Verification of Uncertain Knowledge-Based Systems: An Empirical Verification Approach

Daniel E. O'Leary

Management Science, Vol. 42, No. 12. (Dec., 1996), pp. 1663-1675.

Stable URL:
http://links.jstor.org/sici?sici=0025-1909%28199612%2942%3A12%3C1663%3AVOUKSA%3E2.0.CO%3B2-W

Management Science is currently published by INFORMS.

Verification of Uncertain Knowledge-based Systems: An Empirical Verification Approach

Daniel E. O'Leary
University of Southern California, 3660 Trousdale Parkway, Los Angeles, California 90089-1421
oleary@rcf.usc.edu

A number of different tests and approaches are developed to determine the existence of potential anomalies in rule-based systems that employ MYCIN uncertainty factors (weights). First, the distribution of weights is compared to other systems' distributions and weights are investigated as to their individual meanings, to determine whether any weights are unusual. Second, there is increasing evidence that people are not "good" at developing weights on rules, building in symmetries and redundancies that signal "usual" assumptions about the underlying probabilities. Accordingly, weight symmetries generated from rule pairs are analyzed to determine the existence of anomalies. Third, typically rule-based tools have been developed for application in specific domains, such as medicine. Unique aspects of those domains may limit application of the tools to other domains. Finally, ad hoc, rule-based approaches are suboptimal, and alternative formal probability approaches, such as Bayes' nets, more fully specify the probabilistic nature of knowledge.

The paper is part of the empirical verification literature, where verification is done on an actual system and the system provides data that indicates the kinds of anomalies that can be expected. A case study is used to illustrate each of the verification tests and concerns.
(*Knowledge-based Systems; Expert Systems; Uncertainty in AI; Verification and Validation*)

1. Introduction

MYCIN (Buchanan and Shortliffe 1985) was one of the first expert systems that received substantial attention and analysis. MYCIN-like systems form the basis for most backward chaining expert system shells currently used in developing knowledge-based expert systems. MYCIN was originally developed to diagnose bacterial infections and prescribe treatments for them. The development of MYCIN included an "ad hoc" means of representing uncertainty in medical decisions. A "certainty factor" (CF, also referred to here as "weight") was attached to each rule, and a means of combining those CFs across different rules was specified for the medical domain. (MYCIN is discussed in detail below.)

The purpose of this paper is to generate a verification approach for knowledge bases that employ uncertainty representation. In particular, the focus of this paper is on systems developed for analysis of complex management problems, in a MYCIN rule-based environment, using CFs. The paper elicits verification tests and concerns associated with using CFs. A case study is used to illustrate and motivate that approach and those concerns. In so doing, this paper extends and summarizes the literature of verification of MYCIN-like systems, with MYCIN certainty factors on the rules, and provides examples of the verification issues. This is a critical exercise, since it finds a number of problems in the development of MYCIN-like systems for complex management applications.

This paper is a part of the "empirical" verification and validation literature, that is based on analyzing existing systems. Issues of verification only become issues when someone uses the systems. In addition, analysis of existing systems provides validation issues that we

0025-1909/96/4212/1663$01.25

might not anticipate until they occur (e.g., apparent redundancy in the weights, as discussed below). Accordingly, empirical verification research employs data for empirical analysis where that data (and the systems) come from either practice or the research literature.

This Paper

This paper proceeds as follows. Section 2 provides the background, summarizing related work in uncertainty representation and verification, and summarizing the case study in greater detail. Section 3 investigates general characteristics of certainty factors, such as what particular values mean and finds some of the weights make rules effectively "nil." Section 4 analyzes implementations of weights and finds that developers have difficulty in generating weights. Section 5 investigates the impact of the language used in the rules on the weights and finds ambiguous language impacts weight values. Section 6 summarizes some of the factors that relate to the impact of the domain, and finds that complex business problems may not be well-represented using the MYCIN shell. Section 7 proposes an alternative architecture and §8 summarizes and extends the paper.

2. Background

This section provides a brief discussion of representation of uncertainty, verification of knowledge-based systems and the case study.

Representation of Uncertainty in MYCIN

An important problem in artificial intelligence is the so-called issue of "uncertainty in AI." Accordingly, there have been a number of schemes proposed to capture uncertainty in artificial intelligence, including the Bayesian weight structure of MYCIN (Buchanan and Shortliffe 1985), Prospector (e.g., Duda et al. 1979 and 1976) and Bayes' nets (e.g., Pearl 1989). Generally weights or probabilities are associated with individual rules. Inferencing through the rule-base then leads to sets of weights that need to be combined to determine the weight or probability of evidence associated with a particular line of reasoning. Solutions are then ranked based on the cumulated measure of uncertainty.

MYCIN CF's can be elicited along the scale of -1 to 1 (or -10 to 10, -100 to 100, etc.) by establishing probabilities or using weight estimates. Let $P(\cdot)$ represent probability. Let h represent "hypothesis" and e represent "evidence." Probabilistically, MYCIN CFs are defined by Buchanan and Shortliffe for rules of the sort "if e then h," (1985, p. 248) as follows:

$$\mathrm{CF}[h, e] = MB[h, e] - MD[h, e], \quad \text{where,}[1]$$

$$MB[h, e] = 1 \quad \text{if } P(h) = 1, \text{ and}$$

$$= \frac{\max[P(h|e), P(h)] - P(h)}{\max[1, 0] - P(h)} \quad \text{otherwise,}$$

$$MD[h, e] = 1 \quad \text{if } P(h) = 0, \text{ and}$$

$$= \frac{\min[P(h|e), P(h)] - P(h)}{\min[1, 0] - P(h)} \quad \text{otherwise.}$$

CFs are combined as MYCIN inferences through the rule base using the following cumulative combination function (Buchanan and Shortliffe 1985, p. 216):

$$\begin{aligned} &= X + Y(1 - X) \quad X, Y > 0, \\ \mathrm{CF}(X, Y) &= (X + Y)/(1 - \min(|X|, |Y|) \\ &\quad \text{one of } X, Y < 0, \\ &= -\mathrm{CF}(-X, -Y) \quad \text{both } X, Y < 0. \end{aligned}$$

Verification

Although there has been substantial discussion of CFs in the literature, there has been limited investigation of the verification of certainty factors, the primary concern of this paper. Perhaps the lack of literature is due, at least in part, to the fact that some aspects of the verification task are very difficult. For example, the verification of weights generally is regarded as a "formidable task" (e.g., Baligh et al. 1994, p. 184). A recent summary of the general literature on verification was provided in O'Keefe and O'Leary (1993). A comprehensive summary of the tools to facilitate that analysis is summarized in Murrell and Plant (1996). Unfortunately, there has been no systematic treatment for the verification of MYCIN weights for complex management systems, as is presented in this paper.

[1] This was later modified to $\mathrm{CF} = (MB - MD)/(1 - \min(MB, MD))$, see Buchanan and Shortliffe (1985, p. 216) for discussion.

Case Study System: Organizational Consultant

"Organizational Consultant"—"OC" (Baligh et al. 1994, 1996) is a knowledge-based system designed to aid in the design of organizations. The system

> . . . specifies appropriate organizational contingencies or structures and properties for given organizational situations. . . . The facts it needs to know are those of size, ownership, managerial preferences, environment, strategy and technology. . . . The system recommends a structure and properties such as 'a functional structure with high formalization and many rules.'

OC integrates knowledge from a number of heterogeneous sources in the literature. Different empirical and theoretic research studies become the basis of rules in OC. Each of these resulting rules might be referred to as an "atom" of knowledge, a single evidence-hypothesis rule, "if *A* then *B*." OC assigns a MYCIN CF to each rule and then uses MYCIN combination to combine the heterogeneous studies, as the MYCIN inference engine tries to meet its specified goal. OC uses MYCIN weights through an M.1 implementation (Teknowledge, 1986; see also M.4, Cimflex Teknowledge 1991). All of those rules and their certainty factors available at the time of this paper are summarized in Table 1.

Verification of OC has been investigated (e.g., Baligh et al. 1994), using an approach similar to O'Leary (1987, 1988). In particular, Baligh et al. (1994) were concerned with the factors such as content validity, construct validity, criterion validity, and other factors as discussed in O'Leary (1987). In addition, Baligh et al. (1994) noted that six criterion generated in O'Leary (1988) also needed to be addressed:

1. Analyze the knowledge base for accuracy.
2. Analyze the knowledge for completeness.
3. Analyze the knowledge base weights.
4. Test the inference engine.
5. Analyze the condition-decision for decision quality.
6. Analyze the condition-decision matches to determine whether the right answer was received for the right reasons.

As noted by Baligh et al. (1994, p. 184), "O'Leary's third criterion, analyzing the knowledge base weights is a formidable task." As a result, they choose to investigate using insitu test cases on students and executives, rather than directly analyzing the MYCIN CFs. Unfortunately, using such a black box approach may ignore important available evidence that can facilitate the verification process.

Table 1 Disclosed Rules and Weights

1. If size is large then the formalization is high (cf 20).
2. If technology is routine then complexity is low (cf 20).
3. If the strategy is prospector then the centralization is low (cf 20).
4. If the environmental uncertainty is stable the centralization is high (cf 20).
5. If the preference for microinvolvement is high then centralization is high (cf 40).
6. If the environmental complexity is high and the technology is not routine then the horizontal differentiation is high (cf 60).
7. If the strategy is prospector and the technology is routine then this may cause problems (no cf).
8. If the size is large then the decentralization is high (cf 30).
9. If the strategy is prospector then the decentralization is high (cf 20).
10. If the technology is routine then the organizational complexity is low (cf 20).

3. MYCIN Weights and Their "Meaning"

When we open up the black box and examine the weights, perhaps the first concerns are to try to generate meaning from the distribution of the weights and to determine if there are specific weights that deserve particular concern.

Distributions of Weights

One way of setting expectations is to compare the data to another data set. Buchanan and Shortliffe (1986, p. 218) summarize the distribution of MYCIN weights (disclosed in Table 2) with a number of observations about the data. First, they characterize the distributions as "bimodal" (". . . ignoring for a moment those rules, often definitional, that reach conclusions with certainty"). Second, they note that the bimodal peaks occur at 0.8 and 0.2, indicating that ". . . *experts* tend to focus on strong associations (+0.8 . . .) and many weak associations (+0.2 . . .) . . ." (italics added). Third, in the samples, roughly 15% of the observations were negative.

Case Study. There are many substantial, readily apparent differences that differentiate the case from the MYCIN distributions. That is not to say that I expect the

Table 2 Distribution of Weights

Certainty Factor	MYCIN #2*		MYCIN #1*		Organization Consultant	
	Number	Percent	Number	Percent	Number	Present
1	100	11.75%	36	14.88%		0.00%
0.9	44	5.17%	8	3.31%		0.00%
0.8	78	9.17%	34	14.05%		0.00%
0.7	44	5.17%	24	9.92%		0.00%
0.6	33	3.88%	14	5.79%	1	11.11%
0.5	64	7.52%	14	5.79%		0.00%
0.4	82	9.64%	12	4.96%	1	11.11%
0.3	100	11.75%	20	8.26%	1	11.11%
0.2	110	12.93%	22	9.09%	6	66.67%
0.1	64	7.52%	24	9.92%		0.00%
0	0	0.00%	0	0.00%		0.00%
−0.1	4	0.47%	2	0.83%		0.00%
−0.2	6	0.71%	2	0.83%		0.00%
−0.3	12	1.41%	2	0.83%		0.00%
−0.4	18	2.12%	2	0.83%		0.00%
−0.5	20	2.35%	2	0.83%		0.00%
−0.6	8	0.94%	2	0.83%		0.00%
−0.7	12	1.41%	0	0.00%		0.00%
−0.8	0	0.00%	8	3.31%		0.00%
−0.9		0.00%	2	0.83%		0.00%
−1	52	6.11%	12	4.96%		0.00%
	851	100.00%	242	100.00%	9	100.00%

* Estimated based on Figure 10-2, p. 218, Buchanan and Shortliffe (1985).

data to be the same, but only that differences between the distributions can point to anomalies that deserve further analysis. For example,

- The distribution, for OC, based on the disclosed weights is unimodal and there are no negative weights.
- Two-thirds of the OC weights are at 0.20 and there are no strong associations, such as 0.80.

Unimodality and the lack of strong associations indicate that the expertise captured in OC has a different "structure" than in medical expertise, since as noted above, medical expertise is characterized by both strong associations and weak associations. This leads to the question "Is the nature of expertise for medicine so different than complex management problems that the distribution of weights would be so different?" Unfortunately, there is insufficient information disclosed about OC to fully examine this question, but it does raise a question about difference in expertise captured in the OC and MYCIN systems. Further, since two-thirds of the weights take a single value (0.20), it is important to ask the question, "what does it "mean" to have a weight of 0.20?"

Meaning of Individual Weights

MYCIN attributes a "meaning" to different values of individual weights (Buchanan and Shortliffe 1986, p. 91). The larger the weight, the greater the belief in the specific rule. If CF = 1.0 then the hypothesis is "known to be correct." If CF = −1.0 then that means that the hypothesis ". . . has been effectively disproven." "When CF = 0 then there is either no evidence regarding the hypothesis or the supporting evidence is equally balanced by evidence suggesting that the hypothesis is not true."

There are also other values that make a weight important or not important. For example, 0.2 is used as an important point in the interpretation of the CF's (Buchanan and Shortliffe 1986, pp. 94–97). If the CF is less than or equal to 0.2 and greater than or equal to −0.2 then that region of the CF space is categorized as the "not known" region. That is, there is ". . . so little evidence supporting the hypothesis that there is virtually no reasonable hypothesis currently known." As a result, in general if we have a knowledge base of weights in the "not known" region there are some potential problems because it suggests that there is substantial ambiguity surrounding the knowledge.

Case Study. M.1 and M.4 map the CF's into a range of −100 to 100, the approach used by Baligh et al. (1996). In this paper, the original MYCIN mapping is used, so that, a 20 from the range of −100 to 100 becomes a 0.20 for the range of −1 to 1.

In OC Most Weights are at 20—The "Not Known" Region. Six of the weights are 20, one is 30, one is 40, one is 60 and one is not disclosed. Accordingly, two-thirds of the disclosed weights in the OC system in Baligh et al. (1996) are found in that region of "unknown."

In deterministic versions of MYCIN, rules with $|CF| = <0.2$, are treated as having the value "nil" and are disregarded (Buchanan and Shortliffe 1985, p. 97). In a deterministic version of the system two-thirds of the weights would "go away." As a result, there is some concern where the weights on such a large percentage of the rules are so small as to be equivalent to nil. In

particular, as with other ad hoc approaches, it may be that developers are not good at developing CFs.

4. Developing Weights

There is substantial evidence that people are not good at generating "weights" for expert systems using ad hoc approaches like MYCIN. For example, an analysis of the Bayesian weights of "Prospector," in various knowledge-based systems, found numerous implementation errors (O'Leary 1990). As O'Leary (1990) and others have noted, at least a portion of the problem is definition of weights.

How Are CFs Operationally Defined?

Napolean (1990, p. 20) notes that there is no "operational definition of a certainty factor, that is, the definition of a certainty factor does not prescribe a method for determining a certainty factor." As a result, there is no way to know whether different experts mean different things when assigning certainty factors, a particularly difficult problem when those CF's are generated entirely or in part by different knowledge sources, e.g., such as individual empirical research studies. In contrast, probability theory includes an operational definition. Further, as noted by Heckerman (1986), a common misconception is that certainty factors represent measures of absolute belief. Instead, the weights are meant to represent *changes* in belief.

Case Study. There is evidence that Baligh et al. (1994) view the weights as absolute measures of belief. In particular, Baligh et al. (1994, p. 184) note that "the issue is to assign certainty factors to each of these rules to reflect their relative strengths, or importance to the goals of the organization, both separately and collectively."

Developed Symmetry in the Weights and Rules—What Does It Mean?

Because of these problems of not being able to develop weights appropriately or not knowing what they mean, "anomalous" weights could be developed. For example, pairs of rules and weights can be written to exhibit "symmetry." The existence of symmetry allows analysis of the relationship between the two (or more) rules' CF equations. That relationship can be used to determine whether the symmetry has resulted in "sensible" statements about the CFs. One type of symmetry occurs when the certainty factors are the same for two different mutually exclusive and exhaustive hypotheses that derive from the same evidence. In particular, we have the following theorem.

THEOREM 1. *Assume there are two rules "if e then h" and "if e then ~h" with equal certainty factors. Then* $P(h) = P(h|e)(P(\sim h) = P(\sim h|e))$, *that is the hypothesis is independent of the evidence.*

PROOF. If the weights on the two rules are equal then that implies

$$(P(h|e) - P(h))/(1 - P(h)) = (P(\sim h|e) - P(\sim h))/(1 - P(\sim h)).$$

Further, $P(h) = 1 - P(\sim h)$. Substituting and multiplying, yields

$$P(h)*((P(h|e) - P(h)) = (1 - P(h))*P(\sim h|e) + P(h) - 1).$$

Combining terms yields

$$P(h)*P(h|e) - P(h)*P(h) = P(\sim h|e) + P(h) - 1 - P(h)*P(\sim h|e) - P(h)*P(h) + P(h),$$

which yields

$$P(h)*P(h|e) + P(h)*P(\sim h|e) - P(\sim h|e) = 2P(h) - 1.$$

Thus,

$$P(h)*(P(h|e) + P(\sim h|e)) - P(\sim h|e) = 2P(h) - 1.$$

Since the sum of all conditional probabilities dependent on the same variable add to one,

$$P(h) - P(\sim h|e) = 2P(h) - 1.$$

Eliminating common terms and changing sides,

$$1 - P(h) = P(\sim h|e), \quad \text{but } P(\sim h) = 1 - P(h). \quad \square$$

Intuitively, since the CFs are the same for each of these rules, no information is gained through their usage. Accordingly, the hypothesis is independent of the evidence: the rules are "wasted." In fact, they are misleading and may signal erroneous use of evidence. Further, these are "cumulative CF bloating" rules. They

raise CFs associated with h and $\sim h$, but provide no discrimination between the two.

This theorem can be extended to additional similar symmetry cases for CFs. In addition, the theorem can also be extended to cases where there are other constructs based on conditional probabilities, on the rules.

A closely related situation is that of redundancy. In this case, rules of the type "if e then h (at value v) CF k" and "if e then $\sim h$ (at value $\sim v$) CF k" are redundant. There are a number of reasons for removing redundant rules as noted in O'Keefe and O'Leary (1993). For example, keeping consistency between different redundant rules is difficult. Maintenance could change one and not the other.

Case Study. Apparently, at least one type of redundancy is found in OC. In the OC rules in Table 1, we see the following type of symmetry.

3. If the strategy is prospector then the centralization is low (CF 20).

9. If the strategy is prospector then the decentralization is high (CF 20).

The rules are redundant. Two rules capture the same knowledge.

5. Language in Rules

Another concern with weights derives from the language of the rules. In some situations, that language can be ambiguous. Accordingly, different users of the system may respond to the same rule request for evidence with different evaluations (e or $\sim e$), ultimately providing the same system different input for the exact same situations. If the language results in ambiguity as to whether evidence is true or not, that places another level of uncertainty in the model that might be reflected in the weights.

What Do the Rules Mean?

Consider the rule "If e then h." Suppose that there is ambiguity about whether an input is "e" or "$\sim e$" because of the language. For example, some rules use of terms such as "low" and "high," "increasing" or "decreasing." In order to assess the impact of such ambiguous terms, the author performed a study where 76 consultants from the Los Angeles office of one of the largest consulting firms, were asked to provide an assessment as to the extent to which a sequence of rules were true. That empirical research indicates that such terms are subject to substantial variation in interpretation. As part of the study, the consultants were asked to assess the following case, based on an expert system rule:

Experts indicate that the existence of one ***large*** *delinquent charge has an impact on the collectability of that charge from the customer.*
Sales to client E are about \$30,000 *per year. Client E has a single outstanding payable of* \$1,000. *Indicate whether that account is*

Small						*Large*
1	2	3	4	5	6	7

The results found the following distribution of assessments:

Assessment Value	1	2	3	4	5	6	7
Number of Evaluations	7	20	17	15	10	4	3.

Accordingly, there can be substantial semantic ambiguity associated with certain language, referred to here as semantic ambiguity. That semantic ambiguity creates an additional level of uncertainty that is not accounted for in the certainty factors.

Case Study. There is some concern as to the clarity of some of the rules in OC. For example, rule #1 uses the term "large," which as noted above can contain substantial semantic ambiguity. In addition, other rules also contain semantic ambiguity, e.g., rule #7,

"If the strategy is prospector and the technology is routine then this may cause problems."

What does it mean to "cause problems?" Are the terms "technology" and "routine" subject to interpretation? There are other examples of language ambiguity scattered throughout the other rules.

Rule Language Ambiguity and Weights

Given that there is some difficulty in estimating the MYCIN weights, an important issue is the impact of that language ambiguity on these weights. This section extends previous research on Bayesian weights (O'Leary 1995) to the study of the impact of ambiguity in the language of the rules on MYCIN weights.

Language ambiguity is introduced into the MYCIN models of $MB(h, e)$ and $MD(h, e)$. The rules in a system relate evidence (e) to hypothesis (h). However, a user of the system actually would see data (d) that they would then categorize as evidence (e or $\sim e$). Once categorized, the system would use the rules to inference through the knowledge base.

The approach is to use Bayes' Theorem on the only probability that is influenced by the language of the evidence e, $p(h|e)$, in order to develop $p(h|d)$. This is done by focusing on the data that the user sees, rather than just conceptual evidence e, developing the measures $MB(h, d)$ and $MD(h, d)$. In order to study the behavior of MB and MD in concert with semantic ambiguity, the original versions without the max and min operators are used, so that

$$MB[h, e] = [P(h|e) - P(h)]/[1 - P(h)] \quad \text{and}$$

$$MD[h, e] = [P(h) - P(h|e)]/P(h).$$

THEOREM 2. (A) $P(h|d) = [P(h|e)*P(e)*P(d|e \text{ and } h) + P(h|\sim e)*P(e')*P(d|\sim e \text{ and } h)]/P(d)$.

(B) $P(h|d) = P(e|d)*P(h|e \text{ and } d) + P(\sim e|d)*P(h|\sim e \text{ and } d)$.

Formulation (A) has the advantage of incorporating $P(h|e)$, the original form, allowing direct comparison with $P(h|d)$ in order to assess the impact of accounting for semantic ambiguity. As a result, the primary focus in this paper is on (A). Formulation (B) has the advantage of requiring fewer probabilities and thus, may be easier to use in a real world situation.

The factors $P(d|e \text{ and } h)$ and $P(d|\sim e \text{ and } h)$ contain the information that relates to the ambiguity in the relationship between the evidence and the data. In this paper, it is assumed that the data does not depend on the hypothesis h. In this case those factors reduce to $P(d|e)$ and $P(d|\sim e)$. This allows us to rewrite $P(h|d)$ in formulation (A) as

$$P(h|d) = [P(h|e)*P(e)*P(d|e) + P(h|\sim e)*P(\sim e)*P(d|\sim e)]/P(d).$$

This paper examines one particular case of $P(d|\cdot)$. It is assumed that $P(d|e)$ is symmetric, so that $P(d|e) = p(\sim d|\sim e)$ (e.g. Schum and DuCharme 1971), however, the results could be extended to other sets of assumptions. Some sample values illustrating the impact of semantic ambiguity on the weights are summarized in Table 3. Based on this example, it is apparent that accounting for semantic ambiguity results in weights that are different than weights without accounting for semantic ambiguity.

When is the model that includes semantic ambiguity the same as the model that does not include the semantic ambiguity? Theorem 3 provides a set of conditions. The theorem indicates that when there is no semantic ambiguity the weights under the semantic ambiguity formulation are the same as the definition of CFs.

THEOREM 3 (EQUIVALENCE OF $MB(h, e)$ AND $MB(h, d)$). *If* $P(e) = P(d)$, $P(d|e) = 1$ *and* $P(d|e') = 0$, *then* $P(h|e) = P(h|d)$.

These last three sections have focused on limitations of the weights, the difficulty of developing weights and the impact of language on weights. However, the domain also influences the suitability of MYCIN for use in the modeling of complex business problems.

6. The Impact of the Domain

MYCIN was developed to model medical decision making. The domain influenced the method of combining

Table 3 **Example Impact of Introducing Semantic Ambiguity into MYCIN Weights**

Semantic Ambiguity $P(d\|e)$	$P(h\|e) = 0.7$ MB = 0.5 and MD = −0.75		$P(h\|e) = 0.9$ MB = 0.83 and MB = −1.25	
1.0	0.5	−0.75	0.83	−1.25
0.9	0.4	−0.60	0.70	−1.05
0.8	0.3	−0.45	0.57	−0.85
0.7	0.2	−0.30	0.43	−0.65
0.6	0.1	−0.15	0.29	−0.45
0.5	0.0	0.00	0.16	−0.25
0.4	−0.1	0.15	0.03	−0.05
0.3	−0.2	0.30	−0.10	0.15
0.2	−0.3	0.45	−0.23	0.35
0.1	−0.4	0.60	−0.37	0.55
0.0	−0.5	0.75	−0.50	0.75

Assumptions: $P(e) = P(\sim e) = P(d) = 0.5$, $P(h|\sim e) = 0.1$, $P(h) = 0.4$, $P(d|e) + P(d|\sim e) = 1$.

the weights on the rules in the inference process, the use of diagnostic as opposed to causal reasoning, the use of the acyclic "evidence–hypothesis" rule structure and the use of a single criterion for optimization (i.e., the CFs were used to rank the outcomes).

Combining Weights

As noted in Buchanan and Shortliffe (1985, p. 211) when discussing the development of the certainty factors, "Thus, we sought to show that the CF model allowed MYCIN to reach good decisions comparable to those of experts and intelligible both to experts and to the intended user community of practicing physicians." Further, Buchanan and Shortliffe (1985, p. 234) note that the CF model was ". . . conceived with medical decision making in mind. . . ."

Although Buchanan and Shortliffe (1985) also note that it is potentially applicable to other problem domains, there is concern about its "generalizability" to other settings for a number of reasons. First, unfortunately, one of the major limitations of MYCIN, as noted by Buchanan and Shortliffe (1985, p. 213) is the rapidity with which the cumulative CF will converge to 1, "no matter how small the CFs of the individual rules are." Accordingly, as Buchanan and Shortliffe (1985, pp. 213–214) note,

> For some problem areas, therefore, the combining function needs to be revised. For example, damping factors of various sorts could be devised (but were not) that would remedy this problem in ways that are meaningful for various domains. In MYCIN's domain of *infectious diseases*, however, this potential problem never became serious. (italics added)

The rapidity with which rules for which there is little confidence can be combined to form "substantial" confidence is illustrated in Table 4. For example, the combination of four rules with CF's of 0.2 ("not known—nil") yields about 0.6, a CF with "substantial" confidence.

Second, MYCIN combination properties were found to be stable. It was found that for those CF's > 0.2 (Buchanan and Shortliffe 1985, pp. 224–225), the system was very stable. Ultimately, this was attributed to the domain

> *The observed stability of therapy despite changing organism lists probably results because a single drug will cover many organisms, a property of the domain.* (Buchanan and Shortliffe 1985, p. 219)

Table 4 Combination of MYCIN Weights[1]

	Size of Weight		
x = No. of Rules	0.10	0.20	0.25
2	0.1900	0.3600	0.4375
3	0.2710	0.4880	0.5785
4	0.3439	0.5904	0.6835
5	0.4095	0.6723	0.7626

[1] For example, for $x = 2$, and size of weight = 0.10, two rules each with CF of 0.10 are combined to yield a combined CF of 0.19.

Case Study. Unfortunately, few of the weights in OC were greater than 0.2. As a result, the Buchanan and Shortliffe results on stability are not necessarily appropriate in the context of OC.

Baligh et al. (1996) recognize that the various heterogeneous knowledge sources combined in the system cannot be combined using meta knowledge, so after assembling a set of atomistic rules, Baligh et al. (1996) use MYCIN's combination approach to generate rules that consider a broader base of factors. Accordingly, Baligh et al. (1996) seem to indicate that it is in this way that MYCIN can combine CFs to provide increased evidence of an hypothesis. However, Baligh et al. (1996) provide no apparent analysis or theory as to why MYCIN would be appropriate, as an ad hoc approach, for organization design. Domain makes a difference, since with just a few rules, the CFs converge to 1 or at least very large values. Although, as Buchanan and Shortliffe (1985) noted, the rapid convergence resulting from combining evidence did not seem to make a difference in the medical domain, it is unclear, if the ability to chain together a few rules and rapidly make the CFs approach 1, is appropriate for organizational design.

Further, as noted by Baligh et al. (1996), a critical question is "Given that both the original rules make sense considered separately, do the combined results also make sense?" Does it make sense to use a system designed for combining ad hoc measures of uncertainty in medical decision making to combine rules for purposes of organizational design? In particular, should we use a medical decision making system that allows combination of a number of factors that are "not

known—nil" to generate diagnostic measures that are quite large, for organization design?

Diagnostic Versus Causal Reasoning

There is evidence that medicine places substantial weight on the diagnosis form of reasoning (e.g., Buchanan and Shortliffe 1985). As a result, it is not surprising that MYCIN reasons from evidence (e) to hypothesis (h) using "diagnostic" reasoning (also called "evidential reasoning"). This is in contrast to "causal reasoning," modeled using cause and effect relationships.

There has been some empirical research investigating usage of diagnostic and causal reasoning. Tversky and Kahneman (1980) investigated judgments of conditional probabilities of the type $P(X \mid D)$. They defined X to be some target event and D to be data or evidence. If D is perceived as the cause of the occurrence or non occurrence of X then D was referred to as "casual datum." If X is treated as a possible cause of D then D was referred to as "diagnostic datum." Tversky and Kahneman (1980, p. 50) found that ". . . people assign greater impact to causal than to diagnostic data of equal informativeness." This study led Pearl (1988, p. 151) to suggest that people prefer to encode experimental knowledge in a cause and effect schema.

Case Study. Since OC was developed using MYCIN, OC uses evidential reasoning. If it is true that people prefer to reason using causal reasoning then the evidential-based diagnostic reasoning, used in MYCIN, may not be appropriate or optimal for complex management problems, instead, a causal reasoning approach should be used.

Abductive Reasoning

Human reasoning typically employs what is referred to as "abductive reasoning." If the rule "if e then h," is true, then that makes the value "e" more credible (e.g., Pearl 1989, p. 7). Going from "e" to "h" is diagnostic, while going from "h" to "e" is predictive.

As noted by Pearl (1989, pp. 501–502), "The MYCIN system . . . admits only evidential rules (always pointing from evidence to hypothesis); it can perform simple diagnoses but cannot combine diagnoses with prediction. . . ." Rule-based systems like MYCIN do not allow "cycles" in their reasoning ($e - > h - > e$); they assume that the underlying set of rules is acyclic, since otherwise their inference engines will just cycle continuously through the knowledge. As a result, such systems cannot perform abductive reasoning. Accordingly, there is only diagnostic reasoning, and no predictive inferences can be made about the evidence. As a result, rule-based systems are limited in their ability to model human decision making and limited in the complexity of the decisions that they can be used to model.

Case Study. Since OC uses MYCIN it cannot use abductive reasoning. Complex management problems, such as organization design, rarely only are concerned with diagnosis, but are often concerned with prediction. Accordingly, systems designed to help solve these complex problems are likely to require abductive reasoning. As a result, rule-based systems, such as MYCIN may not be appropriate for modeling systems for support of complex management problems.

Using CFs Means Optimization of a Single Criterion

A system like MYCIN ultimately provides a ranking (by cumulative CFs) of lines of reasoning. As a result, systems such as MYCIN permit the capture of uncertainty, extending problem solving capabilities beyond those of deterministic rule-based systems. However, using certainty factors focuses on optimization of a single criterion, the largest certainty factor. Inevitably, single criterion optimization, ignores a number of important issues, such as cost and benefit, that are critical components of virtually all decision analysis (e.g., O'Leary 1986). Although ranking according to cumulative certainty factor may be appropriate in the design of medical systems, an important aspect of analyzing complex management problems is the ability to investigate multiple criterion, analyze tradeoffs and the ability to put analyses into dollar terms in the context of the model.

Case Study. The role of the OC seems to be limited to ranking organizational design factors based on cumulative certainty factors. Unfortunately, systems for organizational design need to take into account cost benefit tradeoffs and a variety of other criteria that are actually used in organizational design. For example, an organization may choose a particular organization form, because they are willing to incur the cost associated with that form in terms of communications costs, because of benefits such

as flexibility. Unfortunately, a rule-based approach, like that generated using a tool like MYCIN, does not permit analysis of such trade-offs or ranking using any approach other than cumulative certainty factors. A multiple criterion approach would more likely meet the needs of users of a system such as OC.

7. An Alternative Architecture to Rule-based Expert Systems

Most of this paper has discussed verifying uncertain knowledge bases, focusing on, e.g., limitations of using MYCIN's ad hoc CFs. Just as important to finding limitations is the process of proposing an alternative.

Bayesian Models Are Available

Heckerman (1986) was among the first to question the use of ad hoc models of uncertainty in artificial intelligence. In particular, Heckerman (1986) finds that MYCIN has a number of assumptions that are rarely true in practical situations. Accordingly, there is interest in alternative approaches that are not so limited.

One important set of possible models is formal probability theory. As noted by Shortliffe (1991, p. xv), in the 1970s researchers identified a number of problems with using formal probability methods, and accordingly abandoned probability theory. However, he also notes that

> In recent years there has been a resurgence of interest in the use of more formal probabilistic models to handle uncertainty in large artificial intelligence (AI) systems. Investigators have concentrated on knowledge acquisition and on Bayesian inference using the belief network (also called knowledge map), which is a graphical representation of uncertain knowledge based on probability theory.

Shortliffe (1991) goes on to argue that with the development of "Pathfinder," Heckerman (1991) has ". . . demonstrated that, using a probabilistic framework to elicit and encode the knowledge of domain experts, he could construct a useful system." One of the developers of MYCIN recognizes that formal probability theory can now be used instead of the ad hoc MYCIN approach.

Bayes' net models, such as the one in Heckerman (1992), provide an alternative architecture that mitigates many of the limitations discussed above. Although a detailed discussion of Bayes' nets generally is beyond the scope of this paper, I will briefly summarize some of the advantages and costs, as compared to classic expert systems.

- Rather than a classic rule-based system, the Bayes' net approach would be used. Recently, tools that use Bayes' nets have become widely available and very easy to use. Given the right software (e.g., HUGIN,[2] see Lauritzen and Spiegelhalter 1988), Bayes' nets are much easier to draw than it is to generate rules. Graphical structures that represent rules can be generated literally using a "point and click," as compared to rules that must be phrased and typed and debugged for syntactic and semantic content (see below for an example).
- Rather than ambiguous certainty factors, conditional probabilities would be gathered. As seen above, certainty factors are not well-defined and are difficult to generate. Although probabilities have easy-to-use tests of correctness (e.g., they sum to one), with CFs there are no such constraints. As a result,

(a) developers are less likely to come up with many "nil" relationships, since probabilities sum to one and

(b) developers are less likely to establish "unusual" symmetries or redundancies, as has been done with CFs, because there is less confusion over what probabilities are or represent.

- Bayes' nets can explicitly represent semantic ambiguity in the network.
- Bayes' nets use Bayesian combination (rather than an ad hoc approach) and are more "completely" specified probabilistically than a rule-based system that uses certainty factors. For example, given six rules of the type

If A1 then B1 (cf-1) or B2 (cf-2);

If A2 then B1 (cf-3) or B2 (cf-4);

If B1 then C1 (cf-5) or C2 (cf-6);

If B2 then C1 (cf-7) or C2 (cf-8);

If A1 then C1 (cf-9) or C2 (cf-10);

If A2 then C1 (cf-11) or C2 (cf-12);

generates 12 certainty factors, where the consequences (the Bs and Cs) are each solely conditioned on a single

[2] HUGIN is one of many Bayes' net packages available for a "Windows" environment. Information on availability is on the world wide web at http://www.hugin.dk/products.html

variable. However, an equivalent Bayes net would require twelve conditional probabilities, of which the set of eight going in to C1 and C2 would be conditioned on both the As and Bs, as seen in Figure 1. Apparently, the probabilistic structure of expert systems is underspecified, when compared to Bayes' nets. Accordingly, the benefits of using Bayes' nets, include correct specification of the probabilities, according to Bayesian probability.

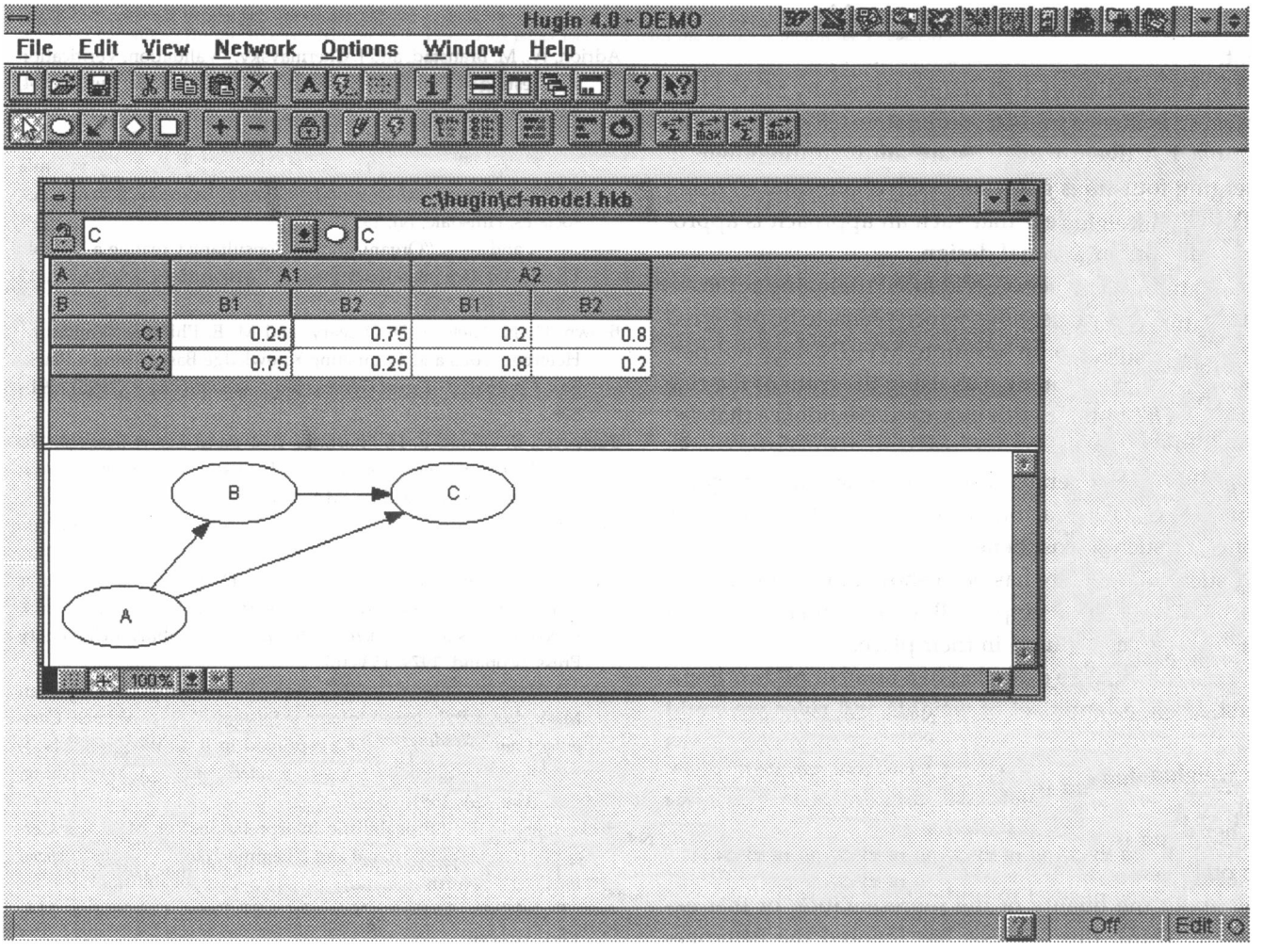
• An extension of Bayes' nets, referred to as influence diagrams (e.g., Pearl 1989) allows maximization of other criteria, e.g., dollars. Thus, models can be developed that provide the important ability to compare alternatives using a well-defined measure that facilitates trade-offs.

8. Summary and Extensions

This paper results in a number of findings about verification of uncertain knowledge-based systems with particular emphasis on MYCIN-like systems in general.

• MYCIN weights in any given system can be analyzed independently for their "meaning," e.g., "not known" or "certain." "Not known" weights are a potential problem. Such weights are given a value of "nil" in deterministic versions of the system and a large number

Figure 1



of "not known" weights suggests that the rules are not contributing knowledge to the problem solution.

• It may be that the distributions of weights indicate something about expertise that can be used to understand whether that expertise has be captured correctly. The weights in MYCIN were bimodal, with both strong and weak associations.

• There is increasing evidence that people are not "good" at developing CFs. Ultimately, people seem to build in symmetries and redundancies that may appear reasonable, but that signal "usual" assumptions about the weights and their underlying probabilities. Rule pairs of the type "if e then h" CF = k and "if e then $\sim h$" CF = k, "bloat" cumulative CFs. Rule pairs of the type "if e then h (at value v) CF k" and "if e then $\sim h$ (at value $\sim v$) CF k" are redundant.

• MYCIN was developed for an application in the medical domain and is domain-specific to a certain extent.

(a) The resulting CF structure defined by MYCIN for merging evidence leads to cumulative CF's that rapidly approach 1. Even if the CFs are all 0.2 ("unknown"), merging four rules generates a CF of almost 0.60 using MYCIN. It is not clear that such an approach is appropriate for organizational design.

(b) MYCIN focuses on evidential reasoning at the exclusion of causal reasoning, which seems to be preferred by decision makers. In addition, MYCIN does not permit abductive reasoning, i.e., using the truth of the rule "if e then h" to infer with increased confidence that "e" is more credible.

(c) MYCIN generates a single criteria ranking, by cumulative certainty factors, which ignores basic business concerns such as cost-benefit tradeoffs.

• Further, MYCIN has been shown to be suboptimal (since it does not employ Bayesian inference), and Bayes' nets can be used in their place.

The paper can be extended in a number of directions. First, given a larger set of weights and rules we could make some statements about outliers and distributions of weights (e.g., O'Leary and Kandelin 1988). Second, further research is needed to understand the nature of expertise captured in distributions of certainty factors or other such measures, and what those distributions say about the quality of the representation of that expertise. Third, this analysis has ignored the heterogeneous nature of the knowledge used to generate the knowledge base. Verification of such knowledge bases has received only limited attention to date (e.g., Brown et al. 1995). However, such an investigation is beyond the scope of this particular paper. Fourth, this paper provided a limited analysis of the impact of semantic ambiguity on MYCIN weights, focusing on a symmetric model. That analysis could be extended to other models, such as the asymmetric case.[3]

[3] The author would like to thank the referees for their comments on an earlier version of this paper. This paper should not be read as a criticism of "Organizational Consultant." It was chosen as a case study since it is a system designed to support management, derived from heterogeneous knowledge sources and employing uncertainty reasoning, through the use of certainty factors. In certain ways the system is bold and innovative.

References

Adrion, W., M. Branstad, and J. Cherniavsky, "Validation, Verification and Testing of Computer Software," *ACM Computing Surveys*, 14, 2 (1982), 159–192.

Baligh, H. H., R. Burton, and B. Obel, "Validating an Expert System that Designs Organizations," in K. M. Carley and M. J. Prietula (Eds.), *Computational Organization Theory*, Laurence Erlbaum Associates, Hillsdale, NJ, 1994.

——, ——, and ——, "Organizational Consultant: Creating a Useable Theory for Organizational Design," *Management Science*, 42, 12 (1996), 1648–1662.

Brown, C., N. Nielson, D. O'Leary, and M. E. Phillips, "Validating Heterogeneous and Competing Knowledge Bases Using a Black-box Approach," *Expert Systems with Applications*, 9 (1995), 591–598.

Buchanan, B. G. and E. H. Shortliffe, *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, Addison-Wesley, Reading, MA, May 1985.

Cimflex Teknowledge, *M.4 User's Guide*, Cimflex Teknowledge, Palo Alto, CA, 1991.

Duda, R., J. Gashnig, and P. Hart, "Model Design in the Prospector Consultant Systems for Mineral Exploration," in D. Mitchie (Ed.), *Expert Systems for the Micro Electronic Age*, Edinburgh University Press, Scotland, 1979, 153–167.

——, P. Hart, N. Nilsson, and G. Sutherland, "Subjective Bayesian Methods for Rule-based Inference Systems," *Proc. National Computer Conf.*, 1976, 1075–1082; reprinted in B. L. Weber and N. J. Nilsson (Eds.), *Readings in Artificial Intelligence*, Tioga Publishing, Palo Alto, CA, 1981.

Heckerman, D. E., "Probabilistic Interpretations for MYCIN's Certainty Factors," in L. Kanal and J. Lemmer, *Uncertainty in Artificial Intelligence*, North-Holland, New York, 1986, 167–196.

——, *Probabilistic Similarity Networks*, MIT Press, Cambridge, MA, 1991, xv–xvii.

Kanal, L. and J. Lemmer, *Uncertainty in Artificial Intelligence*, North-Holland, New York, 1986.

Lauritzen, S. L. and D. J. Speigelhalter, "Local Computations with Probabilities on Graphic Structures and their Applications to Expert Systems," *J. Royal Statistical Society*, 157 (Series B, Methodological), 50 (1988), 205–247.

Murrell, S. and R. Plant, "A Survey of Tools for Validation and Verification 1985–1995," forthcoming, *Decision Support Systems*, 1996.

Napolean, R., *Probabilistic Reasoning in Expert Systems*, John Wiley & Sons, New York, 1990.

O'Keefe, R. M. and D. E. O'Leary, "Expert System Verification and Validation," *Artificial Intelligence Review*, 7 (1993), 3–42.

O'Leary, D. E., "Multiple Criterion Decision Making Evaluation Functions in Expert Systems," *The 6th International Symposium on Expert Systems and Their Applications*, Avignon, France, April 1986, 1017–1035.

——, "Validation of Expert Systems," *Decision Sci.*, 18 (Summer 1987), 468–486.

——, "Methods of Validating Expert Systems," *Interfaces*, 18, 6 (1988), 72–79.

——, "Soliciting Weights or Probabilities from Experts for Rule-based Systems," *International J. Man-Machine Studies*, 32 (1990), 293–301.

——, "The Impact of Semantic Ambiguity on Bayesian Weights," *European J. Oper. Res.*, 84 (1995), 163–169.

—— and N. Kandelin, "Validating the Weights in Rule-based Expert Systems: A Statistical Approach," *International J. Expert Systems: Research and Applications*, 1, 3 (1988).

Pearl, J., *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufman, San Mateo, CA, 1989.

Shortliffe, E. H., "Foreward," in D. E. Heckerman, *Probability Similarity Networks*, MIT Press, Cambridge, MA, 1991, xv–xvii.

Schum, D. and W. DuCharme, "Comments on the Relationship Between Impact of Reliability and Evidence," *Organizational Behavior and Human Performance*, 6 (1971), 111–131.

Teknowledge, *M.1 Reference Manual*, Teknowledge, Palo Alto, CA, December 1986.

Tversky, A. and D. Kahneman, "Causal Schemata in Judgments Under Uncertainty," in M. Fishbein (Ed.), *Progress in Social Psychology*, Lawrence Erlbaum, Hillsdale, NJ, 1980.

Accepted by Gabriel Bitran; received November 1, 1995. *This paper has been with the author* 1 *month for* 1 *revision.*

References

Organizational Consultant: Creating a Useable Theory for Organizational Design
Helmy H. Baligh; Richard M. Burton; Børge Obel
Management Science, Vol. 42, No. 12. (Dec., 1996), pp. 1648-1662.
Stable URL:
http://links.jstor.org/sici?sici=0025-1909%28199612%2942%3A12%3C1648%3AOCCAUT%3E2.0.CO%3B2-S