- {label:$text_1$:$text_2$:…:$text_n$}
  Check for a node named 'label', and print the appropriate text for its current state. $text_1$ corresponds to the first state listed for that node in the HUGIN definition file, $text_2$ corresponds to the second,…and $text_n$ corresponds to the UNKNOWN state. For nodes of type STRING, a '%' can be used in the text stream where the node's value will be inserted. For nodes of type RANGE, a '%.%%%' can be used in the text stream where the node's value will be inserted with the specified precision.

- < BOOLEAN:$text_1$:$text_2$ >
  Pick a random boolean value, and either print $text_1$ or $text_2$ depending on that value ($text_1$ if TRUE, $text_2$ if FALSE).

- [class]
  The 'class' can be one of HISTORY, PHYSICAL-FINDINGS, or DISEASES. A set of nodes will be selected from the specified group, excluding those nodes that have already been selected with {…}. If a node is of type BOOLEAN, its name will be included only if its current state is TRUE. If a node is of type STRING or RANGE, its name and value will always be displayed.

- (class)
  The 'class' can be one of HISTORY, PHYSICAL-FINDINGS, or DISEASES. A set of nodes will be selected from the specified group, excluding those nodes that have already been selected with {…}. If a node is of type BOOLEAN, its name will be included only if its current state is FALSE. If a node is of type STRING or RANGE, its name and value will always be displayed.

Below is the story template used in conjunction with the gallbladder network to generate the example story problem discussed in Section 2.1.3.

{SEX:Mr. Jones:Mrs. Jones:The patient} {AGE:is % years old, and} presents with [HISTORY, and denies (HISTORY).

{SEX:His:Her:The patient's} {TEMPERATURE:temperature is %.%%.}

{SEX:His:Her:The patient's} {WBC-COUNT:white blood cell count is %..} Physical examination reveals [PHYSICAL-FINDINGS], and no evidence of (PHYSICALFINDINGS).

What is the best test to perform to < BOOLEAN:rule in:rule out > [DISEASE]?

5. Related work

Recent research has addressed the issue of generating explanations in Bayesian belief networks from a number of different perspectives. Henrion and Druzdzel [13] present a method of generating scenario-based explanations of inference in belief networks. As motivation for this approach they cite psychological evidence for the prevalence of scenario-based reasoning in human thinking [17]. A scenario in the context of a belief network is complete assignment of values to a set of related random variables in the network. If we have n binary random variables then we have 2^n possible scenarios. In an attempt to make scenarios form a coherent causal story, events are ordered so that effects follow their causes.

Henrion and Druzdzel focus on the problem of generating an explanation of a single hypothesis, given some set of evidence. The hypothesis is an assignment of a value to a random variable. First those nodes relevant to the hypothesis are selected from the network. This is done using the criterion of d-separation. The selected nodes are then used to generate scenarios by enumerating all possible combinations of their values. The probability of each scenario is computed by multiplying the probabilities in the conditional probability matrices for the given set of variable values. Finally, scenarios are divided into two groups: those compatible with the hypothesis and those not compatible. A scenario is compatible with the hypothesis if it includes the particular value of the hypothesis. The explanation is then presented by listing the observed evidence, the most likely scenarios compatible with the hypothesis, the most likely scenarios incompatible with the hypothesis, and a comparison of the differences in probabilities between the most important pairs of scenarios. In general there may be an enormous number of scenarios, so they only display the most probable ones in each list.

In other work, Druzdzel and Henrion [7,13] present a method for generating qualitative verbal explanations of inference in Bayesian belief networks. They assume that the network is a singly connected collection of binary variables and focus on explaining the qualitative effect of observing the state of an evidence variable e on the probability of a target variable t. Generating a qualitative explanation of such reasoning involves three steps. First, a qualitative influence graph is extracted from the original belief network. A qualitative influence graph is a qualitative probabilistic network (QPN) consisting of only active paths from e to t. Second, an algorithm for qualitative sign propagation is invoked to mark all nodes in the qualitative influence graph with the sign of their belief change. Third, a qualitative verbal explanation is generated as a list of elementary qualitative inferences, along with any summary conclusions.

In later work, Druzdzel and Henrion [8] extend their method for qualitative propagation in QPNs to general multiply connected networks and present a polynomial time algorithm.

Madigan et al. [14] describe an approach to explanation in belief networks based on visualizing the propagation of evidence through the network. The described techniques have been implemented in a system called GRAPHICAL-BELIEF, which is a belief network modeling package [2]. They use Good's weight of evidence [10] as the basic metric for the impact of evidence and use the network itself to provide the context for describing the paths along which that impact flows.

The weight of evidence for hypothesis H provided by evidence E is defined as [10]

$$W(H{:}E) = \log \frac{P(E|H)}{P(E|\neg H)}. \tag{3}$$

The conditional probabilities in the formula can be easily computed using standard belief network inference algorithms. One simply instantiates H positively and negatively and observes the posterior probability of E.

Given a test T with outcomes $t_1,\ldots, t_n$, Madigan et al., define the *expected weight of evidence* provided by the test for a target variable as the average weight of evidence of the possible test results when the hypothesis is true:

$$EW(H{:}E) = \sum_{i=1}^{n} W(H{:}t_i)p(t_i|H). \tag{4}$$

As with the BANTER system, the explanation methods they describe can provide two types of explanations: (i) which findings were most influential on the target variable, and (ii) why a particular finding is influential or not influential. They answer the first kind of question by displaying a balance sheet displaying influential findings and their impact on the target variable. Some information in the evidence balance sheet can also be displayed directly on the belief network by coloring nodes according to their weight of evidence for the target variable.

The second type of question, why a particular finding is influential or not, is answered by displaying the paths along which its evidence flows to the target variable. In a singly connected network there will be a unique path between the finding and the target variable, called the *evidence chain*. For multiply connected networks, they provide a technique for transforming them to Berge networks [5], which are chordal graphs with clique intersections of size one. They prove that given any two nodes in such a network, the network can be collapsed to a unique evidence chain connecting the pair. Once the evidence chain is identified, the next step in displaying the evidence flow is to calculate the actual and relevant potential weights of evidence at each step in the chain. This information is then displayed graphically by depicting actual and potential evidence weights with annotations on edges and by using color to distinguish between evidence supporting the positive and negative states of the next node in the chain.

The approach to explanation generation closest to ours is that of Suermondt [19]. He presents a system called INSITE that generates explanations of diagnostic reasoning in Bayesian networks by identifying influential pieces of evidence, as well as the paths along which the influence flows.

Suermondt's [19] algorithm for identifying influential pieces of evidence uses an inverse approach from ours: rather than *instantiating* each piece of evidence individually, Suermondt *removes* each piece of evidence individually and computes the posterior probability of the hypothesis without that piece of evidence. An influential piece of evidence is then one for which the posterior probability without the evidence is significantly lower than with the evidence. By removing individual pieces of evidence, Suermondt is determining which pieces are *necessary* to influence the hypothesis, while our approach determines which pieces are *sufficient*. In this sense, the two approaches are complimentary. Because each focuses on a different type of influence, each can miss detecting influential evidence under certain circumstances.

First consider the case in which $P(H) = 0.1$, $P(H|E_1) = 0.8$, $P(H|E_2) = 0.8$, and $P(H|E_1, E_2) = 0.81$. Here each piece of evidence is sufficient to confirm H but neither is individually necessary. Suermondt's method will flag neither piece of evidence as significant, while our method will flag both as significant. In this case it is clear that our method produces the better explanation.

Consider a second case in which one piece of evidence enables a second piece to influence the hypothesis. Suppose for example that $P(H) = 0.1$, $P(H|E_1) = 0.1$, $P(H|E_2) = 0.2$, and $P(H|E_1, E_2) = 0.01$. In this case our method will not flag E_1 as significant, while Suermondt's method correctly will. Such a situation can occur when one cause explains away another cause in the presence of an observed common effect.

Finally, suppose we have two pieces of evidence E_1 and E_2 and that $P(H) = 0.1$, $P(H|E_1) = 0.8$, $P(H|E_2) = 0.6$, and $P(H|E_1, E_2) = 0.8$. Then Suermondt's method will flag only E_1 as significant, since leaving out E_2 produces no change in the probability of H. Our method will flag both pieces of evidence as significant. In such a case, a combination of both methods would be best since each piece of evidence is sufficient to confirm H but E_2 is not necessary when E_1 is present. It could be argued that if only one method is to be used then Suermondt's method produces the better explanation of influence in the presence of the current pool of evidence. Indeed, in the extreme case that E_1 blocks all paths leading from E_2 to H rendering H independent of E_1, our method would indicate that E_2 is influential but would find no paths of influence. This is due to the fact that we remove all evidence when looking for influential pieces of evidence but leave all evidence in the network when looking for paths. So our explanations indicate what pieces of evidence are influential when present alone and how the influence is exerted in the context of the current evidence pool, which may not always produce a consistent explanation. Fortunately, such pathological cases are highly unlikely to occur in practice since networks are typically designed so that observable variables occur as either root or leaf nodes.

All the above examples of anomolous behavior of the explanation algorithms depend of interaction between multiple pieces of evidence. Suermondt discusses using his technique on all possible subsets of the set of evidence in order to identify sets of evidence that may be collectively significant but for which no single element is individually relevant. As he points out, the combinatorics can make this computationally prohibitive.

Suermondt's ([19], Appendix A) algorithm for identifying direct chains from evidence to a node of interest performs the same function as our path finding algorithm but is somewhat more complex. His algorithm works in two stages: it first removes all barren nodes and nodes d-separated from the node of interest and then among the remaining nodes finds all chains from evidence nodes to the node of interest. By using the traversal chart in Fig. 5 we have combined the step for generating paths with the step for identifying d-separated nodes.

6. Summary and future research

We have presented an educational tool for bringing the information contained in a Bayesian network to the end user in an easily intelligible form. The BANTER system is designed to tutor users in evaluation of hypotheses and selection of optimal diagnostic procedures. The system will function with any network containing nodes that can be classified as hypotheses, observations, and diagnostic procedures. It provides the capability to query the network, quiz the user in the use of diagnostic procedures, and provide explanations. Since almost all the system's reasoning is performed using the Bayesian network knowledge base, configuring the system to work with a given network requires little effort. On the other hand, since nothing in the system's functionality indicates that it is using a Bayesian network for its reasoning, the complex details of the representation are hidden from the user. Thus the BANTER system provides both breadth of applicability and ease of use.

The effectiveness of BANTER can be improved along two main dimensions: allowing the system to provide more active guidance to the user and improving the quality of the explanations. BANTER currently passively waits for input from the user via the provided menus before producing a response. But in a typical tutoring situation a more natural form of interaction would have the system guide the user through a logical sequence of exercises and generally participate in a dialog with the user. One example of a system that offers more structured interaction is Buchanan et al.'s intelligent interactive system for delivering information to patients [6]. The system provides an interactive information sheet containing explanations that can be followed up on by questions from the patient. The system uses an extensive knowledge base containing both general medical terminology, as well as condition-specific knowledge to provide intelligent responses. Work is currently ongoing to add interactive dialog capabilities to the BANTER system [15].

The explanation generation capabilities of BANTER can be improved in several ways. BANTER'S explanations are currently purely text-based. Use of graphics to augment the text could make them significantly more intelligible. For example, for hypotheses with more than two states, tables are currently used to display the influence of important pieces of evidence. While an advanced user may wish to have access to this information, a bar graph would provide a more comprehensible initial display of this information.

The way in which BANTER reports paths of influence can be somewhat misleading. For example, consider the network in Fig. 6 and ignore the numbers labeling

the nodes. Suppose that E_i has a strong influence on H and that path (E_i, A, B, D, H) and path (E_i, A, C, D, H) are equally strong. Then BANTER would report each path individually. But this may be misleading if B and C must occur jointly to significantly increase the probability D. In this case a better explanation would indicate that evidence E_i increases belief in B and C, which confirms D, and hence H. One solution is to follow Madigan, etal [14] and use Berge networks to generate the explanations. On the other hand, when either B or C is sufficient to increase the probability of D, then the current form of explanation is the preferred one.

We can make explanations more informative, as well as more concise in the case of very large networks, by including abstraction information in the network model. In extremely large networks, the current technique for displaying paths of influence will produce explanations that are too lengthly to be useful. Abstraction techniques can be used to cluster nodes in the network in order to reduce the length of explanations. Additionally, abstraction can be used to provide more informative explanations: rather than providing only an explanation of the current scenario, one can move up the abstraction hierarchy to provide an explanation of more general scenarios, of which the current one is an instance. For example, in the case of cholecystitis elevating temperature, one could additionally tell the user that any inflammatory disease, of which cholecystitis is an instance, has the tendency to elevate temperature.

Acknowledgements

We thank Hugin Expert A/S for generously providing us the use of the Hugin software. Thanks to Jun Zhang for lending his expertise in information theory and to Rhea Tombropoulos for providing us with the DAFODILL model. Haddawy was supported in part by NSF grant IRI-9509165. Kahn was supported in part by the 1993 American Roentgen Ray Society Scholarship.

References

[1] B. Abramson, ARCO1: An application of belief networks to the oil market, *Proc. Seventh Conf. on Uncertainty in Artificial Intelligence* (July 1991) 1–8.

[2] R.G. Almond, *Fusion and propagation in graphical belief models: an implementation and and example*, PhD thesis, Department of Statistics, Harvard University, 1990. Report S-130.

[3] S.K. Andersen, K.G. Olesen, F.V. Jensen, and F. Jensen, HUGIN-a shell for building Bayesian belief universes for expert sytems, *Proc. Eleventh Int. Joint Conf. on Artificial Intelligence* (Detroit, Michigan, 1989) 1080–1085.

[4] S. Andreassen, M. Woldbye, B. Falck, and S.K. Andersen, MUNIN-A causal probabilistic network for interpretation of electromyographic findings, *Proc. Tenth Int. Joint Conf. on Artificial Intelligence* (Milan, Italy, August 1987) 366–372.

[5] C. Berge, *Graphs and Hypergraphs* (North Holland, Amsterdam, 1976).

[6] B.G. Buchanan, J.D. Moore, D.E. Forsythe, G. Carenini, S. Ohlsson and G. Banks, An intelligent interactive system for delivering individualized information to patients, *Artif. Intell. Med.* 7(2) (1995) 117–154.

[7] M.J. Druzdzel, Qualitative verbal explanations in bayesian belief networks, *Artif. Intell. Simulation Behav. Q.* 94 (1996) 43–54.
[8] M.J. Druzdzel and M. Henrion, Efficient reasoning in qualitative probabilistic networks, *Proc. Eleventh Nat. Conf. on Artificial Intelligence* (Washington, DC, July 1993) 548–553.
[9] L.P. Fischer and P. Abrahamsen, *Hugin API Reference Manual*. Hugin Expert A/S, Niels Jernes 10, DK-9220 Aalborg, Denmark, 1991.
[10] I.J. Good, Explicativity: a mathematical theory of explanation with statistical applications, *Proc. R. Stat. Soc. London A*, 354 (1977) 303–330.
[11] P. Haddawy, C.E. Kahn and M. Butarbutar, A Bayesian network model for radiological diagnosis and procedure selection: Work-up of suspected gallbladder disease, *Med. Phys.* 21(7) (1994) 1185–1192.
[12] D.E. Heckerman, E.J. Horvitz and B.N. Nathwani, Update on the Pathfinder project, in: L.C. Kingsland, ed., *Proc. Thirteenth Annu. Symp. on Computer Applications in Medical Care* (Los Almitos, CA, 1989) 203–207.
[13] M. Henrion and M.J. Druzdzel, Qualitative propagation and scenario-based approaches to explanation of probabilistic reasoning, in: P.P. Bonissone, M. Henrion, L.N. Kanal and J.F. Lemmer, eds., *Uncertainty in Artificial Intelligence 6*, (Elsevier, Amsterdam, 1991) 17–32.
[14] D. Madigan, K. Mosurski and R.G. Almond, Explanations in belief networks. *J. Computational Graphical Statistics*, 1996 (in press).
[15] S. Mcroy, A. Liu-Perez, J. Helwig and S. Haller, B2: A tutoring shell for bayesian networks that supports natural language interaction. *Working Notes of the AAAI Spring Symposium on Artifcial Intelligence in Medicine: Applications of Current Technologies* (Stanford, March 1996) 114–118.
[16] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Kaufmann, San Mateo, CA, 1988).
[17] N. Pennington and R. Hastie, Explanation-based decision making: Effects of memory structure on judgement, *J. Exp. Psychol.: Learning, Memory and Cognition* 14(3) (1988).
[18] M.A. Shwe, B. Middleton, D.E. Heckerman, M. Henrion, E.J. Horvitz, H.P. Lehmann and G.F. Cooper, Probabilistic diagnosis using a reformulation of the INTERNIST1/QMR knowledge base: I. The probabilistic model and inference algorithms, *Methods Inform. Med.* 30(4) (1991) 241–255.
[19] H.J. Suermondt, *Explanation in Bayesian belief networks*, PhD thesis, Medical Information Sciences, Standford University, March 1992.
[20] R. Tombropoulos, S. Shiffman and C. Davidson, A decision aid for diagnosis of liver lesions on MRI, *Proc. 17th Annu. Symp. of Computer Applications in Medical Care* (New York, 1993) 439–443.

.