TOWARDS A BETTER UNDERSTANDING OF THE INTERNIST-1 KNOWLEDGE BASE*

David Heckerman and Randolph A. Miller

Medical Computer Science Group, Dept. of Medicine
Stanford University School of Medicine, Stanford, CA 94305

Dept. of Medicine
University of Pittsburgh School of Medicine, Pittsburgh PA 15261

We consider two probabilistic interpretations for quantities found in the INTERNIST-1 knowledge base called *evoking strengths*. The first interpretation is based on the definition of evoking strengths while the second is based on the use of evoking strengths in the scoring scheme. We then describe an experiment which can test whether either interpretation is valid. Results of the experiment are presented and discussed. The experiment described in the paper serves as a prototype for other experiments to better understand the INTERNIST-1 knowledge base.

1. INTRODUCTION

INTERNIST-1 is an expert system which has the formidable goal of diagnosing diseases across all of internal medicine [1]. The project was initiated at the University of Pittsburgh by Harry E. Pople, Jr., Ph.D. and Jack D. Myers, M.D. almost 15 years ago; one of the authors (R.A.M.) has been a member of the project for the past 13 years. The INTERNIST-1 project is the core for a continuing research program at the University of Pittsburgh aimed at capturing, representing, and utilizing diagnostic information for computer-based tools in general internal medicine.

One of the impressive features of INTERNIST-1 is its knowledge base which contains a substantial amount of medical information. It is estimated that the current knowledge base encompasses 80% of all of the important diseases of internal medicine. Relationships between over 555 diseases diseases and approximately 4,000 manifestations are represented. Most of the important diseases of internal medicine not currently in the knowledge base are scheduled to be encoded over the next few years. The knowledge base will be maintained in the future as medical knowledge evolves.

A great deal of effort has gone into trying to make this large knowledge base consistent. For example, the encoding of each disease profile has been supervised by a single individual, Dr. Jack Myers, with the assistance of the one of us (R.A.M.) and many other volunteers working on the project. Using the INTERNIST-1 diagnostic consultant program, an estimated 2000-5000 test cases have been analyzed over the past decade, and whenever case analyses resulted in improper diagnostic conclusions, an attempt was made to identify and correct any defects in the knowledge base responsible for erroneous system behavior. For example, in this manner (case analysis) it was discovered that the evoking strengths assigned during the first two years of the project were systematically too high. A retroactive correction was made and subsequently the evoking strengths were adjusted as the knowledge base grew to include more diseases.

Given the general and consistent nature of the INTERNIST-1 knowledge base, it is not surprising that other researchers are interested in adapting it, or portions of it, for their own use. For example, Goldberg and Weiss translated a substantial part of the knowledge base into the EXPERT framework [2] and Nguyen experimented with a subset of the knowledge base adapted to the HELP framework [3]. Indeed, we feel that the INTERNIST-1 knowledge base can become an important resource for medical computer science research.

However, one feature of the current knowledge base may seriously limit its usefulness to other researchers. The definitions of quantities used to represent the strength of association between diseases and patient findings are loosely defined and open to different interpretations. This makes maintenance of the knowledge base difficult for anyone other that the original INTERNIST-1 team. Also, the lack of precise definitions makes it difficult to use the knowledge base with formal inference systems.

In this paper, we focus on one of these quantities intended to represent a strength of association, the *evoking strength*. We will present two possible interpretations for this quantity. These interpretations are precise in that both relate the evoking strength to transformations of subjective probabilities. We then describe an experiment which can determine whether either interpretation accurately reflects the evoking strengths found in the knowledge base.

The results presented are preliminary. In this paper, we wish to emphasize the techniques in such experiments. It is our hope that this discussion will stimulate additional efforts to better understand the INTERNIST-1 knowledge base. We believe such efforts can substantially enhance its use as a tool for medical computer science research.

2. OVERVIEW OF THE KNOWLEDGE BASE

The building block for the INTERNIST-1 knowledge base is the individual disease. For each possible disease

*This work was supported in part by the Josiah Macy, Jr. Foundation, the Henry J. Kaiser Family Foundation, and the Ford Aerospace Corporation. INTERNIST-1 work has been funded through NIH grants RR-01101 and LM-03710. INTERNIST-1 uses the SUMEX-AIM resource under NIH grant RR-00785. Dr. Miller is a recipient of an RCDA grant from the NLM.

entered into the system, a *disease profile* is constructed. A portion of the disease profile for *acute myocardial infarction* is shown in Figure 1. Each profile contains a list of manifestations that can occur with the disease and quantities which represent, in some sense, the degree of association between manifestation and disease. In particular, two clinical variables are associated with each manifestation: an *evoking strength* and a *frequency* (see Figure 1). The evoking strength is intended to be the answer to the question "Given a patient with this finding, how strongly should I consider this diagnosis to be its explanation (or associated with the finding, in the case of predisposing factors)?" The frequency is intended to represent how often patients with the disease have the finding. In this paper, we focus on attempts to better understand the first quantity.

```
Myocardial infarction (acute):

EKG ischemia serial change <s>  4 4
Chest pain unrelieved by nitroglycerin  3 5
EKG ST segment elevation  3 4
Chest pain substernal at rest  2 4
EKG atrial fibrillation  2 2
LDH blood increased  1 4
Onset abrupt  0 4
Cigarette smoking HX  0 2
```

Figure 1: Part of the disease profile for *acute MI*

The evoking strength and frequency for each manifestation are listed after the name of the manifestation.

Evoking strengths range from 0 to 5. Table 1 contains the definition for each possible value of evoking strength that is used by the builders of the INTERNIST-1 knowledge base. We will use the term ES(D,M) to denote the evoking strength for a particular disease D and manifestation M.

0:	Nonspecific - manifestation occurs too commonly to be used to construct a differential diagnosis of individual diagnoses.
1:	Diagnosis is a rarely associated with listed manifestation.
2:	Diagnosis causes a substantial minority of instances of listed manifestation.
3:	Diagnosis is the most common but not the overwhelming condition associated with the listed manifestation.
4:	Diagnosis is the overwhelming consideration in the presence of listed manifestation.
5:	Listed manifestation is pathognomonic for the diagnosis.

Table 1: Definition of evoking strengths

One of the tasks that INTERNIST-1 must perform in order to construct a diagnosis is to *score* all of the diseases in the knowledge base based on the manifestations that have been entered into the system. Evoking strengths play an important role in this scoring process. The score for disease D is determined, in part, by the evoking strengths $ES(D,M_i)$ where M_i ranges over all those manifestations present in the patient that are contained in the profile for disease D. In particular,

$$\text{partial-score}(D) = \Sigma_{M_i} w_{ES}(D,M_i) \qquad (1)$$

The quantity w_{ES} is a *weight* attached to each possible value of evoking strength. These weights, shown in Table 2, were determined empirically. The designers of the system have experimented with several different weighting schemes, but none significantly improved performance over the original scheme shown.

ES	w_{ES}
0	1
1	4
2	10
3	20
4	40
5	80

Table 2: The evoking strength weights

3. TWO PROBABILISTIC INTERPRETATIONS

In this section, we present two interpretations for evoking strengths which can be subjected to formal tests of validity. We begin by stating several assumptions which motivate both interpretations. Firstly, we assume that evoking strengths represents a *degree of uncertainty* in the relationship between disease and manifestation. In particular, we assume that evoking strengths are reflections of the beliefs of Dr. Myers, as he supervised the construction of the entire knowledge base. Secondly, we assume that evoking strengths represent beliefs in a strict sense. That is, we assume utility considerations (i.e., the costs and benefits of treating patients with and without the disease) do not influence the assessment of this quantity. Those familiar with INTERNIST-1 may find this second assumption objectionable. In a later section, we discuss methods for relaxing this assumption.

Given these assumptions, we look for interpretations of the evoking strength based in the theory of subjective probability. Probabilistic interpretations are desirable for theoretical reasons [4] but here we emphasize several practical advantages. One such advantage mentioned earlier is that subjective probabilities have a precise operational definition. Another advantage is that methods for using "hard" data to modify subjective probabilities could be adapted to modify evoking strengths whenever such data become available.

One probabilistic interpretation for evoking strengths can be derived by simply inspecting the definition given in Table 1. Such an inspection suggests that the evoking strength ES(D,M) is directly related to the belief that disease D is present given that manifestation M is present. More formally, the definition suggests that there is some continuous monotonic function f' such that

$$ES(D,M) = f'[p(D|M)] \qquad (2)$$

where p(D|M) is the subjective probability (i.e., belief) that the disease is present given that the manifestation is present. We restrict f' to be monotonic and continuous in order to capture the notion that the evoking strength is directly related to p(D|M). With respect to this and subsequent interpretations, we assume that the relationship between evoking strengths and probabilistic quantities contain noise but this will not be represented explicitly.

Later, we will see that it is convenient to reformulate (2) in terms of *odds*. The odds of an event x, denoted O(x), is a continuous monotonic function of the probability of x. In particular,

$$O(x) = \frac{p(x)}{1 - p(x)} \quad (3)$$

Because the composition of two continuous monotonic functions is another continuous monotonic function, the first interpretation, (2), is equivalent to saying that there is some continuous monotonic function f such that

$$ES(D,M) = f[O(D|M)] \quad (4)$$

where O(D|M) is the odds that disease D is present given that manifestation M is present.

We now consider a second interpretation based on the manner in which evoking strengths are used in the INTERNIST-1 scoring scheme. First, we must introduce a probabilistic quantity known as the *likelihood ratio* and discuss an important property it possesses. Consider Bayes' theorem for the presence and absence of disease D given manifestation M:

$$p(D|M) = p(M|D)p(D) \,/\, p(M)$$

$$p(\sim D|M) = p(M|\sim D)p(\sim D) \,/\, p(M)$$

Dividing the first equation by the second gives

$$\frac{p(D|M)}{p(\sim D|M)} = \frac{p(M|D)}{p(M|\sim D)} \frac{p(D)}{p(\sim D)} \quad (5)$$

Using (3), we see that the term on the far right is the odds that disease D is present before knowing manifestation M is present. This term is often called the *prior* odds and written O(D). Similarly, the term on the left hand side of the equation is the odds that D is present knowing M is present, O(D|M), the *posterior* odds. The term in the middle is often called the *likelihood ratio* and denoted λ(D,M). Therefore, we can rewrite (5) as

$$O(D|M) = \lambda(D,M)\ O(D). \quad (6)$$

Equation (6) is the odds-likelihood form of Bayes' theorem. Notice that the likelihood ratio describes how the prior odds of a disease changes when a manifestation becomes known. That is, the likelihood ratio is a *belief update* as opposed to an *absolute belief*.[1]

When multiple pieces of evidence, M_1, M_2, ... M_n, are conditionally independent given D and its negation, it is not difficult to show that

$$\log\lambda(D, M_1 \ldots M_n) = \Sigma_i \log\lambda(D,M_i) \quad (7)$$

That is, the log-likelihood ratio (or belief update) for the combined evidence is just the sum of the log-likelihood ratios (or belief updates) for each piece of evidence. This simple property of the log-likelihood ratio has been discovered independently many times [5].

We now return to the analysis of the evoking strength. The similarity between the scoring scheme, (1), and the additive property of the log-likelihood ratio, (7), suggests a second interpretation. Formally, this similarity suggests that there is some continuous monotonic function g' such that

$$ES(D,M) = g'[\log\lambda(D,M)]$$

or equivalently, there is some continuous monotonic function g such that

$$ES(D,M) = g[\lambda(D,M)]. \quad (8)$$

In summary, the definition of evoking strength and the use of evoking strengths within INTERNIST-1 suggest two different probabilistic interpretations. The definition suggests that the evoking strength ES(D,M) is related to an absolute belief, O(D|M). The use of evoking strengths in the INTERNIST-1 scoring scheme suggests that ES(D,M) is related to a belief update, λ(D,M).

4. EXPERIMENTAL METHOD

An experiment was performed to determine whether either interpretation is valid, that is, consistent with the evoking strengths found in the INTERNIST-1 knowledge base. The basic idea of the experiment is to assess the quantities O(D|M) and λ(D,M) for many disease-manifestation combinations and compare them with the corresponding evoking strengths in the knowledge base.

We began the experiment by selecting four diseases from INTERNIST-1's knowledge base. The only selection criterion used was that two of the diseases should be common and two should be rare. The reason for this will become apparent shortly. Otherwise, the diseases were selected at random. For each disease, we then randomly selected six to eight manifestations such that a full range of evoking strengths was represented. One of the diseases selected is shown in Figure 1 along with the manifestations selected for that disease.

For each of the selected disease-manifestation pairs, one of us (R.A.M.) assessed p(M|D), the probability that manifestation M is present given that disease D is present, and p(M|~D), the probability that manifestation M is present given that disease D is absent. In assessing these probabilities, it was assumed that other diseases could be present in the patient. In addition to these two quantities, the prior probability of each disease, p(D), was also assessed. The two measures of interest can be computed from these assessments:

$$p(D|M) = \frac{p(M|D)\ p(D)}{p(M|D)p(D) + p(M|\sim D)(1-p(D))}$$

$$\lambda(D,M) = \frac{p(M|D)}{p(M|\sim D)}$$

The probabilities p(M|D), p(M|~D), and p(D) were assessed in lieu of p(D|M) and λ(D,M) because the second author was more comfortable providing the former quantities.

Once these quantities were obtained, plots of λ(D,M) vs. ES(D,M) and O(D|M) vs. ES(D,M) were constructed for each disease.[2] These plots were then used to determine which, if either, interpretation is valid. To see how this is done, once again consider the odds-likelihood form of Bayes' theorem

$$O(D|M) = \lambda(D,M)\ O(D).$$

Suppose that the second interpretation, (8), is valid. In this case, O(D|M) should increase with O(D) as ES(D,M) is held constant. Thus, a plot of O(D|M) vs. ES(D,M) should resemble the upper graph of Figure 2. Each curve in the graph corresponds to the relationship between O(D|M) and ES(D,M) for a particular disease. Curves for diseases with higher prior probabilities should lie above those for diseases with lower priors. In contrast, a plot of λ(D,M) vs. ES(D,M) should show no spread for the different diseases as shown in the lower graph of Figure 2.

But now suppose that the first interpretation, (4), is

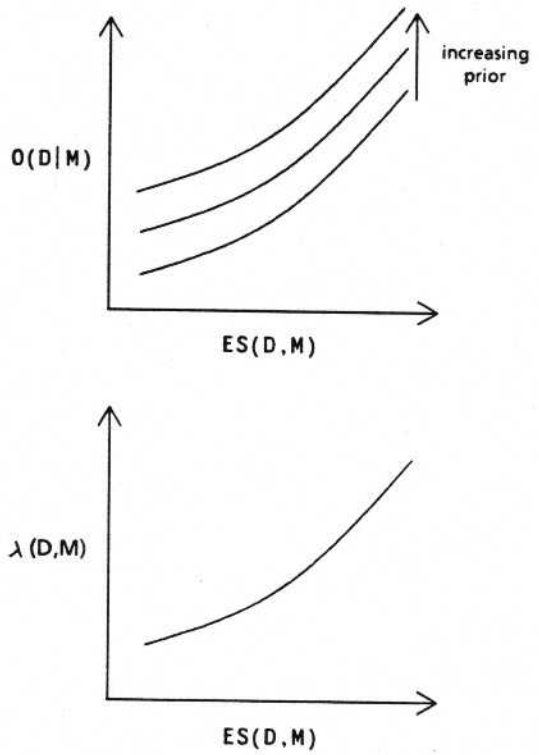

Figure 2: Plots expected if ES(D,M) = g[λ(D,M)]

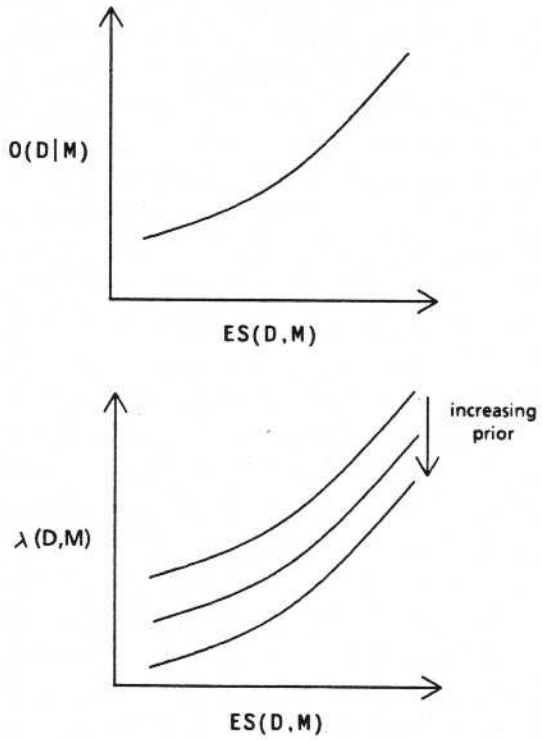

Figure 3: Plots expected if ES(D,M) = f[O(D|M)]

valid. In this case, we expect the graphs shown in Figure 3. A spread in disease curves occurs in the plot of λ(D,M) vs. ES(D,M) and no spread occurs in the plot of O(D|M) vs. ES(D,M). Therefore, we see that these two plots can be used to determine whether either interpretation is valid.

5. RESULTS

Figure 4 shows the results of the experiment just described. In the upper graph of the figure, we see a spread in the disease "curves" which is inconsistent with the first interpretation while in the lower graph of the figure, we see a spread which is inconsistent with the second interpretation. Using a Chi-square test to detect a separation between the curves corresponding to the more prevalent diseases (*influenza* and *acute myocardial infarction*) and the curves corresponding to the less prevalent diseases (*subacute infective left heart endocarditis* and *pheochromocytoma*), we can reject the first hypothesis at a significance level of $p < .01$ and the second hypothesis at a significance level of $p \sim .01$. Thus, our experiment indicates that neither interpretation is valid and, instead, suggests that a third interpretation is possible. In particular, it seems that

$$ES(D,M) = h[\lambda(D,M) + e \cdot O(D)] \qquad (9)$$

where h is yet another continuous monotonic function and where e lies between 0 and 1 and may or may not depend upon the disease and manifestation under consideration.

6. DISCUSSION

We begin with a discussion of one possible explanation of the results. If the first interpretation, (4), were valid, corresponding to e = 1 in (9), the prior probability of the disease would be over-counted in the score for the disease. However, if the second interpretation, (8), were valid, corresponding to e = 0 in (9), the prior probability of the disease would be left uncounted in the score for the disease. Therefore, an e between 0 and 1 seems advantageous given the current scoring scheme. Of course, the best choice for e would depend on the distribution of the number of manifestations entered across cases. We note that if evoking strengths are characterized according to this analysis, a possibly improved scoring scheme could be created wherein the prior probability is counted exactly once in the score for each disease.

The above explanation should be taken as no more than interesting speculation. We again point out that the results of this experiment are preliminary and we wish to emphasize the techniques in such experiments. Therefore, in the remainder of this section, we discuss potential concerns about our experimental method and suggest improvements.

An obvious concern is the amount of data gathered. Each plot contains only 28 data points. This was enough data to reject both initial hypotheses and to suggest (9). However, more data is needed to explore the nature of the functions h and e.

Another concern is about the validity of the probability assessments. Earlier, we assumed that evoking strengths represent beliefs about the association between disease and manifestation. In addition, we assumed that these beliefs are those of Dr. Myers, the individual who supervised the construction of the INTERNIST-1 knowledge base. However, probabilities were assessed by one of us (R.A.M.). Therefore, these assessments can only be valid to the degree that the beliefs within this domain of both individuals are the same. We feel that

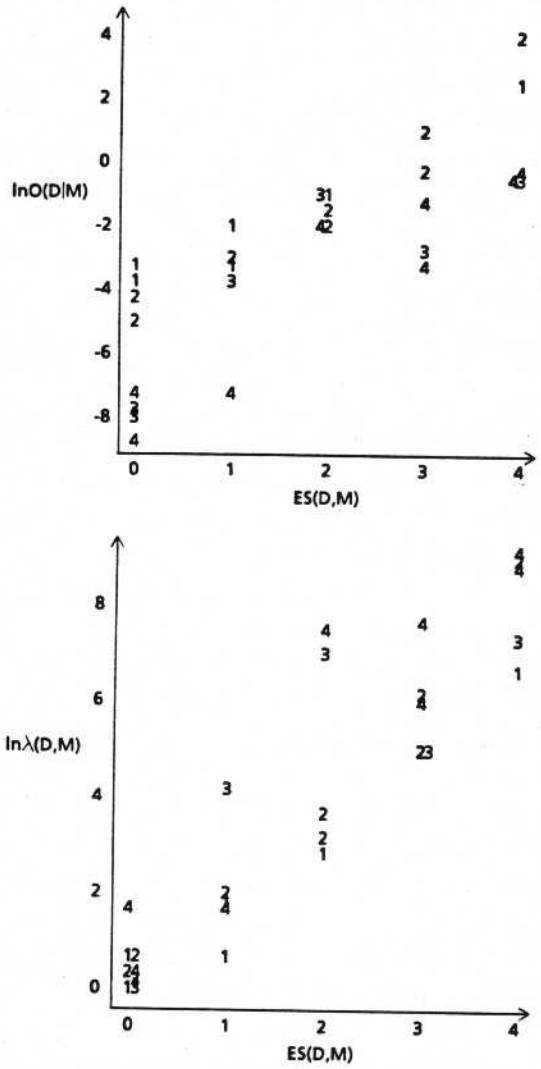

Figure 4: lnO(D|M) and lnλ(D,M) vs. ES(D,M)

Numbers are assigned to each disease in order of decreasing prior probability. Data points corresponding to a particular disease are plotted using the number for that disease. The diseases are (1) *influenza* (prior = 0.02), (2) *acute myocardial infarction* (0.005), (3) *subacute infective left heart endocarditis* (0.0003), and (4) *pheochromocytoma* (0.0001).

the two sets of beliefs are similar to a large extent because the second author has assisted in the construction of the knowledge base for almost the entire duration of the project. Nevertheless, it seems prudent to repeat the experiment with Dr. Myers.

Another potential concern is that knowledge of the purpose of the experiment may have produced biased assessments. However, this could not have occurred because the second author was not aware of the hypotheses being testing when assessing the probabilities. Finally, there is a concern that probability assessments, in general, are subject to systematic biases [6]. In response to this concern, we mention that there are methods for avoiding such biases [7] and several were used in the current experiment. However, we do not claim that all bias was removed and we believe that future experiments of this nature would benefit from greater efforts to avoid bias.

Earlier, we assumed that utility considerations are not factored into assessments of evoking strengths. This assumption may be too strong. To avoid making such an assumption, the experiment can be modified such that diseases with similar utilities (as assessed by members of the INTERNIST-1 team) are considered in groups. Furthermore, by examining differences across groups, it may be possible to quantitate dependencies of the evoking strength on utilities. If this is done, considerations of belief and utility could be separated and perhaps used by a normative reasoning scheme.

We hope the experiment described in this paper will serve as a prototype for future experiments to better understand the INTERNIST-1 knowledge base. A tremendous amount of effort has gone into the creation of the knowledge base. We believe that a small amount of additional effort directed towards a more precise understanding its contents will produce a resource that is invaluable to medical computer science research.

ACKNOWLEDGEMENTS

We would especially like to thank Eric Horvitz for many thought-provoking discussions. We would also like to thank Curt Langlotz, Holly Jimison, Ted Shortliffe, and Larry Fagan for their helpful suggestions.

NOTES

[1]There has been a good deal of confusion between measures of absolute belief and belief update measures in medical computer science literature. For a detailed discussion of this point, see [5].

[2]Actually, logarithms were taken so that the data would be more evenly spread out. This does not change the analysis, however.

REFERENCES

[1] Miller, R.A., Pople, H.E., and Myers, J.D., INTERNIST-1, An Experimental Computer-Based Diagnostic Consultant for General Internal Medicine, *NEJM*, 307 (8), (1982), pp. 468-476.

[2] Goldberg, R.N., Weiss, S.M., An experimental transformation of a large expert knowledge base, *J. of Med. Sys.*, 6 (1), (1982), pp. 41-52.

[3] Nguyen, L.T., "Transferability of a medical knowledge base: a case study between INTERNIST-1 and HELP," PhD dissertation, University of Utah, May 1985.

[4] Cox, R., Probability, Frequency and Reasonable Expectation, *Am. J. of Physics*, 14 (1), January-February (1946), pp. 1-13.

[5] Horvitz, E.J., Heckerman, D.E., Modular belief updates and the inconsistent use of measures of certainty in artificial intelligence research, *In* "Uncertainty in Artificial Intelligence", To be published by North Holland, AAAI/IEEE, 1986.

[6] Tversky, A., and Kahneman, D., Judgement under uncertainty: heuristics and biases, *Science*, 185 (1974), pp. 1124-1131.

[7] Spetzler, C.S., Staeil von Holstein, C.S., Probability encoding in decision analysis, in "The Principles and Applications of Decision Analysis", Strategic Decisions Group, 1984, pp. 603-625.