13. E. Shortliffe. *Computer-Based Medical Consultation: MYCIN.* North-Holland, Amsterdam, 1976.

14. J. Sticklen, B. Chandrasekaran, J.R. Josephson. Control Issues in Classificatory Diagnosis. *Proceedings of the 9th IJCAI*, University of California, Los Angeles, California, August 18-24, 1985, pp. 300-306.

15. A. Tversky, D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science 185*(1974), 1124-1131.

16. A. Tversky, D. Kahneman. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review 90*, 4 (October 1983), 293-315.

17. L.A. Zadeh. Commonsense knowledge representation based on fuzzy logic. *Computer 16*, 10 (October 1983), 61-65.

Uncertainty in Artificial Intelligence
L.N. Kanal and J.F. Lemmer (Editors)

PROBABILISTIC REASONING IN PREDICTIVE EXPERT SYSTEMS

David J. SPIEGELHALTER

MRC Biostatistics Unit
5 Shaftesbury Road
Cambridge CB2 2BW
England

Techniques in developing a coherent probabilistic reasoning system are illustrated with reference to a simplified example. Recent work relating statistical models to graphical representation of causal and associative relationships allows a straightforward means of propagating evidence whilst retaining a probabilistic interpretation for predictive statements. This interpretation allows continual criticism of a system's performance, while imprecise quantitative assessments permit learning from experience. Possible limitations of a formal probabilistic approach are discussed.

1. INTRODUCTION

The pervasive, and often casual, disregard for probabilistic methods for handling uncertainty in expert systems is now being faced by an increasing counter-attack spiritedly mounted by, among others, Pearl, Cheeseman, and Lemmer. Many misconceptions, it is to be hoped, have been laid to rest by Cheeseman (1985) and other contributors to this volume, particularly with regard to the misleading concept of there being 'true' precise probabilities that either must be obtained from extensive data analysis, or - failing that - the whole probabilistic structure must be rejected in place of an ad-hoc formalism.

It could be argued that there is a danger of overstating the universal appropriateness of probabilistic methods, and alienating AI practitioners by the apparent complexity of the numerical techniques required. However, we believe that recent results in theoretical statistics, combined with the literature on subjective Bayesian methods, provide considerable insight into recognising both where probabilistic reasoning is necessary, and how it may be implemented in a relatively straightforward manner.

Contrasts with non-probabilistic methods are made in Spiegelhalter (1986a) and the arguments are briefly summarised in Section 4; there it is concluded that if a system is to be judged, in part, by its ability to make numerical predictions concerning events that can later be verified, then probability is both theoretically and practically appropriate. However, the bulk of this paper consists of taking a small, stylised piece of medical 'knowledge', and stepping through the stages of implementing a probabilistic reasoning system, illustrating a number of issues raised in artificial intelligence research.

The ten stages considered are:-

1. qualitative representation of relationships as a directed graph;
2. quantitative expression of subjective beliefs;
3. efficient storage as an undirected graph;
4. coherent evidence propagation through the graph;
5. 'uncertain' evidence;
6. 'sensitivity' of probabilities due to limited evidence;
7. 'imprecise' probabilities due to limited knowledge;
8. using data to learn about quantitative assessments;
9. using data to learn about qualitative structure;
10. explanation of conclusions.

In this paper, these issues can only be covered briefly in relation to a small example; the discussion is informal, and stages 7-10 are particularly tentative. However, the aim is to emphasise the conceptual clarity and computational ease which can be gained by having a unified, coherent approach to dealing with uncertainty; this point has been particularly strongly made by Pearl (1985), who covers a number of the stages listed above. For more extended discussion with relation to a probabilistic diagnostic system which is in use in a number of gastroenterological clinics, see Spiegelhalter and Knill-Jones (1984).

2. AN EXAMPLE

We consider a deliberately restricted piece of medical 'knowledge' which has been previously dealt with in more detail by Cooper (1984) :

> "Metastatic cancer is a possible cause of a brain tumour, and is also an explanation for increased total serum calcium. In turn, either of these could explain a patient falling into a coma. Severe headache is also possibly associated with a brain tumour."

The stages outlined in the Introduction allow the gradual incorporation of relevant statistical techniques.

2.1. Qualitative representation of relationships as a directed graph

The description above is intended to imitate the type of 'knowledge' contained in the causal network representation used by the CASNET (Weiss et al, 1978) and INTERNIST/CADUCEUS (Pople, 1982) systems. A causal graph, displayed in Figure 1, summarises the relationships.

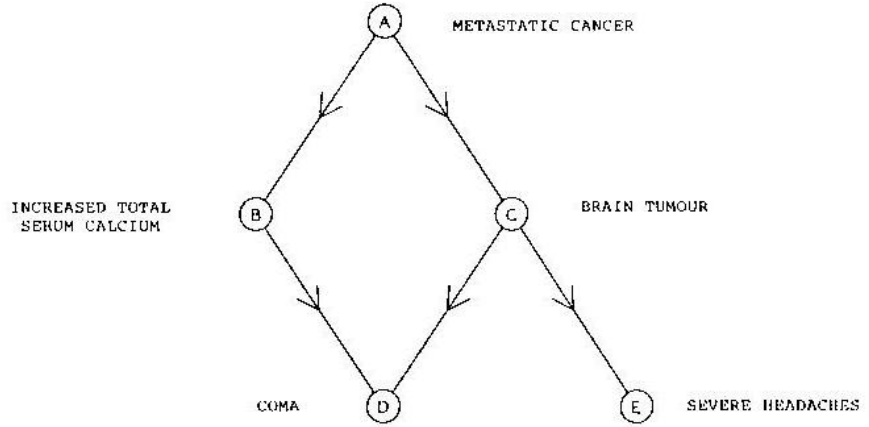

FIGURE 1.

'Causal' model expressed as a directed graph

Here, the term 'causal' is given an extremely wide interpretation and is not necessarily restricted to direct physiological reasoning. The directed links are intended to represent any natural cognitive ordering that will, as we shall see in the next section, allow reasonably confident probability assessments. Naturally, this is a highly stylised representation - for example, no element of 'time' is explicitly incorporated, although it could be. We are also assuming events such as 'increased calcium' are explicitly defined; if levels of increase are required, non-binary variables may be used. We note the graph in Figure 1 is a basic 'static' graph before a new case is encountered; as we shall see later, this graph becomes dynamic in any particular application, (see van Melle et al (1981) for a similar notion in EMYCIN).

Such directed graphs have a long history in the social sciences; see, for example, Wright (1934), Wold (1954) and Blalock (1971). When used as a description of a statistical model, certain conditional independence assumptions can be read directly from the graph, which is also called a 'Bayes network' by Pearl (1986) and an 'influence diagram' by Shachter (1986). Both suggest such graphs as cognitive models rather than necessarily being estimated from data, and we refer to them for further discussion as to their construction.

In brief, a 'node' in the graph represents a random variable, say A, which in our case is assumed to be a proposition taking values 'true' or 'false', denoted by a and $\bar{a}$ respectively. An arrow between two nodes, say A to B, represents a 'causal' relationship, and we shall say that node A 'precedes' node E in the graph, since there is a directed path between A and E. A graph without directed cycles, known as a 'recursive causal graph' (Kiiveri et al, 1984) allows at least one ordering of the nodes which conforms with precedence - the labelling A,B,C,D,E in Figure 1 represents one such ordering in which we shall say A is of 'lower order' than B, etc. Shachter describes the independence assumptions expressed in the graph, where a node is independent of all nodes with lower order, conditional on those that directly precede it. Pearl provides a more graph-theoretic formulation. In our problem, for

example, we are claiming that in patients with confirmed brain tumours, knowledge concerning coma provides no information as to the likelihood of having experienced severe headaches. If this is not reasonable, then the grapical model should be changed to incorporate, say, a directed link between D and E, or an intermediate node below C, conditional on which D and E are independent.

The joint distribution of any five variables A,B,C,D,E may always be factorised into the product form -

$$p(A,B,C,D,E) = p(E|A,B,C,D)\ p(D|A,B,C)\ p(C|A,B)\ p(B|A)\ p(A)$$

but from the graph this can be simplified to

$$p(E|C)\ p(D|B,C)\ p(C|A)\ p(B|A)\ p(A)$$

showing the dependencies explicitly. This type of factorisation, described in detail in Kiiveri et al (1984), links into a rule-based representation, since we need only consider a series of local relationships in order to build up the entire structure. Later we shall see how certain groupings of such 'rules' lead naturally into a frame-based representation.

2.2. Quantitative expression of subjective beliefs

A series of conditional probability assessments are now required. We initially make the unrealistic assumption that these can be made precisely - this will be relaxed in Section 2.7. Eleven numbers are necessary to completely specify our belief about A,B,C,D and E, and fictitious examples are given below together with the features they are intended to represent:

$p(e\|c) = .80$ $p(e\|\bar{c}) = .60$	}	headaches common, but more common if tumour present
$p(d\|b,c) = .80$ $p(d\|b,\bar{c}) = .80$ $p(d\|\bar{b},c) = .80$ $p(d\|\bar{b},\bar{c}) = .05$	}	coma rare, but common if either cause present
$p(b\|a) = .80$ $p(b\|\bar{a}) = .20$	}	increased calcium uncommon, but common consequence of metastases
$p(c\|a) = .20$ $p(c\|\bar{a}) = .05$	}	brain tumour rare, and uncommon consequence of metastases
$p(a) = .20$		incidence of metastatic cancer in relevant clinic

We emphasise that a full assessment of $p(D|B,C)$ has been necessary, and that this may **not** be derived from separate assessments of $p(D|B)$ and $p(D|C)$. The final assessment of $p(a)$ is possibly the most difficult to make, since it is the one most likely to vary between sites, depending on referral policies. We do not deal with the problems of elicitation in detail, but mention that it may be more 'acceptable' to assess likelihood ratios in the manner of PROSPECTOR (Duda et al, 1976) and then solve for the conditional probabilities.

2.3. Efficient storage as an undirected graph

When the system is in operation, we assume that at any time evidence may be received about any of the nodes in the graph. The directed structure described so far is unsuitable for propagating such evidence particularly when there exist multiple paths in the network, as in Figure 1. We therefore require a simple means of storing the elicited relationships in an undirected form.

Fortunately, the statistical theory of 'graphical models' in contingency tables (see, for example, Darroch et al, 1980; Edwards and Kreiner, 1983; Lauritzen, 1982) is of direct relevance, although its application in expert systems does not appear to have been considered previously. This forms part of the general theory of Markov random fields (Isham, 1981) which are of increasing interest in spatial modelling and, in particular, image restoration (Besag, 1986; Geman and Geman, 1984), which in turn owes much to work in statistical mechanics. The qualitative relationships between variables are represented by an **undirected** graph, whose links or 'edges' represent the Markov property of our belief concerning the variables represented by the nodes. Specifically, a variable is independent of all those **not** adjacent to it, conditional on those that **are** adjacent to it. An important subclass of graphical models contains those in which there is no cycle of length 4 edges or more, without a 'short cut'; these are termed 'decomposable' and are particularly important since the relevant joint distribution may be expressed as a simple function of the marginal distributions on the 'cliques' of the graph - that is, the largest subgraphs in which the nodes are all adjacent to each other and hence for which no simplifying independence properties may be assumed. Darroch et al (1980) give a clear exposition of this work, identifying decomposable models as those for which no iterative methods are necessary when estimating parameters from data.

The vital connection between a 'directed/recursive' and an 'undirected/graphical' representation has been provided by Wermuth and Lauritzen (1983). They show that the class of recursive models and the class of graphical models intersect in the class of decomposable models, and a recursive model is a member of this intersection provided it does not have two non-adjacent nodes both directly preceding the same node. Thus Figure 1 cannot be expressed as a decomposable model as it stands, since B and C are unjoined and are both direct predecessors of D. The tree models of Kim and Pearl (1983) are also recursive without necessarily being decomposable.

However, the 'missing link' between B and C is easily introduced as a kind of 'vacuous rule', leaving us able to drop the arrows and leave the undirected representation in Figure 2. (Efficient algorithms for 'filling out' causal models to allow an equivalent undirected representation are provided by Tarjan and Yannakakis (1984) in the context of relational databases). Thus, for example, Figure 2 tells us that E is conditional independent of A, B and D, given C.

The joint distribution may be written, as before, as

$$\begin{aligned} p(A,B,C,D,E) &= p(E|A,B,C,D)\ p(D|A,B,C)\ p(C|A,B)\ p(B|A)\ p(A) \\ &= p(E|C)\ p(D|B,C)\ p(B,C|A)\ p(A) && \text{by the Markov property on Figure 2} \\ &= \frac{p(C,E)}{p(C)} \cdot \frac{p(B,C,D)}{p(B,C)} \cdot p(A,B,C) && \text{by conditional probability law.} \end{aligned}$$

- the product of the marginal distributions on the 'cliques', divided by the product of the distributions on their intersections. The general formula for deriving the joint distribution from the graph is given by Darroch et al (1980). Thus our entire belief structure can be expressed simply in terms of the marginal distributions on the cliques, which may be derived from the assessments made on the causal representation; for example,

$$p(a,b,c) = p(b|a)\ p(c|a)\ p(a) = .8 \times .2 \times .2 = .032.$$

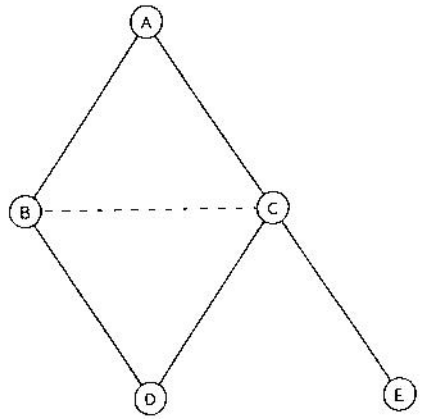

Figure 2.

Re-representation of Figure 1 as an undirected graphical model, showing added edge B-C. The structure can now be stored as three separate cliques, [A,B,C], [B,C,D], [C,E].

The explicit recognition of the conditional independence of B and C, given A, is lost in the re-representation, although it is implicitly retained in the assigned probabilities which obey the original constraints imposed by the initial 'causal' structuring.

These cliques may be best stored as self-contained frames, encapsulating knowledge of intimately related phenonema, and remembering which other frames contain insersecting variables. Thus a complex causal network representation can be broken down into autonomous local components, while retaining a strictly coherent probability model. We have therefore put into an established statistical framework the suggestions of Lemmer and Barth (1982) who also emphasise the advantages of this decomposition into a tree of cliques, using an alternative terminology in terms of 'local event groups'.

Table 1 contains, for future reference, the stored marginal distributions on the cliques.

Table 1: Marginal distributions on the cliques of Figure 2, derived from probability assessments on the causal model in Figure 1

Clique [A,B,C]		Clique [B,C,D]		Clique [C,E]	
$p(a,b,c)$	= .032	$p(b,c,d)$	= .032	$p(c,e)$	= .064
$p(\overline{a},b,c)$	= .008	$p(\overline{b},c,d)$	= .032	$p(\overline{c},e)$	= .552
$p(a,\overline{b},c)$	= .008	$p(b,\overline{c},d)$	= .224	$p(c,\overline{e})$	= .016
$p(\overline{a},\overline{b},c)$	= .032	$p(\overline{b},\overline{c},d)$	= .032	$p(\overline{c},\overline{e})$	= .368
$p(a,b,\overline{c})$	= .128	$p(b,c,\overline{d})$	= .008		
$p(\overline{a},b,\overline{c})$	= .152	$p(\overline{b},c,\overline{d})$	= .008		
$p(a,\overline{b},\overline{c})$	= .032	$p(b,\overline{c},\overline{d})$	= .056		
$p(\overline{a},\overline{b},\overline{c})$	= .608	$p(\overline{b},c,\overline{d})$	= .608		
	1.000		1.000		1.000

The interpretation of Table 1 may not be immediately transparent, but it is not intended to be communicated directly to the system user. From it we may derive, for instance, that our prior probabilities are $p(b) = .32$, $p(c) = .08$, $p(d) = .32$ and $p(e) = .616$. If our assessor did not agree that he would expect, say, 32% of patients under study to lapse into coma, then the source of inconsistency should be investigated and rectified, either by changing the quantitative assessments or the qualitative structure. This is in contrast to the PROSPECTOR approach that allows inconsistent assessments at the expense of bending the probability calculus.

In a more technical sense, we are suggesting expressing the joint relationships as a particular type of log-linear model that, were it being estimated from data, would not require iterative estimation methods (Lauritzen, 1982). In contrast, both Cheeseman (1983) and Geman (1984) suggest a representation in a more general log-linear form requiring complex algorithms for obtaining model parameters from probability assessments. Their methods do, however, explicitly deal with incomplete specification of the probabilistic relationships, and this is discussed in Section 2.7. In that section we also note that in complex problems it may be more reasonable to complete the quantitative assessments after the representation as cliques, so that an appeal to 'maximum entropy' may be made for high-order interactions.

2.4. Coherent evidence propagation through graph

Suppose our only information on a patient is that he suffers from severe headaches (E is true); and that we wish to assess how that changes our belief in him lapsing into coma i.e. $p(d|e)$. Kim and Pearl (1983) describe an efficient algorithm for propagating the implications of evidence through a graph with a tree-structure, that is, no multiple paths of the type exhibited by ABCD in Figure 1. Pearl (1986) suggests a means of dealing with general networks by conditioning on the value of a node, such as A, that 'breaks the cycle', but this requires, in our case, a direct assessment of $p(a|e)$ and it is not clear how this is to be obtained. Shachter (1986) suggests an alternative algorithm, but this appears to lose the localised representation that is essential in constructing complex models.

However, having formulated the original causal model into a decomposable graph, we can return it to a different directed representation to allow

efficient evidence propagation. Specifically, directed dependencies, or 'arrows', may be added to the graph with the realised node as the starting point, retaining the original probabilistic model, but expressed in a new recursive form. We may, for example, use 'maximum cardinality search' (Tarjan and Yannakakis, 1984) to re-order the nodes: label the realised node as '1', then at each step label the node attached to the maximum number of nodes that are already labelled; ties are broken at random. In our case, this can give the ordering displayed in Figure 3, and we note that this creates a sequence of cliques [C,E], [A,B,C][B,C,D] through which evidence can be propagated.

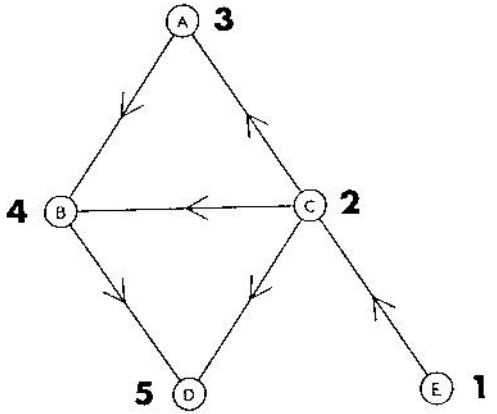

Figure 3.

Having observed E as 'true', the nodes are re-ordered, giving a sequence of cliques to recursively update in a single pass.

Suppose we indicate our current belief, given available evidence, by an asterisk. Then $p^*(c) = p(c|e)$ which may be calculated to be .104 from Table 1. Going on to the next clique, we have, for example, that

$p^*(a,b,c)$	$= p(a,b,c\|e)$	by definition
	$= p(a,b\|c,e)\ p(c\|e)$	by conditioning
	$= p(a,b\|c)\ p^*(c)$	from graphical model
	$= p(a,b,c)\ p^*(c)\ /\ p(c)$	by conditional probability
	$= .032 \times 1.3$	from Table 1
	$= .042$	

In turn, we may calculate

$p^*(b,c,d)$	$= p(b,c,d\|e)$
	$= p(d\|b,c,e)\ p(b,c\|e)$
	$= p(b,c,d)\ p^*(b,c)\ /\ p(b,c)$

which again may be obtained from current beliefs and Table 1. Thus evidence 'chains' through the cliques, each joint probability being multiplied by the ratio of the 'new belief' to the prior belief in the appropriate values of the nodes intersecting with the preceding clique. We emphasise the similarities with the results of both Cheeseman (1983) and Lemmer and Barth (1982); however, Cheeseman's propagation scheme does not provide updated probabilities for all nodes in a single pass, while Lemmer and Barth appeal to an

information-theoretic justification that appears unnecessary when the joint distribution on the 'local event groups' is fully specified.

Having propagated the influence of observing E as true, we find our new marginal probabilities are $p^*(a) = .208$, $p^*(b) = .325$, $p^*(c) = .104$ and, as our primary concern, $p^*(d) = p(d|e) = .333$ compared to our prior $p(d) = .32$; since E had quite a high prior probability, knowledge of its presence does not lead to substantial revision in beliefs, particularly in 'distant' nodes. In practice, evidence propagation could stop when the 'propagation weights' $p^*(\ldots)/p(\ldots)$ are within some small distance of 1 for all realisations of the clique intersections.

The node E can now be 'removed' from the dynamic graph for this patient, which still leaves a decomposable representation with no known nodes. Thus the asterisks can be dropped and new evidence treated as if it were the first obtained. Removal of a 'realised' node may disconnect the graph, as would have happened were C the first item elicited.

2.5. 'Uncertain' evidence

Suppose the user was fairly certain that E was false, but still wished to specify a 10% chance that severe headaches occurred. (We strongly emphasise that such a response does not correspond to a milder form of headaches; if this is a possibility then it should be explicitly included as a possible response). If we can assume that this judgement is made without any consideration of the other nodes in the network, then the dynamic graph for this patient may be temporarily extended to include a node E', as in Figure 4, producing the additional clique $[E,E']$ with, say, $p(e,e') = .030$, $p(\bar{e},e') = .270$, $p(e,\bar{e}') = .586$, $p(\bar{e},\bar{e}') = .114$. This 'coheres' with the prior $p(e) = .616$, and ensures $p(e|e') = .10$. Setting E' as 'true', and propagating its effect using the techniques of Section 2.4, provides a means of using uncertain evidence.

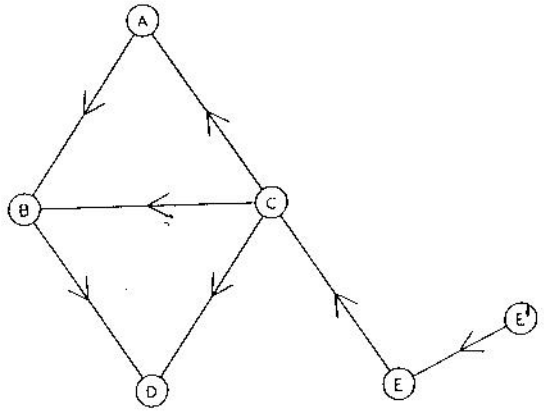

FIGURE 4

Temporary node E' introduced to produce a specified level of uncertainty concerning the truth of E; arrows show propagation of evidence from E'

If E' **does** depend in part on nodes of the original network, then this must either be explicitly represented within the graphical structure, or - preferably - a 'thought experiment' carried out in order to assess what our belief is in E ignoring all information currently represented in the system. See Pearl (1985) for a parallel discussion in terms of likelihood ratios.

2.6. 'Sensitivity' of probabilities due to limited evidence

One reason that has been given for wishing to have a range, rather than a point numerical measure of uncertainty, is to provide a means of expressing one's 'ignorance' concerning the current case in hand. We interpret this as meaning that because little evidence is available, our beliefs are extremely susceptible to change and we would not wish to make a firm probabilistic statement. Thus a patient presenting to a specialist may have a 10% chance of gastric cancer just from the known incidence in that referral clinic. However, one may be unwilling to make a decision until many further questions were asked, after which it may well be reasonable to perform an endoscopy even on the basis of the same 10% belief, since no further interrogation will substantially alter our belief.

Suppose, say, we consider our belief in an increased serum calcium, node B, which before any further evidence arrives is p(b) = .32. However, by the law of conditional probability,

$$p(b) = \sum_{A,C,D,E} p(b|A,C,D,E)\ p(A,C,D,E)$$

We can see that our current belief can be thought of as the expectation of what our belief **could** become on receipt of further evidence. The 'local' nature of the graphical representation makes this sensitivity to further evidence straightforward to investigate, since the belief in a node only depends on our beliefs in adjacent nodes. In the case of node B, this means the above expression simplifies to -

$$p(b) = \sum_{A,C,D} p(b|A,C,D)\ p(A,C,D).$$

We can easily calculate the terms in the summation by noting that the 'potential' belief

$$p(b|A,C,D) = p(A,b,C,D) \,/\, p(A,C,D), \text{ and}$$

$$p(A,C,D) = p(A,b,C,D) + p(A,\bar{b},C,D)$$

$$= \frac{p(A,b,C)\ p(b,C,D)}{p(b,C)} + \frac{p(A,\bar{b},C)\ p(\bar{b},C,D)}{p(\bar{b},C)}$$

the values of which may be obtained from Table 1. For example, our belief could become as low as $p(b|\bar{a},\bar{c},\bar{d}) = .05$ if the patient were found to have no metastases, tumour or coma; a combination we predict with probability $p(\bar{a},\bar{c},\bar{d}) = .608$. In contrast, our belief in B could become as high as $p(b|a,\bar{c},d) = .985$, if the patient were found to have metastases and lapse into coma without having a tumour, a combination of events we predict with probability $p(a,\bar{c},d) = .104$. Our distribution of possible future beliefs in B is shown in Figure 5. Other intermediate values for p*(b) are possible, of course, if A,C or D are not known with certainty.

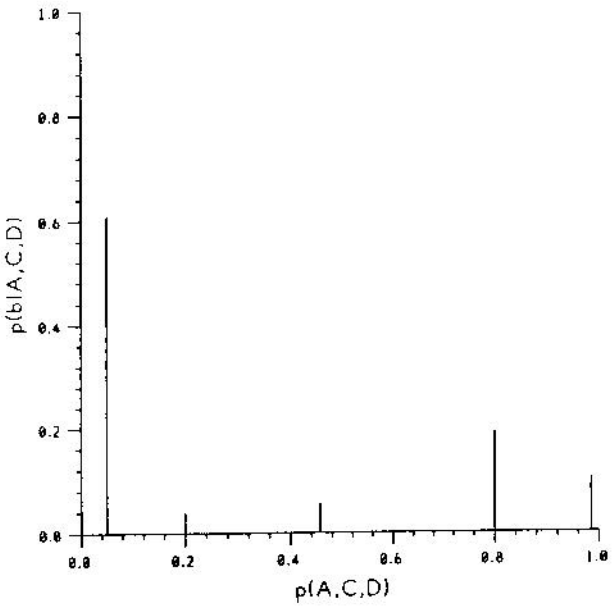

FIGURE 5.

Predictive distribution of what our belief in B could become on receipt of adjacent information - the mean is our current belief, .32

This distribution emphasises the sensitivity to further information and hence expresses our 'ignorance' concerning B. As evidence accumulates, this predictive distribution will tend to get tighter around its mean.

It may be reasonable to use some measure of the spread of this distribution for control purposes in order to guide questionning towards areas where there is still much to be learnt. The primary care system IMMEDIATE (Dodson and Rector, 1985) makes use of the **range** for this purpose, but this seems rather too sensitive to events which may have very low predictive probability.

2.7. 'Imprecise' probabilities due to limited knowledge

We have assumed so far that all probabilities are specified precisely. However, in practice there will usually be considerable doubt as to many of the quantities - in fact, such natural doubt is a prerequisite for the 'quantitative learning' procedure outlined in the next section.

Two situations should be distinguished. Firstly, an 'expert' may claim complete lack of knowledge concerning higher order interactions in cliques with 4 or more members - in this case it seems reasonable to adopt the 'maximum entropy' or 'minimum information' approach (Cheeseman, 1983) of assuming such terms are zero until evidence arrives to the contrary. Lemmer and Barth (1982), Lemmer (1983), Cheeseman (1983) and Geman (1984) all appeal to such a principle in coping with a probability distribution that is incompletely specified; Lemmer (1983) also discusses inconsistent probability

assessments, but our original causal representation avoids this problem. The second situation concerns imprecision concerning the low-level conditional probabilities of greatest importance – for example, our experience may indeed suggest that severe headaches are common in those with brain tumours, but a precise figure of $p(e|c) = .80$ seems to overstate the accuracy of the judgement. How can doubt about this figure be incorporated whilst retaining a coherent probabilistic framework?

We should first follow Cheeseman (1985) in emphasising that it is quite reasonable to allow 'second-order' uncertainty concerning probabilities. This has a respectable tradition (Fisher, 1957; Good, 1965) and, although for most decision-making purposes it is simply the mean probability that is important, there are clear psychological advantages, both for the constructor and user of the system, in allowing a measure of doubt concerning a probabilistic prediction. We simply regard the probabilities of the system as being unknown parameters with limited information concerning their value. Two approaches seem possible.

First approach : 'attached imprecisions'

The simplest way to conceptualise this is by thinking of every probability assessment, say $p(e|c) = .80$, as actually being an observed proportion based on an 'imaginary sample' of n patients with brain tumours, of whom .80n had severe headaches. By standard binomial theory, our doubt concerning $p(e|c)$ follows an approximately Gaussian curve with centre .80 and standard deviation $(.80 \times .20/n)^{\frac{1}{2}}$. Thus, if our expert were particularly ill-at-ease with his assessment of .80, he may state his imaginary sample size is only 25 of whom 20 had headaches, giving an approximate 95% interval for $p(e|c)$ of $.80 \pm 2 \times (.80 \times .20/25)^{\frac{1}{2}} = (.64,.96)$. In fact, it is often more natural to elicit the interval directly and then solve to find n. How this interval will shorten with experience will be discussed in the next section.

In constructing the original causal network, it is therefore conceivable that the directed links should have associated 'precisions' expressed as integers. The problem remains, however, of how to store and propagate such measures of precision within the graphical framework adopted, although it should be kept in mind that the 'second-order' nature of the precisions makes the use of crude approximations more justifiable than for the primary probabilities.

With regard to storage, it is initially attractive to assign a single precision number n to each clique, possibly the minimum precision attached to its constituent links. However, this ignores the fact that some relationships in the clique may be better known than others. Additionally, while the 'local' representation using cliques may be reasonable in view of conditional independence of events, it is feasible that the probabilities of such events are highly dependent. An extreme example is when one has a sequence of conditionally independent events, each occurring with an identical but currently unknown probability. This particular circumstance is discussed in more detail in the next section.

Propagation creates further problems, since a means is required of assessing the imprecision of the propagation weights $p^*(...)/p(...)$ and combining it with the imprecision of the current probabilities. This only seems straightforward under unrealistic independence assumptions.

Second approach : auxiliary nodes

This technique stays within the graphical structure since our doubt about a conditional probability, say, $p(e|c)$, is explicitly represented by creating a new, unobservable, random variable, say X, whose realisation decides the value of $p(e|c)$. Specifically, it might be reasonable to represent the doubt of the expert by saying that X could take on values corresponding to $p(e|c) = .7, .8$ and $.9$ respectively, and that $p(X=.7) = .1$, $p(X=.8) = .8$, $p(X=.9) = .1$, independently of any other evidence. This extends the network to that shown in Figure 6 with conditional probabilities

$$p(e|c,X=.7) = .7,\ p(e|c,X=.8) = .8,\ p(e|c,X=.9) = .9,$$

and hence extending the clique [C,E] to [C,E,X] with joint distribution

$$p(E,C,X) = p(E|C,X)\ p(C)\ p(X)$$

given, for reference, in Table 2.

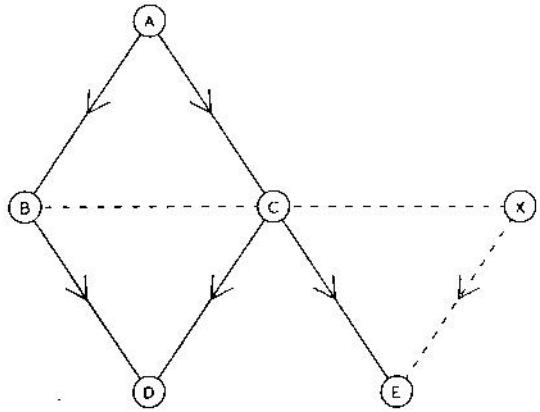

FIGURE 6.

Auxiliary node X introduced to represent doubt about the conditional probability $p(e|c)$. Dashed edges need to be added to original graph in Figure 1.

Table 2. Joint distribution on extended clique [C,E,X]

$p(c,e,X=.7) = .0056$	$p(c,\bar{e},.8) = .0128$
$p(\bar{c},e,.7) = .0552$	$p(\bar{c},\bar{e},.8) = .2944$
$p(c,\bar{e},.7) = .0024$	$p(c,e,.9) = .0072$
$p(\bar{c},\bar{e},.7) = .0368$	$p(\bar{c},e,.9) = .0552$
$p(c,e,.8) = .0512$	$p(c,\bar{e},.9) = .0008$
$p(\bar{c},e,.8) = .4416$	$p(\bar{c},\bar{e},.9) = .0368$

Now it is straightforward to see that when E is observed to be true, our conditional belief $p^*(c) = p(c|e) = .104$ is the same as it was before X was introduced. However, this belief has a certain imprecision, since

$$p(c|e) = \sum_X p(c|e,X)\ p(X|e)$$

and from Table 2 we can calculate that

$$p(c|e,X=.7) = .092\ , \qquad p(X=.7|e) = .099$$
$$p(c|e,X=.8) = .104\ , \qquad p(X=.8|e) = .800$$
$$p(c|e,X=.9) = .115\ , \qquad p(X=.9|e) = .101$$

Thus there is about a 10% chance that $p^*(c)$ could be $\pm$ 1% of its mean value of 10.4%.

We note two important aspects of this means of introducing imprecision:

Firstly, we have remained within the structure developed through this paper. At any time, the imprecision in our belief in a node, say Y, can be obtained by assessing its 'sensitivity', using a slight extension of the technique described in Section 2.6, to alternative realisations of those 'auxiliary' nodes which are either predecessors of Y in the original directed graph, or predecessors of nodes that are successors of Y, that have been observed and are still connected to Y.

Secondly, our belief in the auxiliary node has changed as a result of evidence obtained on the patient. In the next section, we see how this provides a natural mechanism for automatic learning by the system.

There is clearly a need for some consideration of imprecision in probability assessment, and the two approaches outlined above are possible candidates for study. Although both methods introduce additional problems that tend to oppose the attractive local representation obtained through the use of decomposable models, the second approach appears to have considerable potential.

2.8. Using data to learn about quantitative assessments

Any self-respecting expert system should learn by experience in order to overcome the inevitable limits on the knowledge of those who initially developed it. However, as has been previously mentioned, it is only by acknowledging imprecision in the quantitative assessments in a system that learning is possible.

Both approaches outlined in Section 2.7 have little difficulty in incorporating revision of the doubt attached to conditional probabilities. The first approach essentially stores each probability as a fraction, say $p(e|c) = 20/25$ in the given example, rather than as a single number, .8. Thus if a further patient with a tumour is observed with severe headaches, this will increase both the numerator and denominator by 1 to produce a revised conditional probability $p(e|c) = 21/26 = .81$. In the second approach, we have seen how our belief in the 'auxiliary' node X is slightly altered simply by observation of E without even knowing whether C were true or not. If we assume a common 'true' conditional probability applies to all patients, then this revised belief $p^*(x)$ should carry over to future cases to become a new initial value $p(x)$. Thus by identifying auxiliary nodes as those which should retain changes in belief when considering new cases, we have conceptually distinguished the process of 'updating' beliefs in a particular case, from 'learning' from case to case, while uniting them within a single simple computational structure.

2.9. Using data to learn about qualitative structure

It should also be possible to question the qualitative aspects of our belief, as expressed by the graphical structure. This is difficult to do directly, since independence assumptions are neither verifiable nor falsifiable on the basis of a single case. However, if a system can record 'surprise' at what it is being told, then hopefully it will be possible to identify areas of poor qualitative modelling.

Fortunately, a coherent probabilistic approach allows a simple means of monitoring 'surprise', since at any time it is essentially storing a probabilistic prediction of what the response would be, were the user questioned as to the true value of a currently unspecified node, and were able to provide a reply. Such probabilistic predictions may be evaluated by means of a 'proper scoring rule' (Winkler, 1969), which measures the discrepancy between the prediction and the truth. A popular choice is the 'Brier score' in which the squared predictive error is recorded. Thus when E (severe headache) is recorded as true, with a current probability $p(e) = .616$, the Brier score is

$$(1 - .616)^2 = .15.$$

Were headaches absent, the score would have been

$$(0 - .616)^2 = .38, \text{ registering increased surprise.}$$

Naturally, some poor predictions must be expected, even if the model is quite reasonable. However, by continuous monitoring of specific parts of the system over a series of patients, consistent 'surprise' can be identified and used to rectify the structure. Monitoring over all questions asked of a particular patient also should allow identification of 'outlying' cases who produce unexpected findings. In each type of monitoring a 'quality control' approach can be adopted. Specifically, suppose a series of n events have been recorded, to which the system has assigned predictive probabilities $p_1, \ldots, p_n$, giving a total Brier score $B = \Sigma\ (1-p_i)^2$. Were the system giving 'reliable' or 'calibrated' predictions, we would expect a Brier score of $E_0(B) = \Sigma\ p_i(1-p_i)$ with variance $V_0(B) = \Sigma\ (1-2p_i)^2 p_i(1-p_i)$. Thus $[B - E_0(B)]/V_0(B)$ provides a standardised test statistic for calibration, (Hilden et al, 1978; Spiegelhalter, 1986b).

Our essential idea is that when a system interrogates a user, it should already be guessing (probabilistically) what the response will be. If the responses are consistently surprising, either for a series of questions concerning a specific individual, or for a specific question over a series of individuals, then some close inspection of the knowledge-base is in order.

2.10 Explanation of conclusions

In 'shallow' graphs, such as a set of symptoms S_i conditionally independent given a disease D, it is straightforward to use the 'weight of evidence' $\log_e[p(s_i|d)/p(s_i|\bar{d})]$ as a simple summary of the support for or against a diagnosis, and Spiegelhalter and Knill-Jones (1984) and Spiegelhalter (1985) show how this can be extended to cope with dependent symptoms and multiple diseases. Within a more complex graphical structure, it appears the ratios $p^*(\ldots)/p(\ldots)$ propagated through the intersections of the cliques may form a natural explanation of the source of the evidence for revised beliefs, although it remains to be investigated how this can best be implemented. Pearl (1985) makes further suggestions concerning explanation in probabilistic models.

3. 'EXCHANGEABILITY' AND DOUBT ABOUT PROBABILITIES

Cheeseman (1985) introduces an example concerning a subject who, when confronted with a black box that puts out a string of decimal digits, is asked his opinion concerning the probability of the next digit being a 7. Cheeseman states that our initial 'indifference' probability may be .1, but we are clearly not too sure about this in the sense that we might expect this assessment to change as evidence accumulates. However, after observing 10000 digits of which 1000 are 7's, our probability is still .1 but now has small 'standard deviation', in his terms. We present this problem within the framework developed in the previous section.

Let us denote the prospective sequence of digits by $X_1,\ldots,X_{10000},\ldots$. Since our belief in these digits is unaffected by the order in which they are to be observed, we say the sequence is 'exchangeable' (de Finetti, 1974). De Finetti's representation theorem states that we can consider the X_i's as independent variables conditional on some true, unknown probability of a 7 turning up; we shall denote this probability y. Within our framework, y is a realisation of an auxillary, unobservable, node Y in the directed graph in Figure 7, where $p(X_i = 7|y) = y$; we assume y can take on any value between 0 and 1. Since no two unjoined nodes are pointing at a common successor, the arrows may be dropped and the joint distribution expressed in terms of cliques $[X_1,Y],[X_2,Y]$, ...

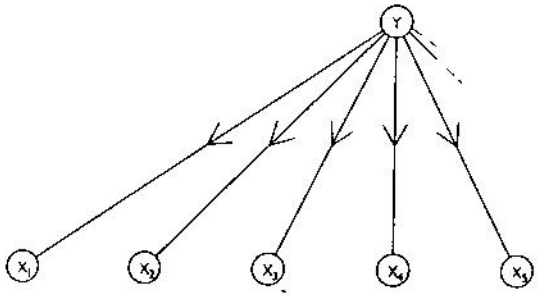

FIGURE 7.

An exchangeable sequence $X_1,\ldots$, expressed in a graphical form

Our initial doubt about $p(X_i = 7)$ is expressed as a distribution over y with mean .1 - a suitable, fairly 'vague' choice is $p(y) = y^{-8/9}/9$ which has standard deviation .21, and is drawn in Figure 8.

After observing 10 7's out of 100 digits (a considerably less extreme example than Cheeseman's) use of the updating techniques of Section 2.4, equivalent to standard Bayesian theory (see, for example, de Groot, 1970), shows that our new belief $p^*(y)$ is a density proportional to $y^{10 - 8/9}(1-y)^{90}$, which has mean .1 and standard deviation .03 and is also shown in Figure 8.

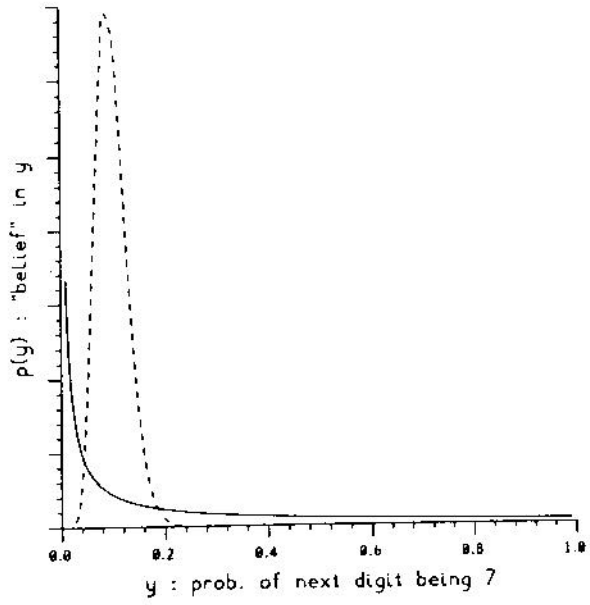

FIGURE 8

Distribution representing our belief in the probability y, before sampling (——), and after observing 10 7's in 100 digits (----). Both distributions have mean .1

The doubt concerning the proportion has considerably reduced, but in spite of our increased confidence, if we had to make a bet on the next digit, we would actually accept the same odds as we would have done before we had observed 100 digits. De Finetti (1974) shows that our belief in the true underlying probability in an exchangeable sequence is simply our belief in what the eventual proportion of 7's will turn out to be were we to continue sampling. Thus, in this particular instance of exchangeable observations, the assessment of 'sensitivity' to unobserved data described in Section 2.6, is precisely equivalent to the assessment of imprecision of Section 2.7, since the unknown quantities on which each assessment conditions coincide in the auxiliary node Y.

4. DISCUSSION: WHEN IS PROBABILITY APPROPRIATE ?

In the preface to his classic treatise, de Finetti (1974) states that "PROBABILITY DOES NOT EXIST". In saying this, he rejects any idea of there being a 'true' objective probability of an event, and argues - as does Cheeseman (1985) - that probabilities may change from person to person, from time to time, as information accumulates; all probabilities are conditional on the evidence deemed to be relevant by the assessor. How, then, can we say what probabilities are? On one level, they are numbers which obey certain basic mathematical laws that ensure mutual consistency over various combinations of events. However, these laws are not axioms, but may be

derived from intuitive behavioural criteria that appear particularly relevant to expert system construction and evaluation.

Lindley (1982) has generalised de Finetti's original specification of the operational justification for probability. Briefly, Lindley describes a situation in which a person has to describe his uncertainty concerning an event E by a real number x. Conditional on a second event F occurring, the person will receive a penalty f(x,E)F, and this penalty is additive over repeated sets of events and assessments, say $x_1, \ldots, x_n$. Then if the person will not choose 'inadmissible' values for x_i, such that there exist alternative choices y_i with guaranteed smaller penalty no matter what events occur, then there exists a transform of x, specified by the form of the penalty f, which obeys the laws of probability. Lindley goes on to argue that confidence intervals, significance tests, 'possibility' measures (Zadeh, 1983) upper and lower probabilities, and belief functions (Shafer, 1986), are 'inadmissible'.

In the published discussion of Lindley's paper, it is not surprising that Shafer and Zadeh, amongst others, criticize Lindley for being too restrictive in his concept of uncertainty. However, little is mentioned of a crucial feature of Lindley's argument: that the measure of uncertainty is applied only to events, that is, propositions which are - at least potentially - verifiable. The theoretical argument for probability would therefore appear to be restricted to situations in which an expert system is to be evaluated in terms of its explicit prediction concerning events that are potentially observable given further investigation. Early diagnostic and classification systems such as MYCIN and PROSPECTOR fit into this category, but it could be argued that in systems concerned, for example, with planning or critiquing proposed courses of action, 'uncertainty' is of a different type.

Two contexts in which there is 'uncertainty' concerning apparently non-verifiable statements occur frequently in the AI literature. The first concerns a statement that is imprecise, such as there being a linguistic qualifier of the extent to which a proposition is true, ("John is fairly tall"). The fuzzy set approach is often argued to be appropriate in assessing the degree to which a particular case-in-hand fulfils a loosely-defined concept. The second context concerns 'uncertainty' about the reasonableness of an action or a conclusion; for example, Cohen (1985, p.52) states that "one's certainty in a result should depend on what the result is wanted for", and goes on to a non-numerical theory of 'endorsements'.

Of course, probability theory can be given a place even within the contexts outlined above, by contriving some suitable decidable proposition (Giles, 1982). For example, a 'test' of an imprecise statement might be put in terms of the probability that a random person, when asked to answer 'yes' or 'no' to whether John was fairly tall, would answer 'yes'. Similarly, the reasonableness of a conclusion could be expressed in terms of the probability of whether an expert in the field would or would not draw that conclusion given the available evidence.

The important feature remains, however, that the procedure by which an expert system is to be judged affects the means by which uncertainty is handled. This has consequences in the design of systems, since 'uncertain rules' that are statistically testable should be carefully distinguished from those that are not objectively verifiable, but are only subject to peer review.

It might be said that it is unreasonable to evaluate an expert system solely on the basis of numerical predictions, and that clarity of explanation, ease of assessment, ability to learn, and transparency of the knowledge

representation are also desirable aims. In this paper we hope to have shown that probabilistic reasoning, in addition to being **theoretically** necessary in any predictive context, is also **practically** justified in relation to the above criteria, as well as providing efficient evidence propagation, operational interpretation of outputs, and systematic criticism of performance.

ACKNOWLEDGEMENTS

I am indebted to Steffen Lauritzen, Peter Cheeseman and Wally Gilks for their useful discussions.

REFERENCES

Besag, J (1986) On the statistical analysis of dirty pictures. J.Royal Statistical Society B (to appear)

Blalock, H M (1971) Causal Models in the Social Sciences. Macmillan: London

Cheeseman, P (1983) A method of computing generalised Bayesian probability values for expert systems. In Proceedings of 8th International Joint Conference on Artificial Intelligence, Karlsruhe, West Germany, p 198-202.

Cheeseman, P (1985) In defense of probability. IJCAI-85, 1002-1009

Cohen, P R (1985) Heuristic Reasoning about Uncertainty : an Artificial Intelligence Approach, Pitmans: Boston.

Cooper, G F (1984) NESTOR: a computer-based medical diagnostic and that integrates causal and probabilistic knowledge. Report HPP-84-48, Stanford University.

Darroch, J N, Lauritzen, S L and Speed, T P (1980) Markov fields and log-linear models for contingency tables. Annals of Statistics, 8, 522-539.

De Groot (1970) Optimal Statistical Decisions. New York: McGraw-Hill.

Dodson, D C and Rector, A L (1985) Importance-driven distributed control of diagnostic inference. In Research and Development in Expert Systems (Bramer, MA, ed.) Cambridge University Press : Cambridge.

Duda, R O, Hart, P E and Nilsson N J (1976) Subjective Bayesian methods for rule-based inference systems. Proc. AFIPS Nat.Comp.Conf., 47, 1075-82.

Edwards, D and Kreiner, S (1983) The analysis of contingency tables by graphical models. Biometrika, 70, 553-65.

de Finetti, B (1974) Theory of Probability Vols. 1 and 2. New York: Wiley

Fisher, R A (1957) The underworld of probability. Sankhya, 18, 201-210.

Geman S (1984) Stochastic relaxation methods for image restoration and expert systems. In Automated Image Analysis: Theory and Experiments, (D B Cooper, R L Launer, D E McClure, eds) New York: Academic Press.

Geman S and Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Machine Intell., 6, 721-741.

Giles, R (1982) Semantics for fuzzy reasoning.
Int. J.Man-Machine Studies, 17, 401-415.

Good, I J (1965) The Estimation of Probabilities, MIT Press.

Hilden J, Habbema J D F and Bjerregaard B (1978). The measurement of performance in probabilistic diagnosis. III - Methods based on continuous functions of the diagnostic probabilities. Methods of Information in Medicine, 17, 238246.

Isham, V (1981) An introduction to spatial point processes and Markov random fields, International Statistical Review, 49, 21-43.

Kiiveri H, Speed T P, Carlin J B (1984) Recursive causal models. J. Austral. Math. Soc. (Series A), 36, 30-52.

Kim, J H and Pearl, J (1983) A computational model for causal and diagnostic reasoning in inference systems. Proceedings 8th International Joint Conference on Artificial Intelligence, Karlsruhe, West Germany : p190-193.

Lauritzen, S L (1982) Lectures on Contingency Tables, 2nd Ed. University of Aalborg Press.

Lemmer, J F and Barth, S W (1982) Efficient minimum information updating for Bayesian inferencing in expert systems. Proc.2nd.Conf. on AI, Pittsburgh, pp 424-427.

Lemmer, J F (1983) Generalised Bayesian updating of incompletely specified distributions. Large Scale Systems, 5, 51-68.

Lindley, D V (1982) Scoring rules and the inevitability of probability. International Statistical Review, 50, 1-26.

Pearl J (1985) How to do with probabilities what people say you can't. Technical Report CSD-R-49, UCLA.

Pearl J (1986) A constraint-propagation approach to probabilistic reasoning. (This volume)

Pople, H E (1982) Heuristic methods for imposing structure on ill structured problems : the structuring of medical diagnosis. In Artificial Intelligence in Medicine (P. Szolovits, ed). pp 119-185. Colorado: Westview Press.

Shachter, R D (1986) Intelligent probabilistic inference. (This volume).

Shafer, G (1986) Probability judgement in artificial intelligence. (This volume)

Spiegelhalter D J and Knill-Jones R P (1984) Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology (with discussion). J. Royal Statistical Soc., B, 147, 35-77.

Spiegelhalter, D J (1985) Statistical methodology for evaluating gastrointestinal symptoms. Clinics in Gastroenterology, 14, 489-515.

Spiegelhalter, D J (1986a) A statistical view of uncertainty in expert systems. In Artificial Intelligence and Statistics (W Gale, ed). Addison-Wesley.

Spiegelhalter, D J (1986b) Probabilistic prediction in patient management and clinical trials. Statistics in Medicine (to appear)

Tarjan, R E and Yannakakis, M (1984) Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. SIAM J. Comput. 13,566-579.

van Melle, W, Scott A C, Bennett J S, Peairs MAS (1981) The EMYCIN Manual. Rep. HPP-81-16. Computer Science Dept, Stanford University, California.

Weiss, S M, Kulikowski, C A, Amarel, S and Safir, A (1978). A model-based method for computer-aided medical decision-making. Artificial Intelligence, 11, 145-172.

Wermuth N and Lauritzen S L (1983) Graphical and recursive models for contingency tables. Biometrika, 70, 537-52.

Winkler R L (1969) Scoring rules and evaluation of probability assessors. J. American Statist. Assoc, 64, 1073-1078.

Wold, H D A (1954) Causality and econometrics. Econometrica, 28, 443-63.

Wright, S (1934) The method of path coefficients. Ann.Math.Statist., 5, 161-215.

Zadeh, L A (1983) The role of fuzzy logic in the management of uncertainty in expert systems. Fuzzy Sets and Systems, 11, 199-228.