Dealing with the expert inconsistencies: the sooner the better.

Stefano Monti   Giuseppe Carenini
Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA 15260
{smonti,carenini}@isp.pitt.edu

Abstract

In [Carenini *et al.*, 1995] we present an information based Bayesian strategy for history-taking, aimed at optimizing the evidence-gathering process. We have applied this strategy to a history-taking module developed as part of a system for patient education in the clinical domain of chronic non-organic headaches. The weakest phase in the development of our prototype application was the knowledge acquisition for the construction of the belief network that we used as our knowledge base. In this paper, we describe our additional elicitation effort, aimed at refining the numbers used in the belief network. We apply known probability elicitation techniques, based on the use of bets and lotteries. We also adapt a technique, not specifically devised for the elicitation of probabilities, to compute an analytical measure of the consistency of the assessments. The technique is in our opinion very promising, and lends itself to be applied more extensively to probability elicitation.

1 Introduction

In [Carenini *et al.*, 1995], 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 non-organic headaches [Buchanan *et al.*, 1993].

The weakest phase in the development of our prototype application was the knowledge acquisition. For the construction of the belief network that we used as our knowledge base, the medical literature only provided information for structuring the network. Since quantitative information was not available to automatically estimate the relevant probabilities, we had to resort to a domain expert. Our expert was already familiar and confident in expressing his subjective probabilities choosing from a predetermined set of probability intervals. Therefore, this was the method of choice. At the project onset, we considered such an approximation as acceptable, since our main goal was to inform history-taking strategies, not to automate the diagnostic process.

In this paper, we consider the belief network structure as given, and focus only on the issues of assessment and criticism of the conditional probabilities. In Section 2 we describe the belief network used as knowledge base for the prototype application described in [Carenini *et al.*, 1995], and the knowledge acquisition process for its construction. In Section 3, we describe the sensitivity analysis that focused further knowledge acquisition. For the knowledge acquisition, we apply well known probability elicitation techniques, based on bets and lotteries. We also apply techniques adapted from the analytical hierarchy process for the elicitation of preferences proposed in [Saati, 1980]. This method provides an explicit measure of the consistency of the assessments. In Section 4 we give a preliminary subjective evaluation of the different elicitation methodologies. In particular, we examine the confidence of the expert with the different elicitation techniques, the expert's reaction to the techniques and their results, changes in the expert's perception of the different elicitation techniques, and the expert's reaction to his inconsistencies. In the final section, we draw some conclusions about the elicitation process based on our experience.

2 Background

The belief network used by the history-taking strategy is the one depicted in Figure 1. As knowledge sources for the knowledge acquisition phase, we used both current medical literature [Saper, 1993; Solomon and Fraccaro, 1991], and the neurologist in our group. The medical literature was the main knowledge source for the definition of the belief network structure. Unfortunately, it did not provide any numbers. We considered a mutually exclusive and exhaustive set of three diseases (hypotheses): Migraine, Cluster, and Tension headaches[1]. The assumption that these diseases are mutually exclusive is a reasonable approximation in our case, since the only significant overlap is a 10% between migraine and tension. The assumption of exhaustiveness is also reasonable because Migraine, Cluster, and Tension headaches cover about 90% of all non-organic headaches.

[1]The assumptions of mutually exclusive and exhaustive set of diseases allow us to group the diseases in a single node, which is the one in the center of Figure 1.

Figure 1: The belief network for the clinical domain of chronic non-organic 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. 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 assessing the conditional probability tables for the network, we did not use any sophisticated knowledge acquisition techniques. At that stage of the project our main concern was the history-taking strategy, rather than the accuracy of the assessments, and our expert was already familiar and confident in expressing his subjective probabilities choosing from a predetermined set of six adjacent probability intervals. We translated these intervals into point values by taking the midpoints of the given intervals, as suggested in [Spiegelhalter *et al.*, 1990]. A small subset of these subjective assessments are shown in Table 1. We considered all these approximations as acceptable, since our main goal was to inform history-taking strategies, not to automate the diagnostic process.

	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	$0.00 - 0.10$	$0.00 - 0.10$	$0.67 - 0.99$
...			

Table 1: Probability assessments for the old network.

After a preliminary evaluation of the strategy using 5 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 Sensitivity Analysis and Probability Elicitation

In order to focus our attempt to increase the diagnostic accuracy of the belief network, we performed a simple one-way sensitivity analysis. We varied the posterior probability of each evidence node in the network from 0 to 1 (maintaining all the other probabilities constant) and we computed the corresponding variation in the probabilities of the three values of the disease node conditioned on the evidence node[2]. 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

[2] 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.

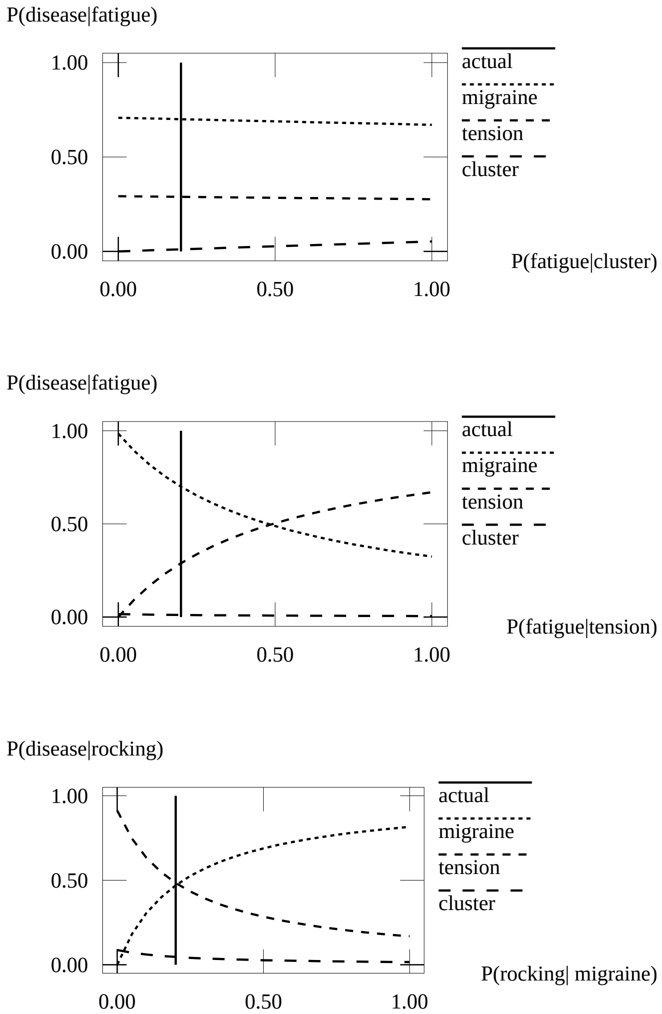

Figure 2: Three example of the graphs used for sensitivity analysis. The top graph plots *P(disease|fatigue)* as a function of *P(fatigue|cluster)*. The middle graph plots *P(disease|fatigue)* as a function of *P(fatigue|tension)*. The bottom graph plots *P(disease|rocking)* as a function of *P(rocking|migraine)*. *Actual* is the thick line parallel to the y axis, and it corresponds to the original assessment.

degrees of sensitivity. The graph at the top of Figure 2 shows a typical case in which the diagnostic decision is not sensitive to variation in the assessment; changing the posterior probability $P(fatigue|cluster)$ 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 Figure 2 shows a typical case of low sensitivity. Although changing the posterior probability $P(fatigue|tension)$ from the original assessment (0.2) can change the most likely disease from migraine to tension, such a change occurs only when $P(fatigue|tension)$ 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 Figure 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 last group, which require further assessment. Twelve of these 32 nodes correspond to binary nodes and the remaining 20 to multi-value 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

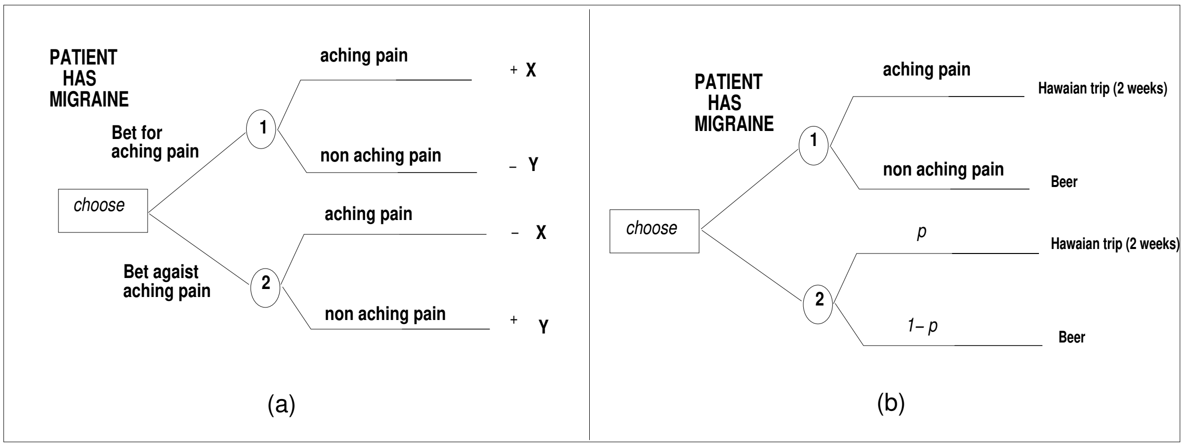

Figure 3: Decision tree representation for assessing subjective probabilities via (a) the betting method; and (b) the probability equivalent method.

Hierarchy Process (AHP) method for the assessment of preferences introduced by Saati in [Saati, 1980], to the task of probability elicitation. Since the technique is based on pairwise comparison, it cannot be used to assess the actual probability to be plugged into the network. However, it is a useful technique for evaluating the consistency of the assessor. A detailed description of these techniques follows.

Standard techniques [3]

The methods we used are exemplified in Figure 3. 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 Figure 3.a. 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 Figure 3.a, 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* | *migraine)*) by solving 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 Figure 3.b. Each of the lottery 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 Figure 3.b, 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.

Analytical Hierarchy Process revisited

The Analytical Hierarchy Process (AHP) technique proposed in [Saati, 1980] was devised for eliciting the decision maker's preferences over multi-attribute alternatives. This technique is based on the elicitation from the decision maker of relative or *pairwise* judgements of the importance of the different attributes of interest. The technique has a well-founded mathematical justification, and provides an analytical measure of the consistency of the decision maker's assessments. We present here our adaptation of the technique to probability elicitation (some of these ideas are introduced in [Schocken, 1993]). Below, we give a short description of the method, by adapting the terminology to the probabilistic setting. To this purpose, 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, we need to define a scale, based on which we can compare pairwise the events of interest. The scale proposed by Saati is given in Table 2. It considers 9 values from "equally likely" (1) to "absolutely more likely"

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

Table 2: The agreed upon scale for the pairwise comparisons.

[3] For a detailed description of these methods the reader can refer to Chapter 4 in[von Winterfeldt and Edwards, 1986].

Figure 4: The matrix of pairwise comparisons of the probabilities conditioned on tension headache, and the corresponding eigenvector of priorities.

(9). We also need to define a square matrix, indexed by the events under examination, which we call *likelihood matrix*. That is, given events e_1 and e_2, the matrix entry m_{12} accounts for the comparison between e_1 and e_2. If for example the entry is set to 3, it means that e_1 is "weakly more likely" than e_2, according to the scale of Table 2. Notice that once we obtain the entry m_{12}, we can also fill in the entry m_{21} with the inverse of m_{12}, i.e., $1/m_{12}$.

Next, we ask our expert to compare the events pairwise with respect to their likelihood, by using the agreed upon scale, and we insert the agreed upon numbers in the square matrix. 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, we need a minimum of $n-1$ pairwise comparisons to fill in the matrix. However, by limiting the assessments to $n-1$ comparisons only, we obtain a perfectly consistent matrix, since we do not ask the expert for any redundant comparisons. For our purposes, we considered it sufficient to fill in the elements above the diagonal by direct assessment (therefore, our expert assessed $n(n-1)/2$ comparisons). Since the matrix is reciprocal, the remaining elements of the matrix can be easily derived.

The matrix in Figure 4 accounts for the likelihood of the different events (pieces of evidence) conditioned on the patient having tension headache, *P(evidence | 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 non-decreasing numbers from the top to the bottom. As you can see in Figure 4, this rule is often violated.

The exact magnitude of the inconsistency can be precisely measured as proposed by Saati, 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. Deviation from consistency can be determined by computing the *consistency index* C.I. given by $(\lambda_{max} - n)/(n-1)$. According to [Saati, 1980], a C.I. of 0.10 or less is considered acceptable. For the computation of the eigenvector and the eigenvalue, we use the third of the approximate method described in [Saati, 1980], pag. 19.

Intuitive justification of the AHP technique [4]

We can interpret the likelihood matrix as accounting for the likelihood ratio of different pairs of events. That is, the entry m_{ij} in the matrix, represents the quantity p_i/p_j, where p_i and p_j are the likelihoods of the events e_i and e_j respectively. The problem to solve is, given the matrix of m_{ij}, to determine the value of the p_i.

The above interpretation of the matrix entries assumes perfect (and perfectly consistent) measurements of the p_i, which would make the problem of finding the p_i (when m_{ij} are given) unsolvable, unless the following equivalence holds:

$$p_i = m_{i1}p_1 = m_{i2}p_2 = \ldots = m_{in}p_n \quad i = 1, \ldots, n$$

A more realistic relation, which takes into account the possible perturbation in the assessments, would be the following:

$$p_i = \text{average of } (m_{i1}p_1, m_{i2}p_2, \ldots, m_{in}p_n) \;\; i = 1, \ldots, n$$

which in mathematical terms, takes the form:

$$p_i = \frac{1}{n}\sum_{j=1}^{n} m_{ij}p_j \quad i = 1, \ldots, n \tag{1}$$

We thus have n equations in n unknowns. However, this formulation of the problem has a solution only if the m_{ij} are consistent. That is, the problem as it is

[4] Both intuitive and mathematically sound justifications to the technique can be found in [Saati, 1980], Chapters 2 and 3.

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

Table 3: An example of a comparison table similar to the ones submitted to the expert.

now stated allows to discount imperfect measurements, but it still assumes consistent measurements. In order to deal with inconsistency, we need to allow the value of n in Equation 1 to change. λ_{max} denotes this value of n. The final formulation of the problem is thus:

$$p_i = \frac{1}{\lambda_{max}} \sum_{j=1}^{n} m_{ij} p_j \quad i = 1, \ldots, n \qquad (2)$$

The solution to Equation 2 is unique, and the problem as formulated in the equation is the well-known eigenvalue problem.

4 Evaluation

The actual elicitation was carried out 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 Bayesian network formalism. Therefore, a lengthy motivational protocol was unnecessary for him to understand the goals of the elicitation.

Standard techniques

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

Pairwise comparisons

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 3. We prepared two of 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 Figure 4. The C.I. for that matrix is 0.052. The matrix for the probabilities conditioned on migraine proved even more inconsistent, with a C.I. of 0.062. This results however, were not made available to the expert during the session.

Confronting the expert with his inconsistencies

The expert was not only inconsistent in the pairwise comparisons just described. He was also 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 Figure 5. 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

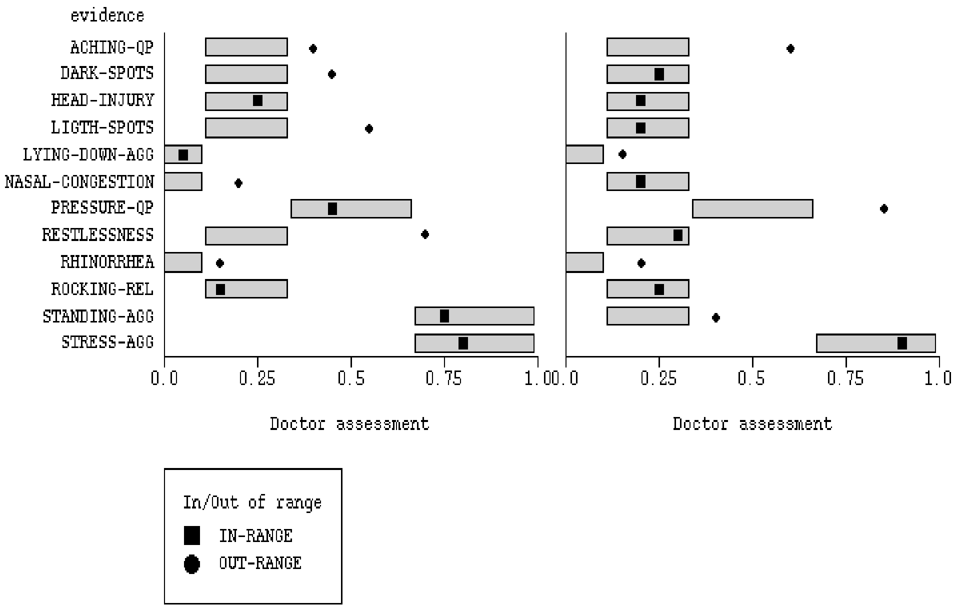

Figure 5: A comparison between the expert's assessments for the preliminary version of the network (horizontal bars), and the expert's assessments obtained by means of the standard lottery-equivalent method (marks).

asked to assess *P(aching | migraine)*, *P(aching | tension)*, and *P(aching | 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 Figure 5, 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 extreme, it can easily happen that the expert selects the adjacent interval.

5 Conclusions and Further Work

In our experience in probability elicitation we recognize some significant limitations. First, the medical domain of chronic non-organic headaches is considered a difficult one, because no well established causal model exists. Second, we had access to one expert only, therefore we could not make any inter-expert comparisons. Finally, our expert was available for four relatively short sessions only.

Even recognizing these limitations, we believe we can draw some interesting conclusions. If the expert is going to provide highly inconsistent assessments, it is probably not worth going through an extensive and possibly expensive elicitation process. The sooner a high degree of inconsistency in the expert's assessments is recognized, the sooner the elicitation process can be abandoned. Our suggestion is to use the AHP technique as a method to evaluate the appropriateness of a domain to probability elicitation. The strategy is to assess the pairwise comparisons with different experts. If the CI of the experts is too high, give up further elicitation. Otherwise, proceed with other techniques.

One idea worth exploring is whether it can be appropriate 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.

Another idea is to use the AHP technique for the actual assessment of probabilities of mutual exclusive and exhaustive events. In fact, the eigenvector of the consistency matrix is such that its elements sum to 1. Therefore, these priority values could plausibly be used as probabilities.

Finally, we envision the use of the AHP technique for sensitivity analysis. 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 a measure of the sensitivity of the event indexing the column.

Acknowledgments

We would like to thank Gordon Banks for his collaboration in the construction of the belief network. We also thank Marek Druzdzel and the anonymous reviewers for their useful comments on the contents of this paper.

References

[Buchanan *et al.*, 1993] B. G. Buchanan, J. Moore, D. Forsythe, G. Carenini, G. Banks, and S. Ohlsson. Using medical informatics for explanation in the clinical setting. Technical Report CS-93-16, University of Pittsburgh, 1993. (To appear in the AI in Medicine journal).

[Carenini *et al.*, 1995] G. Carenini, S. Monti, and G. Banks. An Information-based Bayesian approach to history-taking. In *5th European Conference of AI in Medicine*, Pavia, Italia, 1995. To appear.

[Saati, 1980] T.L. Saati. *The Analytic Hierarchy Process.* McGraw-Hill, Inc., 1980.

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

[Schocken, 1993] S. Schocken. Ratio-scale elicitation of subjective degrees of support. NYU-CRIS Working Paper IS-93-30, 1993.

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

[Spiegelhalter *et al.*, 1990] D.J. Spiegelhalter, R. C. G. Franklin, and K. Bull. Assessment, criticism, and improvement of imprecise subjective probabilities for a medical expert system. In M. Henrion, R. Shacter, L.N. Kanal, and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, volume 5, pages 285–294. North Holland, 1990.

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