Dealing with the Expert Inconsistency in Probability Elicitation

Stefano Monti and Giuseppe Carenini

Abstract—In this paper, we present and discuss our experience in the task of probability elicitation from experts for the purpose of belief network construction. In our study, we applied four techniques. Three of these techniques are available from the literature, whereas the fourth one is a technique that we developed by adapting a method for the assessment of preferences to the task of probability elicitation. The new technique is based on the Analytic Hierarchy Process (AHP) proposed by Saaty [12], [13], and it allows for the quantitative assessment of the expert inconsistency. The method is, in our opinion, very promising and lends itself to be applied more extensively to the task of probability elicitation.

Index Terms—Bayesian belief networks, analytic hierarchy process, probability elicitation.

1 Introduction

Practical experience shows that the acquisition of knowledge from experts is a costly and time-consuming task. It is very rare that an effective methodology is selected at the onset and that its application straightforwardly leads to the appropriate knowledge base. More often, considerable time is spent studying the available methodologies, iteratively applying the most promising ones and assessing and comparing their outcomes. Sometimes, when no available methodology adequately supports the current knowledge engineering task, adapting existing methodology or developing new ones is the only choice.

In this paper, we present and discuss our experience in the knowledge acquisition task of probability elicitation from experts for the purpose of belief network (BN) construction. In our study, we have applied four techniques. Three of these techniques are available from the literature [10], [19], whereas the fourth one is a technique that we developed by adapting the *Analytic Hierarchy Process* (AHP), a method for the assessment of preferences and their consistency introduced by Saaty in [12], [13].

Although our experience in probability elicitation was mainly driven by practical needs, we believe that this study has generated some valuable insights into the development and the application of knowledge acquisition techniques to the task of probability elicitation. The evolution of our investigation from practical needs to more general insights can be summarized as follows: At the onset of our project, only rough approximations of the probabilities of interest were needed. As a consequence, a simple and fast technique of probability elicitation was adopted. Subsequently, because of a shift in the goals of our project, the initial probability assessments were refined by means of more reliable techniques. The refinement process led to the discovery of significant inconsistencies in the expert's assessments. It was at this stage of our study that we started to search for more sophisticated elicitation techniques, and we eventually devised and applied a new elicitation technique that we believe to be very promising. In particular, our technique allows the analyst to measure reliably the degree of inconsistency in the expert's assessments, and to make the expert face his/her inconsistencies as soon as they arise.

The belief network which is the object of this work is aimed at modeling the domain of chronic nonorganic headaches. In order to acquire the relevant conditional probabilities in this clinical domain, we had to rely on elicitation techniques from domain experts. In fact, although information for structuring the network is abundant in the medical literature [14], [16], the medical literature does not provide sufficient information to quantify the probabilistic relationships among the domain variables. Furthermore, no data was available to our group for the automatic estimation of the probabilities of interest.[1]

The discussion of our study is organized as follows: In Section 2, we first motivate the construction of a belief network in the clinical domain of chronic nonorganic headaches. We then examine the structure of the belief network we built, and discuss the initial knowledge acquisition process for its construction. In Section 3, we describe the sensitivity analysis that focused the subsequent knowledge acquisition and the application of well-known probability elicitation techniques, based on bets and lotteries. We then discuss the elicitation technique we adapted from the *Analytic Hierarchy Process* (AHP) [12], [13]. One of the notable features of this technique is its providing for an explicit measure of the consistency of the probability assessments. In Section 4, we report a preliminary subjective evaluation of the different elicitation methodologies.

- *S. Monti is with the Robotics Institute, 3311 Doherty Hall, Carnegie Mellon University, Pittsburgh, PA 15213-3890. E-mail: smonti+@cs.cmu.edu.*
- *G. Carenini is with the Intelligent Systems Program, University of Pittsburgh, 3939 O'Hara St., Pittsburgh, PA 15260. E-mail: carenini@isp.pitt.edu.*

Manuscript received 15 Sept. 1998; revised 4 Oct. 1999; accepted 17 Dec. 1998.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number 112149.

1. Learning belief networks from data, both their structure and their parameterization, is an active field of research, and several methods have been devised to this end [3], [7], [9], [17], [8]. However, in many domains there is no availability of data on which to base the learning process.

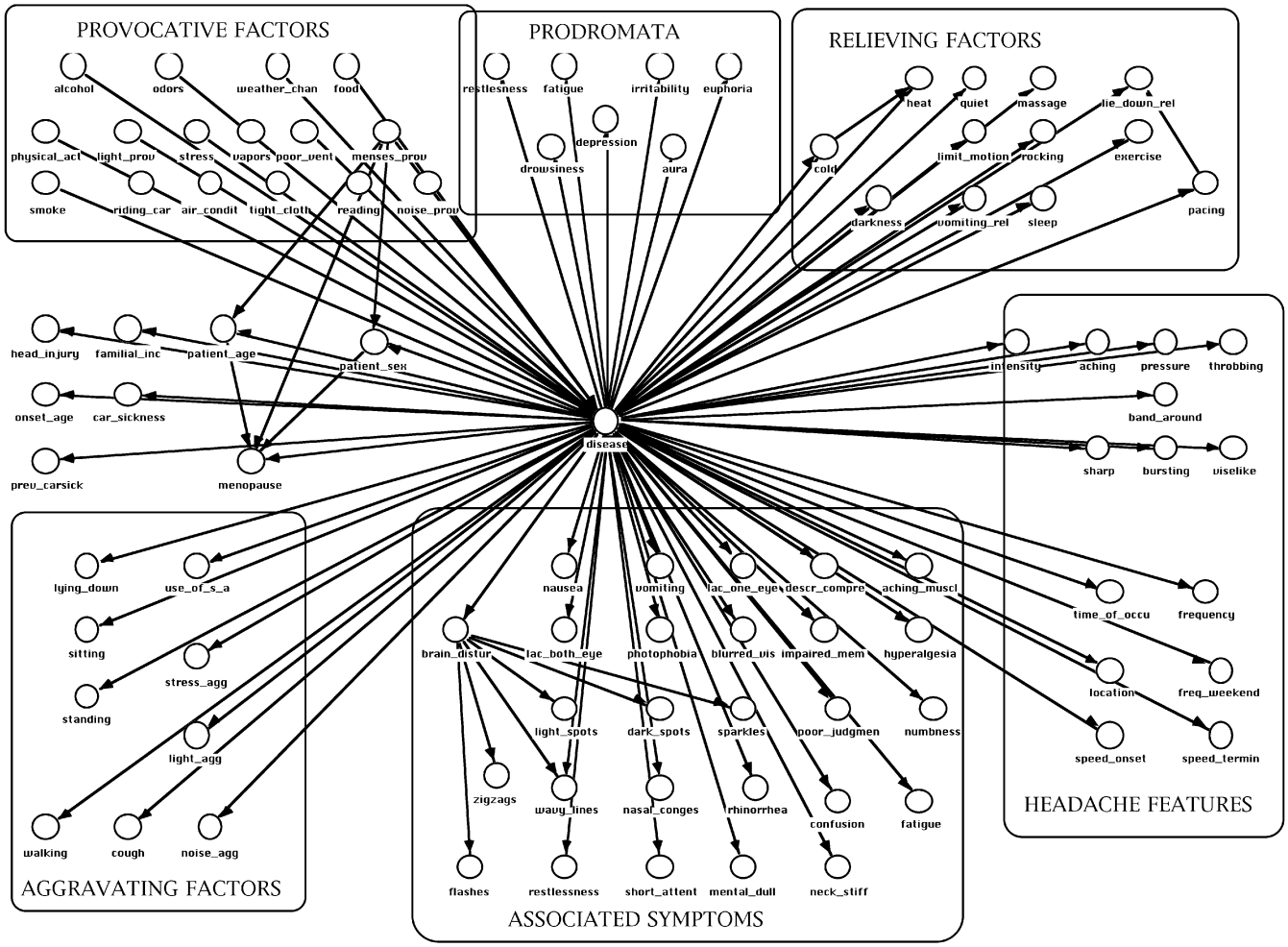


Fig. 1. The belief network for the clinical domain of chronic nonorganic headaches.

In particular, we examine the confidence of the expert with the different elicitation techniques, the expert's reaction to the techniques and the results they produced, changes in the expert's perception of the different elicitation techniques, and the expert's reaction to his inconsistencies. In the final sections, we elaborate on the possible roles of the AHP technique in the elicitation task and we draw some conclusions about our experience in the elicitation process.

2 Background

In [2], we present an information-based Bayesian strategy for history-taking, aimed at optimizing the evidence-gathering process. We apply this strategy to a history-taking module developed as part of a system for patient education in the clinical domain of chronic nonorganic headaches [1].

The knowledge base for the proposed history-taking strategy is built around a belief network that models the domain of interest. The belief network structure is as depicted in Fig. 1. We considered a mutually exclusive and exhaustive set of three diseases (hypotheses): Migraine, Cluster, and Tension headaches.[2] The assumption that these diseases are mutually exclusive is a reasonable approximation in our case, since the only significant overlap is 10 percent between migraine and tension headaches. The assumption of exhaustiveness is also reasonable because Migraine, Cluster, and Tension headaches cover about 90 percent of all nonorganic headaches.

The links diverging from the disease node correspond to findings, aggravating and relieving factors, and other patient features, whereas the links converging to the disease node are typical headache provocative factors (disease etiologies). Although many links in the network can be given a causal interpretation, this is not true for all of them. For some nodes, the connection simply states a correlation and the choice of the link direction is solely determined by the easiness in the elicitation of the corresponding conditional probabilities.

In the development of our prototype application, the elicitation of the numeric probabilities was the most difficult task. In fact, the medical literature provided sufficient information for the definition of the belief network structure [14], [16] (which was later validated by our expert). However, it did not provide the information necessary to quantify the probabilistic relationships among the domain variables. For this task, we relied exclusively on the domain expert. At this stage of the project, our main

2. The assumption of a mutually exclusive and exhaustive set of diseases allows us to group the diseases in a single node, which is the one in the center of Fig. 1.

TABLE 1
Probability Assessments for the Old Network

	Conditioning Disease		
Evidence	*Migraine*	*Tension*	*Cluster*
head injury	$0.11 - 0.33$	$0.00 - 0.1$	$0.11 - 0.33$
stress aggrav.	$0.67 - 0.99$	$0.11 - 0.33$	$0.67 - 0.99$
nasal congestion	$0.00 - 0.10$	$0.00 - 0.10$	$0.67 - 0.99$
...			

concern was the history-taking strategy rather than the accuracy of the assessments, and our expert was already familiar with a very simple elicitation technique based on the selection of his subjective probabilities from a predetermined set of six adjacent probability intervals ranging from impossible to certain.[3] Therefore, this was the method of choice. We translated the six probability intervals into point values by taking the midpoints of the given intervals, as suggested in [18]. A small subset of these subjective assessments is shown in Table 1. We considered all these approximations as acceptable since, as already pointed out, our main goal was to inform history-taking strategies rather than to automate the diagnostic process, a task for which a higher level of accuracy of the estimated probabilities would be deemed appropriate.

After a preliminary evaluation of the strategy using five patient cases, we felt it necessary to better understand the relation between the diagnostic accuracy of the belief network and the effectiveness of the history-taking strategy. This was the main motivation of the additional probability elicitation effort that we describe in this paper.

3 Methodology

This section is devoted to the description of the overall elicitation strategy we adopted. We first illustrate the sensitivity analysis we performed to focus the subsequent probability elicitation effort. We then describe, in detail, the techniques of probability elicitation we adopted, with particular attention to the adaptation of the AHP to the task at hand.

3.1 Sensitivity Analysis

In order to focus our attempt to increase the diagnostic accuracy of the belief network, we performed a simple one-way sensitivity analysis.[4] We varied the probability of each evidence node (conditioned on the disease node) from 0 to 1 (maintaining all the other probabilities constant), and we computed the corresponding variation in the probabilities

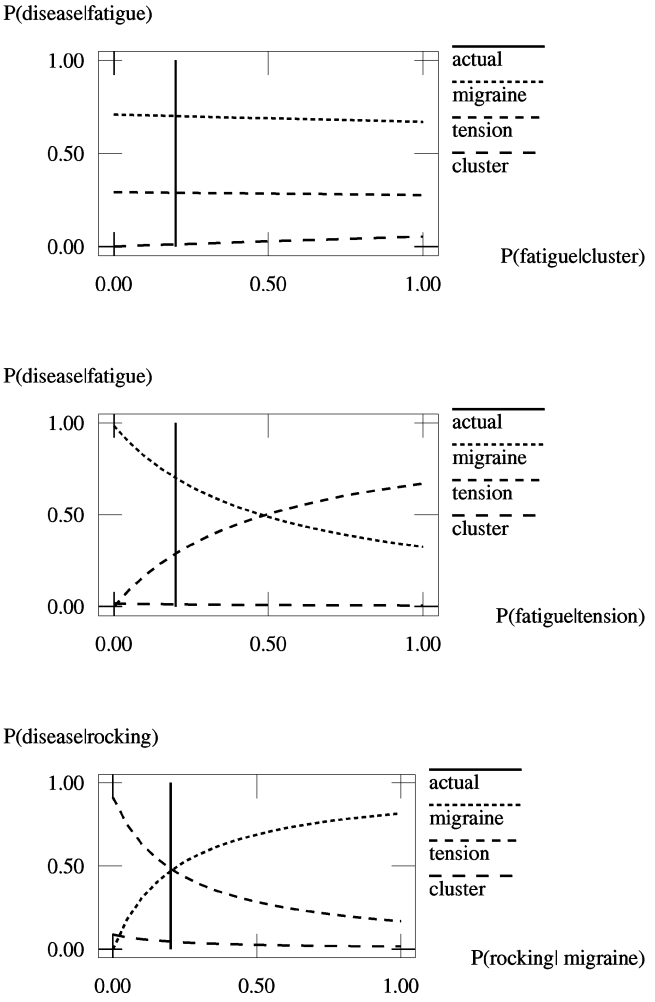


Fig. 2. Three examples of the graphs used for sensitivity analysis. The top graph plots $P(disease \mid fatigue)$ as a function of $P(disease \mid rocking)$. The middle graph plots $P(disease \mid fatigue)$ as a function of $P(fatigue \mid tension)$. The bottom graph plots $P(disease \mid rocking)$ as a function of $P(rocking \mid migraine)$. The vertical line labeled "actual" corresponds to the original assessment.

of the three values of the disease node conditioned on the evidence node.[5]

Examining the graphs generated by plotting the results of this analysis, it is possible to group the posterior probabilities into three different classes, corresponding to three very different degrees of sensitivity. The graph at the top of Fig. 2 shows a typical case in which the diagnostic decision is not sensitive to variations in the assessment; changing the posterior probability $P(fatigue \mid cluster)$ from the original assessment (0.2) cannot change the most likely disease (migraine). Clearly, in this case, no further assessment is required. The graph in the middle of Fig. 2 shows a typical case of low sensitivity. Although changing the posterior probability $P(fatigue \mid tension)$ from the original assessment (0.2) can change the most likely disease from migraine to tension, this change occurs only when

3. Our expert was involved in the development of the INTERNIST-I knowledge base [10]. In INTERNIST-I, the association between a manifestation and a disease is represented as a variable called *evoking strength*. The evoking strength answers the question: "Given a patient with this `finding`, how strongly should I consider this `diagnosis` to be its explanation?" The evoking strength is expressed as a number on a scale from 0 to 5 [10].

4. The task of sensitivity analysis in the context of belief network construction is illustrated in [4], [11], among others.

5. We are adopting a categorical utility model, which assigns utility 1 to the correct diagnosis, and utility 0 to any misdiagnosis. Therefore, the change in probability of a given diagnosis translates into a corresponding change in utility of the diagnosis, which is the measure sensitivity analysis should focus on.

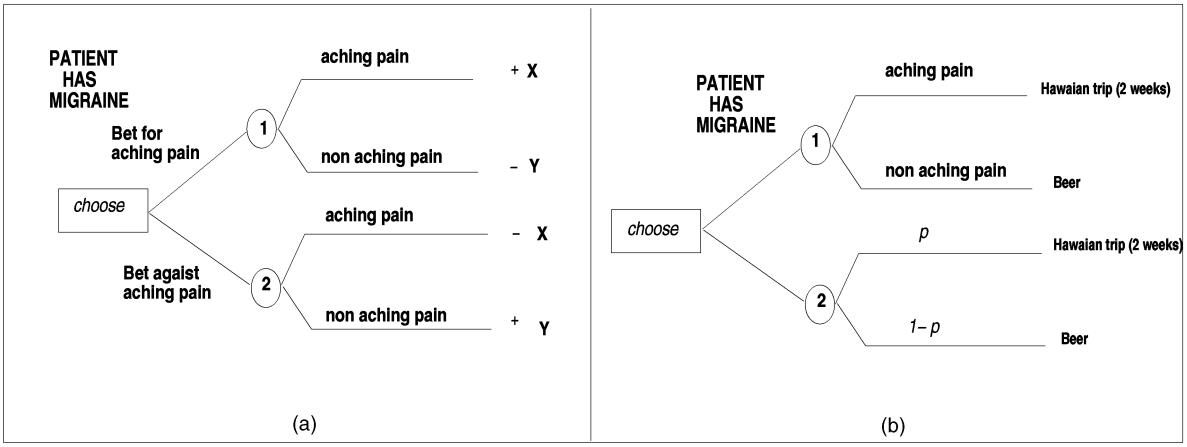


Fig. 3. Decision tree representation for assessing subjective probablities via (a) the betting method; and (b) the probability equivalent method.

$P(fatigue \mid tension)$ is greater than 0.5, that is, 0.3 off the original assessment. In this case, further assessment is required only when the original assessments are expected to be highly inaccurate. Finally, the graph at the bottom of Fig. 2 shows a typical case of high sensitivity; a tiny variation of the original assessment would change the diagnosis from cluster to migraine. In this case, further assessment is necessary.

We found that the conditional probabilities of 32 nodes in our belief network belong to the group of highly sensitive assessments, for which further assessment is required. Twelve of these 32 nodes correspond to binary nodes and the remaining 20 to multivalued nodes. The preliminary refinement effort discussed in this paper focuses on the binary nodes only. Furthermore, since all 12 binary nodes depend on the disease node only, and the change in diagnosis in the sensitivity analysis is always from migraine to tension or vice versa, a total of 24 assessments must be performed ($P(node_i \mid migraine)$ and $P(node_i \mid tension)$).

For the refinement of the "sensitive" probabilities, we assessed the expert's subjective probabilities by means of well established elicitation techniques based on the use of bets and lotteries. Moreover, we adapted the Analytic Hierarchy Process (AHP) method for the assessment of preferences introduced by Saaty in [12], [13] to the task of probability elicitation. A notable feature of the AHP is that, besides facilitating the expert's assessments, it also provides us with a technique to evaluate the consistency of the assessor. A detailed description of these techniques follows.

3.2 Elicitation: Standard Techniques

The methods we used are exemplified in Fig. 3.[6] In the first of these methods, also called *betting method*, the expert is asked about the bets he would be willing to place for or against the occurrence of a certain event. A decision tree representation of this assessment is depicted in Fig. 3a. The elicitation strategy consists of adjusting the amounts of money X and Y until the expert is indifferent about betting for or against the occurrence of the event of interest. In our case, the events of interest are the manifestation of specific aggravating, relieving, or associated factors, assuming that the patient has a certain disease. In Fig. 3a, the event is "the quality of pain is *aching*" (for brevity, *aching*), assuming that the disease is *migraine*. Once we know X an Y, we can compute the expert's subjective probability for the event of interest (in our example, $P(aching \mid, migraine))$ by computing the simple expression $Y/(X+Y)$.

In the second method, often referred to as *equivalent-lottery method*, the expert is asked to compare two lottery-like games as depicted in Fig. 3b. Each of the lotteries can result in either a highly desirable prize (e.g., a trip to Hawaii), or a "consolation" prize (e.g., a glass of beer), but the outcome of the first lottery depends on the occurrence of the event of interest, while the outcome of the second lottery depends on the known probability p. In Fig. 3b, the top lottery reads "win a trip to Hawaii if the event *aching* occurs;" while the bottom lottery reads: "win a trip to Hawaii with known probability p." The elicitation strategy consists of adjusting the parameter p of winning the highly desirable prize until the expert is indifferent between the two lotteries. The final p is the expert's assessment of the probability of the event.

3.3 Analytical Hierarchy Process for Probability Elicitation

The Analytical Hierarchy Process (AHP) proposed in [12], [13] has been devised as a mathematical-based technique to analyze complex situations and assist in decision making. Since its introduction, the method has received wide application in a variety of areas [6]. Commercial software packages are also available [5].

One of the notable features of the method is its providing for a well-founded technique of elicitation of preferences over multiattribute alternatives, and for the mathematical tools to assess the consistency of the elicited preferences. The technique is based on the elicitation from the decision maker of relative or *pairwise* judgements of the importance of the different attributes of interest. From these pairwise judgements, a priority ordering of the attributes of interest can be derived, together with a measure of the expert inconsistency. It is important to emphasize that while the mathematical foundations of the method are by no means trivial, the utilization of the method for preference elicita-

6. For a detailed description of these methods the reader can refer to [19, Chapter 4].

TABLE 2
The Agreed Upon Scale for the Pairwise Comparisons

A and B are equally likely	1
- undecided between 1 and 3	2
A is weakly more likely than B	3
- undecided between 3 and 5	4
A is strongly more likely than B	5
- undecided between 5 and 7	6
A is demonstrably or very strongly more likely than B	7
- undecided between 7 and 9	8
A is absolutely more likely than B	9

tion is straightforward, which is one of the reasons that make the method so appealing.

Although the method is devised for the purpose of preference elicitation, we believe it lends itself naturally to the task of probability elicitation. In this section, we first give a brief description of the method and show its relevance to the task of probability elicitation (some of these ideas are introduced in [15]). We then analyze some of the drawbacks that may arise from its use, as well as possible ways to circumvent them. Finally, we give an intuitive justification of the method, following closely [13].

For the description of the method, we adapt the original terminology for utilities and preferences to the probabilistic setting. In particular, we substitute the term *attribute* with the term *stochastic event* (for brevity, *event*), and the term *importance* with the term *likelihood*. With these modifications, the method can be described as follows:

First, a scale of comparison must be established, based on which the expert can compare pairwise the events of interest. The scale proposed by Saaty is given in Table 2. It considers nine values from "equally likely" 1 to "absolutely more likely" 9. This is also the scale we use in our evaluation. However, we see no reasons not to adopt alternative scales should the expert and/or the analyst deem it appropriate.[7]

Once the scale is defined, the expert can compare pairwise the events under consideration according to the given scale. Since every event is compared with every other event, the resulting assessments can be arranged in a square matrix that we refer to as the *likelihood matrix*. Given events e_i and e_j, the matrix entry m_{ij} (i.e., the matrix element in row i and column j) accounts for the comparison between event e_i and event e_j. For example, if the entry is set to three, it means that the event e_i is "weakly more likely" than the event e_j, according to the scale of Table 2. Notice that once we obtain the entry m_{ij}, we can also fill in the entry m_{ji} with the inverse of m_{ij}, i.e., $1/m_{ij}$.

We thus obtain a reciprocal matrix with diagonal elements all set to 1 (since m_{ii} is equal to 1 for all i). Assuming that n is the number of events of interest, a minimum of $n-1$ pairwise comparisons is needed to fill in the matrix. This can be easily shown by noticing that once the assessments m_{ik} and m_{kj} are available, we can derive the comparison between e_i and e_j as $m_{ij} = m_{ik}m_{kj}$ (that is, if e_i is judged to be twice as likely as e_k, and e_k to be twice as likely as e_j, consistency would imply that e_i is four times as likely as e_j). However, by limiting the assessments to $n-1$ comparisons only, we are guaranteed to obtain a perfectly consistent matrix, as the expert does not need to perform any redundant comparison. Since measuring the consistency of the expert's assessments is one of the objectives of the elicitation process based on the AHP technique, it is clear that we need to directly elicit more than the required minimum of $n-1$ comparisons. In our experiments, as suggested in [13], the comparisons necessary to fill in the elements above the diagonal were obtained by direct assessment (therefore, our expert assessed $n(n-1)/2$ comparisons), while we filled in the elements below the diagonal with the reciprocal of the direct assessments. Notice that with this approach we need to elicit $O(n^2)$ comparisons from the expert, for the assessment of n quantities (the n original probabilities). However, the rationale behind this approach is that the elicitation of the pairwise comparisons is easier and better accepted by the expert than the direct elicitation of the probabilities of interest, as also confirmed by the evaluation described in Section 4.

As an example, Table 3 presents the likelihood matrix accounting for the comparison of the likelihood of the different events (pieces of evidence) introduced in Section 2, conditioned on the patient having tension headache, $P(evidence \mid tension)$. The events are arranged in the matrix from the least likely to the most likely, according to the probability assessments elicited with the standard methods previously described. This arrangement helps in giving a qualitative interpretation of the data. In fact, every column can be interpreted as a ranking of the likelihood of the different events according to a different ratio-scale. If the expert showed consistency in assessing his subjective probabilities with the standard methods and with the AHP technique, every column should be a list of nondecreasing quantities from the top to the bottom. As it is shown in Table 3, this rule is often violated.

The priority order for the events in the matrix can be derived by computing the matrix's eigenvector with the largest eigenvalue. The eigenvector provides the priority ordering, and the eigenvalue is a measure of the consistency of the judgments. Let λ_{max} be the eigenvalue corresponding to the priority vector, and let n be the number of events in the matrix. The closer λ_{max} is to n, the more consistent is the result (in the next section, we will see that λ_{max} is always greater than or equal to n). Deviation from consistency can be determined by computing the *consistency index* C.I. of the likelihood matrix, given by $(\lambda_{max} - n)/(n-1)$, and by comparing it with the C.I. of a randomly generated reciprocal matrix, referred to as the *random index* (R.I.). The ratio of the C.I. to the R.I. of a matrix of equivalent dimension is defined as the *consistency ratio* (C.R.), and can be adopted as the measure of the expert inconsistency. According to [12], a C.R. of 0.10 or less is considered acceptable. It is not clear, however, what is the justification for selecting this threshold, and our experience provides evidence that even lower values of the C.R. may still reflect large levels of inconsistency in the expert assessments, a point to which we will return.

7. See [13, p. 72–73] for a brief discussion on the selection of a scale.

TABLE 3
The Matrix of Pairwise Comparisons of the Probablities Conditioned on Tension Headache, and the Corresponding Eigenvector of Priorities

	lye	rhino	nasal	head	light	rock	dark	rest	stand	ache	press	stress
lye down	1.00	0.50	0.50	0.50	0.33	0.50	0.50	0.50	0.20	0.33	0.20	0.14
rhino	2.00	1.00	1.00	0.33	0.33	1.00	0.50	1.00	0.50	0.33	0.20	0.14
nasal cg.	2.00	1.00	1.00	0.33	0.33	1.00	0.50	1.00	0.50	0.33	0.20	0.14
head inj.	2.00	3.00	3.00	1.00	2.00	2.00	2.00	4.00	3.00	2.00	0.33	0.25
light spots	3.00	3.00	3.00	0.50	1.00	3.00	2.00	4.00	2.00	0.50	0.33	0.20
rocking	2.00	1.00	1.00	0.50	0.33	1.00	0.50	1.00	1.00	0.33	0.20	0.20
dark spots	2.00	2.00	2.00	0.50	0.50	2.00	1.00	2.00	2.00	0.50	0.33	0.20
restless	2.00	1.00	1.00	0.25	0.25	1.00	0.50	1.00	0.50	0.25	0.20	0.14
standing	5.00	2.00	2.00	0.33	0.50	1.00	0.50	2.00	1.00	0.33	0.25	0.20
aching	3.00	3.00	3.00	0.50	2.00	3.00	3.00	4.00	3.00	1.00	0.33	0.25
pressure	5.00	5.00	5.00	3.00	3.00	5.00	5.00	5.00	4.00	3.00	1.00	0.50
stress	7.00	7.00	7.00	4.00	5.00	5.00	5.00	7.00	5.00	4.00	2.00	1.00
priority	0.025	0.033	0.033	0.100	0.083	0.038	0.059	0.032	0.053	0.096	0.182	0.267

3.3.1 Intuitive Justification of the AHP Technique

In this section, we briefly illustrate the motivation behind the application of the AHP technique for probability elicitation and consistency measurement. For a more comprehensive treatment of these issues, see [12, chapters 2,3].

As previously explained, the entries of the likelihood matrix M represent pairwise comparisons, so that entry m_{ij}, in the matrix, measures how much more (or less) likely event e_i is than event e_j, where all assessments are according to an agreed upon ratio scale.

If we knew the actual likelihood of all the events, we could straightforwardly fill in the likelihood matrix M so as to have each of the entry correspond to the ratio of the likelihood of two events. That is, each entry m_{ij} in the matrix, would represent the quantity p_i/p_j, where p_i and p_j are the likelihoods of the events e_i and e_j, respectively. In matrix notation,

$$M \equiv \begin{bmatrix} p_1/p_1 & p_1/p_2 & \cdots & p_1/p_n \\ p_2/p_1 & p_2/p_2 & \cdots & p_2/p_n \\ \cdots & \cdots & \cdots & \cdots \\ p_n/p_1 & p_n/p_2 & \cdots & p_n/p_n \end{bmatrix}.$$

Under this interpretation, if we denote with $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)^T$ the column vector of likelihoods of the single events, then the following relationship should hold:

$$\begin{bmatrix} p_1/p_1 & p_1/p_2 & \cdots & p_1/p_n \\ p_2/p_1 & p_2/p_2 & \cdots & p_2/p_n \\ \cdots & \cdots & \cdots & \cdots \\ p_n/p_1 & p_n/p_2 & \cdots & p_n/p_n \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \cdots \\ p_n \end{bmatrix} = n \begin{bmatrix} p_1 \\ p_2 \\ \cdots \\ p_n \end{bmatrix},$$

or, in a more compact notation,

$$\boldsymbol{M}\boldsymbol{p} = n\boldsymbol{p} \qquad \text{i.e.,} \sum_{j=1}^{n} m_{ij} p_j = \sum_{j=1}^{n} \frac{p_i}{p_j} p_j = n p_i, \quad i = 1, \ldots, n, \tag{1}$$

which yields a system of n equations in n unknowns. For this system of equation to have a solution, n must be an eigenvalue of $\boldsymbol{M}$, and $\boldsymbol{p}$ the corresponding eigenvector.

For the relationship of (1) to hold, the matrix M needs to satisfy certain properties. In particular, it needs to be *consistent*, whereby consistency is defined as follows:

Definition 1. [13, p. 48]. *A matrix $\boldsymbol{M} = \{m_{ij}\}$ is consistent if:*

$$m_{ik} m_{kj} = m_{ij} \qquad i, j, k = 1, 2, \ldots, n. \tag{2}$$

For example, if the assessor judges event e_i to be twice as likely as event e_k, and event e_k to be twice as likely as event e_j, then, to be perfectly consistent, she would have to consider event e_i to be four times as likely as event e_j.

It can be shown that for the matrix M to be consistent, it must have the ratio form $\boldsymbol{M} = (p_i/p_j)$ [13, Theorem 2.1, p. 49]. In general, however, we do not have access to the actual likelihood of the single events, and all we have are the pairwise comparisons m_{ij}. These comparisons are not necessarily consistent (with regard to the above example, it may well happen that the assessor will not consider e_i to be four times as likely as event e_j).

The system of (1) allows for imperfect measurements, since it constrains the assessments to be precise only *on average*, that is,

$$\begin{aligned} &p_i = \text{average of}(m_{i1}p_1, m_{i2}p_2, \ldots, m_{in}p_n) \\ &\text{or, more formally, } p_i = \frac{1}{n}\sum_j m_{ij} p_j \quad i = 1, \ldots, n, \end{aligned} \tag{3}$$

and does not enforce the stricter constraints

$$p_i = m_{i1}p_1 = m_{i2}p_2 = \ldots = m_{in}p_n$$

that should be satisfied if we assumed perfect measurements. However, even allowing for this relaxation, the system of (1) has a solution if and only if the likelihood matrix $\boldsymbol{M}$ is consistent, and n is its principal eigenvalue (with $\boldsymbol{p}$ the associated eigenvector) [13, Theorem 2.2, p. 49].

In order to be able to accommodate inconsistent assessments, we need to allow the value of n in (1) to change. If we denote with λ_{max} this value, the final formulation of the problem is:

$$\begin{aligned} p_i &= \frac{1}{\lambda_{\max}} \sum_{j=1}^{n} m_{ij}p_j \quad i = 1, \ldots, n, \\ \text{or, in matrix notation,} \quad & \boldsymbol{Mp} = \lambda_{\max}\boldsymbol{p}. \end{aligned} \tag{4}$$

The solution of the system of (4) is unique, and the problem as formulated is an instance of the eigenvalue problem. $\lambda_{\max}$ is the largest or principal eigenvalue of $\boldsymbol{M}$, and $\boldsymbol{p}$ is the associated eigenvector representing the vector of priorities. It can be shown that $\lambda_{\max}$ is always greater than or equal to n [13, Theorem 2.9, p. 60], and that it is equal to n if and only if $\boldsymbol{M}$ is consistent [13, Theorem 2.10, p. 60].

Notice that, if we obtain the eigenvector $\boldsymbol{p}' = (p'_1, p'_2, \ldots, p'_n)^T$ by solving (4), the matrix whose entries are p'_i/p'_j is a consistent matrix, and it can be interpreted as a (consistent) estimate of the original matrix $\boldsymbol{M}$ of pairwise comparisons. Correspondingly, the eigenvector $\boldsymbol{p}'$ can be interpreted as an approximation of the actual priority vector, and $\lambda_{\max}$ as an estimate of n. Therefore, the closer $\lambda_{\max}$ is to n, the higher the consistency of the assessments.

To summarize, the problem of finding a priority vector $\boldsymbol{p}$ from a matrix of pairwise comparisons $\boldsymbol{M}$ is solved by assuming that the matrix $\boldsymbol{M}$ is a perturbation of a consistent matrix of likelihood ratios (p_i/p_j), where the p_i's are the unknown likelihoods we want to assess. By solving (4), we derive the eigenvector $\boldsymbol{p}'$, and we construct the consistent matrix $\boldsymbol{M}' = (p'_i/p'_j)$, which we then compare to the original matrix $\boldsymbol{M}$.

To measure the level of inconsistency, we can use the difference between the maximum eigenvalue $\lambda_{\max}$ and n. In particular, the quantity $(\lambda_{\max} - n)/(n-1)$, referred to as the consistency index (C.I.), corresponds to the variance of the error incurred in estimating the m_{ij} [13, p. 83]. As such, it can be used as a measure of inconsistency in the assessments. Furthermore, by taking the ratio of $\lambda_{\max} - n$ to its average value over a large number of reciprocal matrices whose entries are randomly (and uniformly) selected from the interval $[1/n, n]$—a measure referred to as *consistency ratio* (C.R.)—we obtain a relative measure of consistency, which compares the consistency of a set of informed assessments to the consistency of a set of random assessments. Clearly, we would expect the former to be much higher than the latter.

We conclude this section by briefly discussing some of the problems that may arise from the use of pairwise comparisons for the purpose of probability (or preference) elicitation.

The first issue we address has to do with the sensitivity of the AHP to the introduction of new alternatives to the pool of alternatives already considered. That is, the inclusion of a new alternative may result in an alteration of the ranking, or priority order, of the "old" alternatives. This phenomenon is usually referred to as *rank reversal* in the literature (see, e.g., [13, Chapter 5]). While we do not deal with the issue of rank reversal and rank preservation in this paper, experimental studies show that there exist situations when the decision maker is susceptible to rank reversal and other situations when she is not. Accordingly, the AHP provides for modes of operation able to handle both situations.

Another delicate issue is the comparison of widely different elements (in the context of probability elicitation, the comparison of extreme probabilities) and the reliability of the corresponding assessments. In general, when comparing widely different elements, people are unable to provide reliable assessments. To deal with this problem, the AHP works by aggregating elements into *homogeneous* clusters, such that elements belonging to the same cluster must be of the same order of magnitude, with a common element shared by two consecutive clusters. The relative measurements within each clusters can be related and the clusters combined because of the presence of the common element between consecutive clusters. Given two consecutive clusters, the largest element in the small cluster is included as the smallest element in the large cluster. To relate the elements of the two clusters, the relative weights of the elements in the second cluster are all divided by the relative weight of the common element, and multiplied by its relative weight in the smaller cluster [13, p. 58–59]. This procedure also helps to explain the choice of the scale of Table 2, which enforces the requirement that the coefficients of the comparison in the likelihood matrix are of the same order of magnitude, that is, between one and nine.

4 Evaluation

We applied the standard elicitation techniques and the AHP-based technique in two sessions of about 90 minutes each. Our expert is very familiar with our work, and has a good understanding of probability theory and of the belief network formalism. Therefore, a lengthy motivational protocol was unnecessary for him to understand the goals of the elicitation. The first session was devoted to the direct elicitation of the probabilities of interest by means of the standard techniques described in Section 3.2. In the second session, we focused on the elicitation of the pairwise comparisons according to the AHP as described in Section 3.3. It is important to emphasize that while the first session involved the elicitation of n quantities, and the second session involved the elicitation of $n(n-1)/2$ quantities, the time needed for both sessions was approximately the same.

4.1 Standard Techniques

In the first session, we applied the standard techniques only. We performed some warm-up assessments, during which the expert preferred the lottery-equivalent method to the betting method. Since he was comfortable with assessing probabilities explicitly, he felt that the betting method would have forced him to compute mentally the amount of money corresponding to a given probability. In

TABLE 4
An Example of a Comparison Table Similiar to the Ones Submitted to the Expert

	Abso-lute		Very Strong		Strong		Weak		Equal		Weak		Strong		Very Strong		Abso-lute	
light spots	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	aching
light spots	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	nasal congestion
light spots	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	rhinorrea

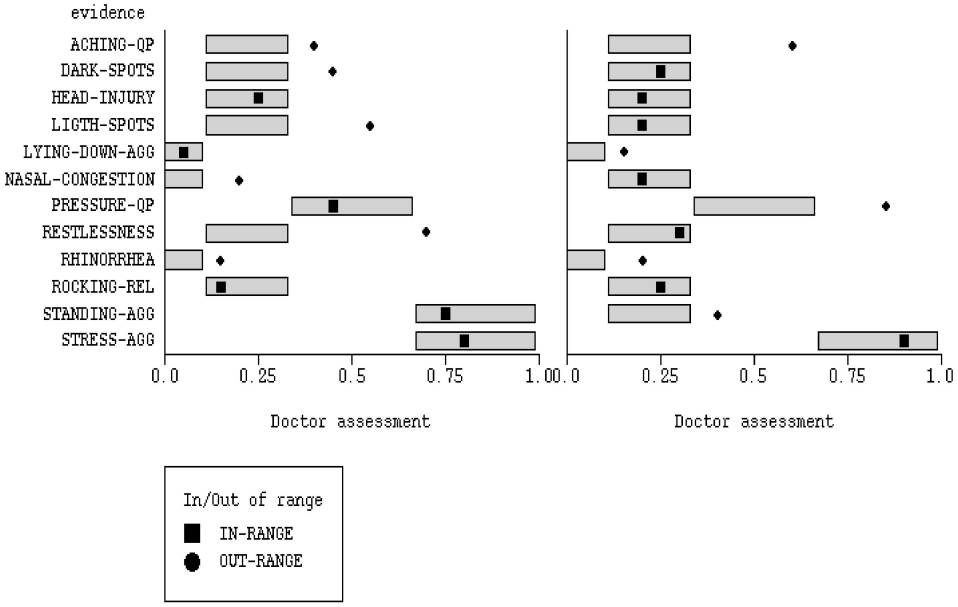


Fig. 4. A comparison between the expert's assessments for the preliminary version of the network (horizontal bars), and the expert's assessments obtained by means of the standard lottery-equivalent method (marks).

other words, the betting method would have simply introduced an unnecessary level of indirectness by increasing the possibility of noise due to the wrong computation of the appropriate monetary quantities. On the other side, the application of the lottery-equivalent method seemed quite effective. Although the method was sometimes unnecessary (the expert would directly give a numeric assessment), it proved a useful tool to force the expert to carefully consider his assessments.

4.2 Pairwise Comparisons (The AHP Technique)

In the second session, we performed the elicitation of the pairwise comparisons. To this purpose, the expert had to fill in two tables whose format is similar to the one depicted in Table 4. We prepared two such tables, one for the probabilities conditioned on migraine, and one for the probabilities conditioned on tension headache. Both tables have 66 entries. The pairwise judgments given by the expert were then plugged into two matrices. The matrix for the probabilities conditioned on tension headache is presented in Table 3. The consistency ratio (C.R.) for this matrix is 0.05, and the C.R. for the matrix of probabilities conditioned on migraine is 0.06. While these results are considered as providing evidence for sufficient consistency according to the 0.10 threshold proposed in [12], direct examination of the matrix' entries suggests that this is not the case (see the matrix of Table 3, where virtually every column manifests inconsistency. See, e.g., the columns indexed by the events "rhino" and "restless," where the entries markedly depart from a nondecreasing order). Further evidence of the expert inconsistency is provided by the comparison of the expert's assessments according to different elicitation techniques, as shown in Fig. 4. These results suggest that a threshold lower than 0.10 might be appropriate, at least for the purpose of probability elicitation.

4.3 Confronting the Expert with His Inconsistencies

The expert was highly inconsistent in his assessments elicited by means of the standard equivalent-lottery method, when compared with his assessments for the preliminary version of the network (the intervals described in Section 2, Table 1). A diagram highlighting these inconsistencies is shown in Fig. 4, where the horizontal bars represent the assessments based on the intervals of Table 1, and the point-value probabilities elicited by means of bets and lotteries are represented by means of filled squares or circles, depending on whether they fall within the corresponding probability interval (squares) or not (circles). The left diagram plots the probabilities conditioned on the disease being migraine. The right diagram plots the probabilities conditioned on the disease being tension headache. We can see that there are ample

inconsistencies between the two classes of assessments. Some of the point-value probabilities assessed by the expert even fall out of the probability intervals.

We were interested in the expert's reactions to his inconsistencies. In particular, we were interested in how confronting the expert with his inconsistencies would change his assessments and/or his perception of the elicitation techniques. To this purpose, we showed the diagram to the expert. When presented with this diagram and the apparent inconsistencies in it, the expert was not able to clearly privilege one class of assessments over the other. For each of the assessments for which the point-value falls out of the corresponding interval, he revised his assessment by picking the middle point between the point-value and the closest extreme of the interval. The expert did not seem to privilege the assessments he gave by means of the method that he was originally familiar and confident with (choice of probability intervals).

We propose some possible explanations of the expert's behavior just described. First, it seems to confirm our expert's opinion. Since the beginning, he claimed that it is very hard to assess sensible numbers for the domain we selected. He also warned that forcing the refinement of the initial intervals he assessed did not necessarily mean we would obtain better assessments.

There are, however, other plausible explanations to the expert's inconsistencies. A possible partial explanation is to be found in the different elicitation methodologies used. For the assessment of the probability intervals, our expert was presented with triples of events to be assessed. Each triple accounted for the same event conditioned on each of the three diseases (e.g., the expert was asked to assess

$$P(aching \mid migraine),$$

$P(aching \mid tension)$, and $P(aching \mid cluster)$). Apparently, the expert was often unable to discriminate to a fine enough grain the causal support of migraine and tension headache to the same event. This is evident by looking at the sets of intervals for the two diseases plotted in Fig. 4, which are almost identical. This similarity is not replicated with the point-value assessments.

Another possible explanation is to be found in the predefined and limited set of intervals the expert could consider in the first of our elicitation efforts. When the actual probability is highly skewed on one of the interval's extremes, it can easily happen that the expert selects the adjacent interval.

5 The Many Roles of AHP for Probability Elicitation

Our experience in probability elicitation has been mainly driven by practical considerations rather than by an attempt to systematically develop and compare different elicitation techniques. However, in retrospect, we believe that the most valuable outcome of our study has been the preliminary development and application of the AHP technique to the probability elicitation task. In fact, in light of this experience, we now recognize several potential roles for the AHP technique in the elicitation task.

Immediate feedback: The AHP technique allows the analyst to make the expert face his inconsistencies as soon as they arise. To this purpose, we propose the use of the AHP technique to help the expert make his priorities explicit. Once agreed on the priority vector (the eigenvector computed from the matrix), the analyst should refer to this vector whenever the assessor gives assessments incompatible with it.

Probability elicitation: The AHP technique can also be used for the actual assessment of probabilities of mutually exclusive and exhaustive events. In fact, since the priority eigenvector is such that its elements are all nonnegative and sum to 1, they can be directly adopted as the probabilities of the corresponding events.

Focus further elicitation: We envision the use of the AHP technique to focus further elicitation. The matrix and the eigenvector generated by the AHP technique can be used for this purpose. As previously explained, each column of the consistency matrix accounts for the comparison of the different events with respect to a fixed element (the event indexing the column). Assuming that the events in the matrix are ordered according to the priorities given by the eigenvector, by plotting each column as a curve we can interpret the deviation of the curve from monotonicity as an indication that further elicitation is necessary for that event.[8]

Early inconsistency detection: The AHP technique allows the analyst to measure the degree of inconsistency in the expert's assessments. Furthermore, elicitation by pairwise comparisons is usually easily assimilated by the expert, and it is relatively economical in terms of time-requirements. Based on these considerations, we envision its use for the purpose of gauging the expected difficulty of the elicitation process. If the C.R. (consistency ratio, see Section 3.3 for its definition) of an expert for the domain of interest is very high, this suggests that the domain is hard to quantify reliably (or, possibly, that the expert is not an "expert" after all, possibility that could be ruled out should the C.R. of other experts result to be as high), and it might be appropriate to consider interviewing other experts, or adopting a possibly more time-consuming but more reliable elicitation technique.

However, to be able to rely on the measure of inconsistency provided by the AHP, it is necessary to better understand the relation between the value of the C.R. and the level of inconsistency deemed acceptable for the knowledge base being built. In this paper, the manifest inconsistency showed by the expert's assessments based on different elicitation techniques provided us with evidence that the 0.10 C.R. was not appropriate.

6 Conclusions and Future Work

In our experience in probability elicitation, we recognize some significant limitations. First, the medical domain of chronic nonorganic headaches is considered particularly difficult to formalize, because no well-established causal

8. In [13], other methods for focusing further elicitation are proposed. However, those methods are concerned with selecting the pairwise comparisons—rather than the probability of single events— for which a refinement is needed.

model exists. Second, we had access to one expert only, therefore we could not make any interexpert comparisons. Finally, our expert was available for four relatively short sessions only.

Even recognizing these limitations, we believe our investigation led to some valuable insights into the probability elicitation task.

On the practical side, the main insight is the realization that confronting the expert with the inconsistencies in his assessments after the expert had already gone through the whole elicitation process proved to be ineffective. As suggested in Section 5, an alternative approach worth investigating is to confront the expert with his inconsistencies as soon as they arise during the elicitation process.

On the methodological side, we devised and applied the new AHP-based technique. This technique appears to be extremely useful in several aspects of the probability elicitation task, although more work is necessary to empirically verify its power in the elicitation of probabilities from different experts in different domains, as well as to better understand the relationship between the value of the consistency ratio (C.R.) introduced in Section 3.3 and the level of inconsistency deemed acceptable for the knowledge base being built.

Some of the general directions of further research worth exploring have been discussed in Section 5. In the context of our project (the development of a system for patient education in the clinical domain of chronic nonorganic headaches), a natural extension of the elicitation effort reported here would be to use the insights from the AHP-based analysis to refine our probabilistic model. A possible course of action would thus include: 1) the identification— based on the AHP-analysis—of the critical assessments to be refined; 2) further elicitation sessions with the expert to refine his assessments based on the results of his pairwise comparisons; and 3) the incorporation of the refined probabilistic model into our system for history-taking, in an effort to improve its performance.

ACKNOWLEDGMENTS

The authors would like to thank Gordon Banks for his collaboration in the construction of the belief network, and the participants to the IJCAI-95 Workshop, "Building Probabilistic Networks: Where Do the Numbers Come From?" for useful discussions on the content of a preliminary version of this paper. This research was done while the first author was at the Intelligent Systems Program, University of Pittsburgh.

REFERENCES

[1] B.G. Buchanan, J. Moore, D. Forsythe, G. Carenini, G. Banks, and S. Ohlsson, "An Intelligent Interactive System for Delivering Individualized Information to Patients," *Artificial Intelligence in Medicine,* vol. 7, no. 2, pp. 117–154, 1995.

[2] G. Carenini, S. Monti, and G. Banks, "An Information-based Bayesian Approach to History-taking," *Proc. Fifth Conf. AI in Medicine, Europe,* pp. 129–138, 1995.

[3] G.F. Cooper and E. Herskovits, "A Bayesian Method for the Induction of Probabilistic Networks from Data," *Machine Learning,* vol. 9, pp. 309–347, 1992.

[4] V. Coupé and L.C. van der Gaag, "Practicable Sensitivity Analysis of Bayesian Belief Networks," *Proc. Joint Session Sixth Prague Symp. Asymptotic Statistics and the 13th Prague Conf. Information Theory, Statistical Decision Functions and Random Processes,* pp. 81–86, 1998.

[5] E.H. Forman and T.L. Saaty, Expert Choice, Inc., http://www.expertchoice.com/.

[6] B.L. Golden, E.A. Wasil, and P.T. Harker, eds. *The Analytic Hierarchy Process: Applications and Studies.* Springer Verlag, 1989.

[7] D. Heckerman, D. Geiger, and D.M. Chickering, "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data," *Machine Learning,* vol. 20, pp. 197–243, 1995.

[8] M.I. Jordan, ed., *Learning in Graphical Models.* NATO Science Series, Kluwer Academic, 1998.

[9] W. Lam and F. Bacchus, "Learning Bayesian Belief Networks—An Approach Based on the MDL Principle," *Computational Intelligence,* vol. 10, no. 4, pp. 269–293, 1994.

[10] R.A. Miller, H. Pople, and J. Meyers, "Internist-1, An Experimental Computer-Based Diagnostic Consultant for General Internal Medicine," *New England J. Medicine,* vol. 307, no. 8 1982. Also in *Readings in Medical Artificial Intelligence: The First Decade,* W.J. Clancey and E.H. Shortliffe, eds. Reading, Massachusetts: Addison-Wesley, pp. 190–209.

[11] M. Pradhan, M. Henrion, G. Provan, B. Del Favero, and K. Huang, "The Sensitivity of Belief Networks to Imprecise Probabilities: An Experimental Investigation," *Artificial Intelligence,* vol. 85, no. 1–2, pp. 363–397, 1996.

[12] T.L. Saaty, *The Analytic Hierarchy Process.* McGraw-Hill, 1980.

[13] T.L. Saaty, *Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process.* Pittsburgh, Pennsylvania: RWS Publications, vol. VI, 1994.

[14] J. Saper, *Handbook of Headache Management.* Williams and Wilkins, 1993.

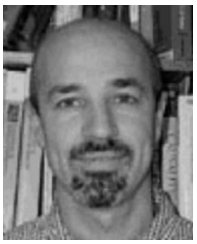
[15] S. Schocken, "Ratio-scale Elicitation of Subjective Degrees of Support," NYU-CRIS Working Paper IS-93-30, 1993.

[16] S. Solomon and S. Fraccaro, *The Headache Book.* Consumers Union of the United States, 1991.

[17] D. Spiegelhalter, A. Dawid, S. Lauritzen, and R. Cowell, "Bayesian Analysis in Expert Systems," *Statistical Science,* vol. 8, no. 3, pp. 219–283, 1993.

[18] D. Spiegelhalter, R.C.G. Franklin, and K. Bull, "Assessment, Criticism, and Improvement of Imprecise Subjective Probabilities for a Medical Expert System," *Uncertainty in Artificial Intelligence,* M. Henrion, R. Shachter, L. Kanal, and J. Lemmer, eds., North Holland, vol. 5, pp. 285–294, 1990.

[19] D. von Winterfeldt and W. Edwards, *Decision Analysis and Behavioral Research.* Cambridge University Press, 1986.



Stefano Monti is currently a post-doctoral fellow at the Robotics Institute, Carnegie Mellon University. He received his PhD degree from the Intelligent Systems Program, University of Pittsburgh. His research interests are in uncertain reasoning, machine learning, and statistics, with an emphasis in learning probabilistic graphical models.

Giuseppe Carenini received a Laurea degree in computer science from the University of Milan (Italy). He joined, as a research associate, the NLP and Communication Group at IRST (Trento, Italy). At IRST, he developed his interests in natural language generation and interactive systems. He worked on the design and development of the Alfresco system, an interactive system about Italian frescoes and monuments. In 1992, he joined the Migraine Project at the computer science department of the University of Pittsburgh, PA, to study how to design interactive systems following sound knowledge acquisition techniques. Since 1993, he has been a graduate student in the Intelligent System Program at the University of Pittsburgh, PA. He is also a member of the Visualization and Intelligent Interfaces Group at Carnegie Mellon University. His current research interest is on knowledge representation and acquisition for the automatic generation of presentations combining text and information graphics.