Chapter 4

Model of Inexact Reasoning in Medicine†

4.1 Introduction

Efforts to develop techniques for modeling clinical decision making have had a dual motivation. Not only has their potential clinical significance been apparent, but the design of such programs has required an analytical approach to medical reasoning that has in turn led to a distillation of decision criteria that in some cases had never been explicitly stated before. It is a fascinating and educational process for experts to reflect on the reasoning steps that they have always used when providing clinical consultations.

As discussed in § 1.3, several programs have successfully modeled the diagnostic process [Gorry, 1968a, 1973; Warner, 1964]. Each of these examples has relied upon statistical decision theory as reflected in the use of Bayes' Theorem for manipulation of conditional probabilities. Use of the theorem, however, requires either large amounts of valid background data or numerous approximations and assumptions. The success of Gorry and Barnett's early work [Gorry, 1968a], and a similar study by Warner *et al.* using the same data [Warner, 1964], depended to a large extent upon the availability of good data regarding several individuals with congenital heart disease. Gorry *et al.* [Gorry, 1973b] have had similar access to data relating the symptoms and signs of acute renal failure to the various potential etiologies.

†Much of the material in this chapter has appeared in an article in *Mathematical Biosciences* [Shortliffe, 1975a]. That paper was co-authored with Dr. Bruce Buchanan who contributed substantially to the development of the model.

Although conditional probability provides useful results in areas of medical decision making such as those I have mentioned, vast portions of medical experience suffer from so little data and so much imperfect knowledge that a rigorous probabilistic analysis, the ideal standard by which to judge the rationality of a physician's decisions, is not possible. It is nevertheless instructive to examine models for the less formal aspects of decision making. Physicians seem to use an ill-defined mechanism for reaching decisions despite a lack of formal knowledge regarding the interrelationships of all the variables that they are considering. This mechanism is often adequate, in well-trained or experienced individuals, to lead to sound conclusions on the basis of a limited set of observations. (Intuition may also lead to unsound conclusions, as noted by Schwartz *et al.* [Schwartz, 1973].)

These intuitive and inexact aspects of medical reasoning are reflected in an argument expounded by Helmer and Rescher [Helmer, 1960]. They assert that the traditional concept of "exact" versus "inexact" science, with the social sciences accounting for the second class, has relied upon a false distinction usually reflecting the presence or absence of mathematical notation. They point out that only a small portion of natural science can be termed exact—areas such as pure mathematics and subfields of physics in which some of the exactness "has even been put to the ultimate test of formal axiomatization." In several areas of applied natural science, on the other hand, decisions, predictions, and explanations are only made after exact procedures are mingled with unformalized expertise. Society's general awareness regarding these observations is reflected in the common references to the "artistic" components in the "science of medicine."

This chapter examines the nature of such nonprobabilistic and unformalized reasoning processes, considers their relationship to formal probability theory, and proposes a model whereby such incomplete "artistic" knowledge might be quantified. We have developed this model of inexact reasoning in response to MYCIN's needs; i.e., the goal has been to permit the opinion of experts to become more generally available to nonexperts. The model is, in effect, an approximation to conditional probability. Although conceived with MYCIN's problem area in mind, it is potentially applicable to any domain in which real world knowledge must be com-

bined with expertise before an informed opinion can be obtained to explain observations or to suggest a course of action.

The presentation begins with a brief discussion of Bayes' Theorem as it has been utilized by other workers in this field. The theorem serves as a focus for discussion of the clinical problems that we would like to solve by using computer models. The potential applicability of the proposed decision model is then introduced in light of MYCIN's rule-based design. Once the problem has been defined in this fashion, the criteria and numerical characteristics of our quantification scheme are proposed. The chapter concludes with a discussion of how the model is being used by MYCIN when it offers opinions to physicians regarding antimicrobial therapy selection.

4.2 Problem Formulation

The medical diagnostic problem can be viewed as the assignment of probabilities to specific diagnoses after analyzing all relevant data. If the sum of the relevant data (or evidence) is represented by E, and D_i is the ith diagnosis (or "disease") under consideration, then $P(D_i|E)$ is the conditional probability that the patient has disease i in light of the evidence E. Diagnostic programs have traditionally sought to find a set of evidence that allows $P(D_i|E)$ to exceed some threshold, say .95, for one of the possible diagnoses. Under these circumstances the second ranked diagnosis is sufficiently less likely ($<.05$) that the user is content to accept disease i as the diagnosis requiring therapeutic attention. (Several programs have also included utility considerations in their analyses. For example, an unlikely but lethal disease that responds well to treatment may merit therapeutic attention because $P(D_i|E)$ is nonzero, even though very small.)

Bayes' Theorem is useful in these applications because it allows $P(D_i|E)$ to be calculated from the component conditional probabilities:

$$P(D_i|E) = \frac{P(D_i)P(E|D_i)}{\sum_{j=1}^{n} P(D_j)P(E|D_j)}$$

In this representation of the theorem, D_i is one of n disjoint diag-

noses. $P(D_i)$ is simply the *a priori* probability that the patient has disease i before any evidence has been gathered. $P(E|D_i)$ is the probability that a patient will have the complex of symptoms and signs represented by E, given that he has disease D_i.

I have so far ignored the complex problem of identifying the "relevant" data that should be gathered in order to diagnose the patient's disease. Evidence is actually acquired piece-by-piece, the necessary additional data being identified on the basis of the likely diagnosis at any given time. Diagnostic programs that mimic the process of analyzing evidence incrementally often use a modified version of Bayes' Theorem that is appropriate for sequential diagnosis [Gorry, 1968a]:

Let E_1 be the set of all observations to date, and S_1 be some new piece of data. Furthermore, let E be the new set of observations once S_1 has been added to E_1. Then

$$P(D_i|E) = \frac{P(S_1|D_i \& E_1)P(D_i|E_1)}{\sum_{j=1}^{n} P(S_1|D_j \& E_1)P(D_j|E_1)}$$

The successful programs that use Bayes' Theorem in this form required huge amounts of statistical data, not merely $P(D_i|S_k)$ for each of the pieces of data S_k in E, but also the interrelationships of the S_k within each disease D_j. For example, although S_1 and S_2 are independent over all diseases, it may be true that S_1 and S_2 are closely linked for patients with disease D_i. Thus relationships must be known within *each* D_j; overall relationships are not sufficient. The congenital heart disease programs [Gorry, 1968a; Warner, 1964] were able to acquire all the necessary conditional probabilities from a survey of several hundred patients with confirmed diagnoses and thus had nonjudgmental data on which to base their Bayesian analyses.

Edwards has summarized the kinds of problems that can arise when an attempt is made to gather the kinds of data needed for rigorous analysis [W. Edwards, 1972]:

... My friends who are expert about medical records tell me that to attempt to dig out from even the most sophisticated hospital's records the frequency of association between any particular symptom and any particular diagnosis is next

to impossible—and when I raise the question of complexes of symptoms, they stop speaking to me. For another thing, doctors keep telling me that diseases change, that this year's flu is different from last year's flu, so that symptom-disease records extending far back in time are of very limited usefulness. Moreover, the observation of symptoms is well-supplied with error, and the diagnosis of diseases is even more so; both kinds of errors will ordinarily be frozen permanently into symptom-disease statistics. Finally, even if diseases didn't change, doctors would. The usefulness of disease categories is so much a function of available treatments that these categories themselves change as treatments change—a fact hard to incorporate into symptom-disease statistics.

All these arguments against symptom-disease statistics are perhaps somewhat overstated. Where such statistics can be obtained and believed, obviously they should be used. But I argue that usually they cannot be obtained, and even in those instances where they have been obtained, they may not deserve belief.

An alternative to exhaustive data collection is to use the knowledge that an expert has about the disease—partly based upon experience and partly on general principles—to reason about diagnoses. In the case of this judgmental knowledge acquired from experts, the conditional probabilities and their complex interrelationships cannot be acquired in an exhaustive manner. Opinions can be sought and attempts made to quantify them, but the extent to which the resulting numbers can be manipulated as probabilities is not clear. We shall explain this last point more fully as we proceed. First, let us examine some of the reasons that it might be desirable to construct a model that allows us to avoid the inherent problems of explicitly relating the conditional probabilities to one another.

As was pointed out in § 3.2, a conditional probability statement is, in effect, a statement of a decision criterion or rule. For example, the expression $P(D_i|S_k)=X$ can be read as a statement that there is a $100X$ % chance that a patient observed to have symptom S_k has disease D_i. Stated in rule form:

IF: THE PATIENT HAS SIGN OR SYMPTOM Sk
THEN: CONCLUDE THAT HE HAS DISEASE Di WITH PROBABILITY X

I shall often refer to statements of conditional probability as decision rules or decision criteria in the diagnostic context. The value of X for such rules may not be obvious (e.g., "y strongly suggests that z is true" is difficult to quantify), but an expert may be able to offer an

estimate of this number based upon clinical experience and general knowledge, even when such numbers are not readily available otherwise.

A large set of such rules obtained from textbooks and experts would clearly contain a large amount of medical knowledge. It is conceivable that a computer program could be designed to consider all such general rules and to generate a final probability of each D_i based upon data regarding a specific patient. Bayes' Theorem would only be appropriate for such a program, however, if values for $P(S_1 \mid D_i)$ and $P(S_1 \mid D_i \& S_2)$ could be obtained. As has been noted, these requirements become unworkable, even if the subjective probabilities of experts are used, in cases where a large number of diagnoses (hypotheses) must be considered. The first would require acquiring the inverse of every rule, and the second requires obtaining explicit statements regarding the interrelationships of all rules in the system.

In short, we would like to devise an approximate method that allows us to compute a value for $P(D_i \mid E)$ solely in terms of $P(D_i \mid S_k)$, where E is the composite of all the observed S_k (see § 4.5 and 4.6). Such a technique will not be exact, but since the conditional probabilities reflect judgmental (and thus highly subjective) knowledge, a rigorous application of Bayes' Theorem will not necessarily produce accurate cumulative probabilities either. Instead we look for ways to handle decision rules as discrete packets of knowledge and for a quantification scheme that permits accumulation of evidence in a manner that adequately reflects the reasoning process of an expert using the same or similar rules.

4.3 Mycin's Rule-Based Approach

As has been discussed, MYCIN's principle task is to determine the likely identity of pathogens in patients with infections and to assist in the selection of a therapeutic regimen appropriate for opposing the organisms under consideration. In Chapter 3, we explained how MYCIN models the consultation process, utilizing judgmental knowledge acquired from experts in conjunction with certain statistical data that are available from the clinical microbiology laboratory and from patient records. MYCIN's decision rules are similar in form to those just introduced in § 4.2.

It is useful to consider the advantages provided by a rule-based system for computer use of judgmental knowledge. It should be emphasized that we see these advantages as being sufficiently strong in certain environments that we have devised an alternative and approximate approach that parallels the results available from using Bayes' Theorem. I do not argue against the use of Bayes' theory in those medical environments in which sufficient data are available to permit adequate use of the theorem.

The advantages of rule-based systems for diagnostic consultations include:

(1) the use of general knowledge (from textbooks or experts) for consideration of a specific patient; even well-indexed books may be difficult for a nonexpert to use when considering a patient whose problem is not quite the same as those of patients discussed in the text;
(2) the use of judgmental knowledge for consideration of very small classes of patients with rare diseases about which good statistical data are not available;
(3) ease of modification; since the rules are not explicitly related to one another and there need be no prestructured decision tree for such a system, rule modifications and the addition of new rules need not require complex considerations regarding interactions with the remainder of the system's knowledge;
(4) facilitated search for potential inconsistencies and contradictions in the knowledge base; criteria stored explicitly in packets such as rules can be searched and compared without major difficulty;
(5) straightforward mechanisms for explaining decisions to a user by identifying and communicating the relevant rules;
(6) an augmented instructional capability; a system user may be educated regarding system knowledge in a selective fashion, i.e., only those portions of the decision process that puzzle him need be examined.

One of MYCIN's rules, which I shall use for illustrative purposes throughout this chapter, is the following:

IF: 1) THE STAIN OF THE ORGANISM IS GRAM POSITIVE, AND
2) THE MORPHOLOGY OF THE ORGANISM IS COCCUS, AND
3) THE GROWTH CONFORMATION OF THE ORGANISM IS CHAINS
THEN: THERE IS SUGGESTIVE EVIDENCE (.7) THAT THE IDENTITY OF THE ORGANISM IS STREPTOCOCCUS

This rule was acquired from an expert in infectious disease therapy and reflects his belief that gram positive cocci growing in chains are apt to be streptococci. When asked to weight his belief in this conclusion, he indicated a 70% belief that the conclusion was valid. In the English language version of the rules, the program uses phrases such as "suggestive evidence" as in the above example. However, the numbers following these terms, indicating degrees of certainty, are all that is used in the model. The English phrases are not given by the expert and then quantified; they are, in effect, "canned-phrases" used only for translating rules into English representations. The prompt used for acquiring the certainty measure from the expert is: "On a scale of 1 to 10, how much certainty do you affix to this conclusion?"

Translating to the notation of conditional probability, the rule above at first seems to say $P(H_1 \mid S_1 \& S_2 \& S_3)=.7$ where H_1 is the hypothesis that the organism is a streptococcus, S_1 the observation that the organism is gram positive, S_2 that it is a coccus, and S_3 that it grows in chains. Questioning of the expert gradually reveals, however, that despite the apparent similarity to a statement regarding a conditional probability, the number .7 differs significantly from a probability. The expert may well agree that $P(H_1 \mid S_1 \& S_2 \& S_3)=.7$, but he becomes uneasy when he attempts to follow the logical conclusion that therefore $P(\text{not}.H_1 \mid S_1 \& S_2 \& S_3)=.3$. The three observations are evidence (to degree .7) *in favor* of the conclusion that the organism is a streptococcus and should not be construed as evidence (to degree .3) *against* streptococcus. I shall refer to this problem as Paradox 1 and return to it later in the exposition after the interpretation of the .7 in the rule above has been introduced.

It may at first seem tempting to conclude that the expert is irrational if he is unwilling to follow the implications of his probabilistic statements to their logical conclusions. Another interpretation, however, is that the numbers he has given should not be construed as probabilities at all, that they are judgmental measures that reflect a level of belief. The nature of such numbers, and the very existence of such concepts, have interested philosophers of science for the last half century. Some of these philosophical issues are briefly discussed in § 4.4. I then proceed to a detailed presentation of the proposed quantitative model. In the last section of this

chapter, I shall show how the model has been implemented for ongoing use by the MYCIN program.

4.4 Theoretical Background

Although probability is a familiar concept defined axiomatically in any introductory statistics book [Parzen, 1960], the *P*-function has been subjected to a variety of interpretations [Swinburne, 1973; Harré, 1970; Ramsey, 1931; Savage, 1954; deFinetti, 1972; Keynes, 1921; Carnap, 1950]. I shall not describe all of these because, as has been observed, imperfect knowledge and the dependence of decisions on individual judgments make the *P*-function no longer seem entirely appropriate for modeling many of the decision processes in medical diagnosis.

Carnap [Carnap, 1950] and Hempel [Hempel, 1945] discuss an interpretation of probability known as confirmation. Carnap distinguishes confirmation from the traditional *P*-function, defining the former as the degree to which an hypothesis is supported by an evidence statement. Thus, it should be noted that the term confirmation does not indicate that an hypothesis is proven but rather that an observation lends credence to it. The measure of support is commonly represented by the notation $C[h,e]$, i.e., the degree of confirmation of the hypothesis h based upon the observation e.

Quantifying confirmation and then manipulating the numbers as though they are probabilities quickly leads to apparent inconsistencies or paradoxes [Carnap, 1950; Hempel, 1945; Barker, 1957; Salmon, 1973, 1966]. Carl Hempel [Hempel, 1945] presented his famous Paradox of the Ravens early in his discussion of the logic of confirmation. Let h_1 be the statement that "All ravens are black" and h_2 the statement that "All nonblack things are nonravens." Clearly h_1 is logically equivalent to h_2. If one were to draw an analogy with conditional probability, it might at first seem valid, therefore, to assert that $C[h_1,e]=C[h_2,e]$ for all e. However, it appears counter-intuitive to state that the observation of a green vase supports h_1 even though the observation does seem to support h_2. $C[h,e]$ is therefore different from $P(h|e)$ for it seems somehow wrong that the observation of a vase could logically support an

assertion about ravens. A re-examination of this paradox in light of our proposed quantification scheme is included as an appendix to this chapter (Appendix 4.A).

In their analyses of confirmation, several authors [Harré, 1970; Carnap, 1950; Hempel, 1945; Barker, 1957; Salmon, 1973, 1966] note that $C[h,e]$ does not equal $1-C[\text{not}.h,e]$, an observation reminiscent of our Paradox 1 from § 4.3. Furthermore, they recognize the need for an independently introduced disconfirmation function because, as Harré puts it [Harré, 1970], "to confirm something to ever so slight a degree is not to disconfirm it at all, since the favourable evidence for some hypothesis gives no support whatever to the contrary supposition in many cases."

The inadequacies of probability in the analysis of real-world problems have led to a variety of alternate approaches. These include the theory of "fuzzy sets" [Zadeh, 1965; Goguen, 1968], the theory of "choice" [Tversky, 1972; Luce, 1965], and the logic of "surprise" [Shackle, 1952, 1955]. However, the theory of confirmation seems to parallel more closely the kind of decision task involved in medical diagnosis. We have therefore sought to develop a quantification scheme that reflects the observations of philosophers who have dealt with the logic of confirmation. However, the scheme I propose meets desiderata derived intuitively from the problem at hand rather than from a formal list of acceptability criteria. Such criteria are proposed by several authors such as Carnap [Carnap, 1950], Swinburne [Swinburne, 1970], Salmon [Salmon, 1966], and Törnebohm, [Törnebohm, 1966]. Although our model was not developed with any such list of criteria as guidance, I shall show (§ 4.5 and 4.6) that the technique we propose satisfies Tornebohm's criteria in light of the approximation mechanisms that we have introduced for the combination of incrementally acquired evidence.

4.5 Proposed Model of Evidential Strength

This section introduces our quantification scheme for modeling inexact medical reasoning. It begins by defining the notation that we use and by describing the terminology. A formal definition of the quantification function will then be presented. The remainder of the section discusses the characteristics of the defined functions. It closes

with consideration of the model when it is compared to Törnebohm's criteria for acceptability of a quantification technique regarding evidential strength [Törnebohm, 1966].

Although the proposed model has several similarities to a confirmation function such as those mentioned above, I shall introduce new terms for the measurement of evidential strength. This convention will allow me to clarify from the outset that I seek only to devise a system that captures enough of the flavor of confirmation theory that it can be used for accomplishing our computer-based task. We have chosen "Belief" and "Disbelief" as our units of measurement, but these terms should not be confused with their formalisms from epistemology. The need for two measures was introduced above in our discussion of a disconfirmation measure as an adjunct to a measure for degree of confirmation. The notation will be as follows:

(1) MB[h,e]=X means "The measure of increased Belief in the hypothesis h, based on the evidence e, is X"
(2) MD[h,e]=Y means "The measure of increased Disbelief in the hypothesis h, based on the evidence e, is Y"

The evidence e need not be an observed event, but may be a hypothesis (itself subject to confirmation). Thus, I may write MB[h_1,h_2] to indicate the measure of increased Belief in the hypothesis h_1 given that the hypothesis h_2 is true. Similarly MD[h_1,h_2] is the measure of increased Disbelief in hypothesis h_1 if hypothesis h_2 is true.

To illustrate in the context of the sample rule from MYCIN, consider e = "The organism is a gram positive coccus growing in chains" and h = "The organism is a streptococcus." Then, MB[h,e]=.7 according to the sample rule given us by the expert. The relationship of the number .7 to probability will be explained as I proceed. For now let me simply state that the number .7 reflects the extent to which the expert's Belief that h is true is increased by the knowledge that e is true. On the other hand, MD[h,e]=0 for this example, i.e., the expert has no reason to increase his Disbelief in h on the basis of e.

In accordance with subjective probability theory, it may be argued that the expert's personal probability $P(h)$ reflects his Belief in h at any given time. Thus $1-P(h)$ can be viewed as an

estimate of the expert's Disbelief regarding the truth of h. If $P(h|e)$ is greater than $P(h)$, the observation of e increases the expert's Belief in h while decreasing his Disbelief regarding the truth of h. In fact, the proportionate decrease in Disbelief is given by the ratio:

$$\frac{P(h|e)-P(h)}{1-P(h)}$$

This ratio is called the measure of increased Belief in h resulting from the observation of e, i.e., MB[h,e].

Suppose, on the other hand, that $P(h|e)$ were less than $P(h)$. Then the observation of e would decrease the expert's Belief in h while increasing his Disbelief regarding the truth of h. The proportionate decrease in Belief is in this case given by the ratio:

$$\frac{P(h)-P(h|e)}{P(h)}.$$

We call this ratio the measure of increased Disbelief in h resulting from the observation of e, i.e., MD[h,e]. Törnebohm suggests a similar measure of evidential strength [Törnebohm, 1966], but uses $C(H)$ instead of $P(H)$, where $C(H)$ is the amount of information contained in H.

To summarize these results in words, we consider the measure of increased Belief, MB[h,e], to be the proportionate decrease in Disbelief regarding the hypothesis h that results from the observation e. Similarly, the measure of increased Disbelief, MD[h,e], is the proportionate decrease in Belief regarding the hypothesis h that results from the observation e, where Belief is estimated by $P(h)$ at any given time and Disbelief is estimated by $1-P(h)$. These definitions correspond closely to the intuitive concepts of confirmation and disconfirmation that we have discussed above. Note that since one piece of evidence cannot both favor and disfavor a single hypothesis, when MB[h,e]>0, $MD[h,e]=0$ and when MD[h,e]>0, MB[h,e]$=0$. Furthermore, when $P(h|e)=P(h)$ the evidence is independent of the hypothesis (neither confirms nor disconfirms) and MB[h,e]$=$ MD[h,e]$=0$.

The above definitions may now be specified formally in terms of conditional and *a priori* probabilities:

$$\mathrm{MB}[h,e] = \begin{cases} 1 & \text{if } P(h)=1, \\ \dfrac{\max[P(h|e),P(h)]-P(h)}{\max[1,0]-P(h)} & \text{otherwise,} \end{cases}$$

$$\mathrm{MD}[h,e] = \begin{cases} 1 & \text{if } P(h)=0, \\ \dfrac{\min[P(h|e),P(h)]-P(h)}{\min[1,0]-P(h)} & \text{otherwise.} \end{cases}$$

Note that here $P(h)$ is used to denote *a priori* probabilities. More correctly they might be written as $P(h|0)$, i.e., the probability of h on no evidence. Examination of these expressions will reveal that they are identical to the definitions introduced above. The formal definition is introduced, however, to demonstrate the symmetry between the two measures. In addition, we define a third measure, termed a certainty factor (CF) that combines the MB and MD in accordance with the following definition:

$$\mathrm{CF}[h,e] = \mathrm{MB}[h,e] - \mathrm{MD}[h,e]$$

The certainty factor thus is an artifact for combining degrees of Belief and Disbelief into a single number. Such a number is needed in order to facilitate comparisons of the evidential strength of competing hypotheses. The use of this composite number will be described below in greater detail. The following observations help to clarify the characteristics of the three measures that I have defined (MB, MD, CF):

Characteristics of Belief Measures

(1) Range of degrees:
   (a) $0 \leqslant \mathrm{MB}[h,e] \leqslant 1$.
   (b) $0 \leqslant \mathrm{MD}[h,e] \leqslant 1$.
   (c) $-1 \leqslant \mathrm{CF}[h,e] \leqslant +1$.

(2) Evidential strength and mutually exclusive hypotheses†:
   If h is shown to be certain $[P(h|e)=1]$:

†There is a special case of characteristic (2) that should be mentioned. This is the case of logical truth or falsity where $P(h|e)=1$ or $P(h|e)=0$, regardless of e. Popper has also suggested a quantification scheme for confirmation [Popper, 1959] in which he uses $-1 \leqslant C[h,e] \leqslant +1$, defining his limits as:

(a) $MB[h,e]=[1-P(h)]/[1-P(h)]=1$.
(b) $MD[h,e]=0$.
(c) $CF[h,e]=1$.

If the negation of h is shown to be certain $[P(\text{not}.h|e)=1]$:

(a) $MB[h,e]=0$.
(b) $MD[h,e]=[0-P(h)]/[0-P(h)]=1$.
(c) $CF[h,e]=-1$.

Note that this gives $MB[\text{not}.h,e]=1$ if and only if $MD[h,e]=1$ in accordance with the definitions of MB and MD above. Furthermore, the number 1 represents absolute Belief (or Disbelief) for MB (or MD). Thus if $MB[h_1,e]=1$ and h_1 and h_2 are mutually exclusive, $MD[h_2,e]=1$.

(3) Lack of evidence:
(a) $MB[h,e]=0$ if h is not confirmed by e (i.e., e and h are independent or e disconfirms h).
(b) $MD[h,e]=0$ if h is not disconfirmed by e (i.e., e and h are independent or e confirms h).
(c) $CF[h,e]=0$ if e neither confirms nor disconfirms h (i.e., e and h are independent).

We are now in a position to examine Paradox 1 (§ 4.3), the expert's concern that although evidence may support a hypothesis with degree X, it does not support the negation of the hypothesis with degree $1-X$. In terms of our proposed model, this reduces to the assertion that, when e confirms h:

$$CF[h,e]+CF[\text{not}.h,e]\neq 1.$$

$$-1=C[\text{not}.h,h]\leqslant C[h,e]\leqslant C[h,h]=+1.$$

This proposal led one observer [Harré, 1970] to assert that Popper's numbering scheme "obliges one to identify the truth of a self-contradiction with the falsity of a disconfirmed general hypothesis and the truth of a tautology with the confirmation of a confirmed existential hypothesis, both of which are not only question begging but absurd." As I shall demonstrate in § 4.6, we avoid Popper's problem by introducing mechanisms for approaching certainty asymptotically as items of confirmatory evidence are discovered.

This intuitive impression is verified by the following analysis:

$$\begin{aligned}\mathrm{CF}[\text{not.}h,e] &= \mathrm{MB}[\text{not.}h,e]-\mathrm{MD}[\text{not.}h,e]\\ &= 0-\frac{P(\text{not.}h|e)-P(\text{not.}h)}{-P(\text{not.}h)}\\ &= \frac{[1-P(h|e)]-[1-P(h)]}{1-P(h)}=\frac{P(h)-P(h|e)}{1-P(h)},\end{aligned}$$

$$\begin{aligned}\mathrm{CF}[h,e] &= \mathrm{MB}[h,e]-\mathrm{MD}[h,e]\\ &= \frac{P(h|e)-P(h)}{1-P(h)}-0.\end{aligned}$$

Thus,

$$\begin{aligned}\mathrm{CF}[h,e]+\mathrm{CF}[\text{not.}h,e] &= \frac{P(h|e)-P(h)}{1-P(h)}+\frac{P(h)-P(h|e)}{1-P(h)}\\ &= 0.\end{aligned}$$

Clearly this result occurs because (for any h and any e) $\mathrm{MB}[h,e]=\mathrm{MD}[\text{not.}h,e]$. This conclusion is intuitively appealing since it states that evidence that supports a hypothesis disfavors the negation of the hypothesis to an equal extent.

We noted earlier that experts are often willing to state degrees of belief in terms of conditional probabilities but they refuse to follow the assertions to their logical conclusions (e.g., Paradox 1 above). It is perhaps revealing to note, therefore, that when the *a priori* belief in a hypothesis is small (i.e., $P(h)$ is close to zero), the CF of a hypothesis confirmed by evidence is approximately equal to its conditional probability on that evidence:

$$\mathrm{CF}[h,e]=\mathrm{MB}[h,e]-\mathrm{MD}[h,e]=\frac{P(h|e)-P(h)}{1-P(h)}-0\approx P(h|e),$$

whereas, as shown above, $\mathrm{CF}[\text{not.}h,e] \approx -\mathrm{P}(h|e)$ in this case. This observation suggests that confirmation, to the extent that it is adequately represented by CF's, is close to conditional probability (in certain cases) although it still defies analysis as a probability measure.

We believe, then, that the proposed model is a plausible representation of the numbers an expert gives when asked to quantify the

strength of his judgmental rules. He gives a positive number (CF>0) if the hypothesis is confirmed by observed evidence, suggests a negative number (CF<0) if the evidence lends credence to the negation of the hypothesis, and says there is no evidence at all (CF=0) if the observation is independent of the hypothesis under consideration. The CF combines knowledge of both $P(h)$ and $P(h|e)$. Since the expert often has trouble stating $P(h)$ and $P(h|e)$ in quantitative terms, there is reason to believe that a CF that weights both the numbers into a single measure is actually a more natural intuitive concept (e.g., "I don't know what the probability is that all ravens are black, but I *do* know that every time you show me an additional black raven my belief is increased by X that all ravens are black.")

If we therefore accept CF's rather than probabilities from experts, it is natural to ask under what conditions the physician's behavior based upon CF's is irrational. We know from probability theory, for example, that if there are n mutually exclusive hypotheses h_i, at least one of which must be true, then $\Sigma^n P(h_i|e)=1$ for all e. In the case of certainty factors, we can also show that there are limits on the sums of CF's of mutually exclusive hypotheses. Judgmental rules acquired from experts must respect these limits or else the rules will reflect irrational quantitative assignments. (Note we assert that behavior is irrational if actions taken or decisions made contradict the result that would be obtained under a probabilistic analysis of the behavior.)

Sums of CF's of mutually exclusive hypotheses have two limits—a lower limit for disconfirmed hypotheses and an upper limit for confirmed hypotheses. The lower limit is the obvious value that results because $CF[h,e]\geqslant -1$ and because more than one hypothesis may have CF=−1. Note first that a single piece of evidence may absolutely disconfirm several of the competing hypotheses. For example, if there are n colors in the universe and C_i is the ith color, then ARC_i may be used as an informal notation to denote the hypothesis that all ravens have color C_i. If we add the hypothesis ARC_0 that some ravens have different colors from others, we know $\Sigma_0^n P(ARC_i)=1$. Consider now the observation e that there is a raven of color C_n. This single observation allows us to conclude that $CF[ARC_i,e]=-1$ for $1\leqslant i\leqslant n-1$. Thus, since these $n-1$ hypotheses are absolutely disconfirmed by the observation e, $\Sigma_1^{n-1} CF[ARC_i,e]=-(n-1)$. This analysis leads to the general statement that, if k

mutually exclusive hypotheses h_i are disconfirmed by an observation e:

$$\sum_{i=1}^{k} \mathrm{CF}[h_i,e] \geq -k \qquad \text{(for } h_i \text{ disconfirmed by } e\text{)}.$$

In the colored raven example, the observation of a raven with Color C_n still left two hypotheses in contention, namely ARC_n and ARC_0. What, then, are $\mathrm{CF}[\mathrm{ARC}_n,e]$, $\mathrm{CF}[\mathrm{ARC}_0,e]$, and the sum of $\mathrm{CF}[\mathrm{ARC}_n,e]$ and $\mathrm{CF}[\mathrm{ARC}_0,e]$? The values of $\mathrm{CF}[\mathrm{ARC}_n,e]$ and $\mathrm{CF}[\mathrm{ARC}_0,e]$ are intimately related with the Paradox of the Ravens as discussed in Appendix 4.A. The limit on their sum, however, is important here as we attempt to characterize the rational use of CF's. In fact, it can be shown that, if k mutually exclusive hypotheses h_i are confirmed by an observation e, the sum of their CF's does not have an upper limit of k but rather:

$$\sum_{i=1}^{k} CF[h_i,e] \leq 1 \qquad \text{(for } h_i \text{ confirmed by } e\text{)}.$$

In fact, $\Sigma_{i=1}^{k}\ \mathrm{CF}[h_i,e]$ is *equal* to 1 if and only if k=1 *and* e implies h_1 with certainty, but the sum can get arbitrarily close to 1 for small k and large n. The analyses that lead to these conclusions are included as Appendix 4.B.

The last result allows us critically to analyze new decision rules given by experts. Suppose for example, we are given the following rules: $\mathrm{CF}[h_1,e]$=.7 and $\mathrm{CF}[h_2,e]$=.4 where h_1 is "The organism is a streptococcus", h_2 is "The organism is a staphylococcus", and e is "The organism is a gram positive coccus growing in chains." Since h_1 and h_2 are mutually exclusive, the observation that $\Sigma_1^2\ \mathrm{CF}[h_i,e] > 1$ tells us that the suggested certainty factors are inappropriate. The expert must either adjust the weightings or we must normalize them so that their sum does not exceed 1. In other words, because behavior based on these rules would be irrational, we must change the rules.

In concluding this section, I shall briefly examine Törnebohm's criteria for acceptability of a theory of confirmation [Törnebohm, 1966]. He states that:

It would be desirable to have a measure of evidential strength or degree of confirmation Dc satisfying the following conditions:

Dc1. If E L-implies H, then Dc($H|E$)=max.
Dc2. If E L-implies not H, the Dc($H|E$)=min.
Dc3. Dc($HE|E$) = Dc($H|E$)
Dc4. If E and H are independent of each other, then Dc($H|E$)=0.

Unfortunately it does not seem possible to construct a reasonable measure satisfying all these conditions . . .

Note that CF[H,E] satisfies Dc1, Dc2, and Dc4 for max=1 and min=−1. However, it can be shown† that CF[HE,E]=CF[H,E] if and only if P($E|H$)=1. Thus, despite its intuitive appeal, the CF we have defined fails to satisfy all four acceptability criteria suggested by Törnebohm. I shall point out later, however, that the conventions we have adopted for combining CF's allow us to satisfy Dc3.

4.6 Model as Approximation Technique

Certainty factors provide a useful way to think about confirmation and the quantification of degrees of belief. However, I have not yet described how the CF model can be usefully applied to the medical diagnosis problem. The remainder of this chapter will explain conventions that we have introduced in order to utilize the certainty factor model. Our starting assumption is that the numbers given us by experts who are asked to quantify their degree of Belief in decision criteria are adequate representations of the numbers that

†I shall demonstrate the result for E confirming H. The proof for E disconfirming H is similar.

$$\begin{aligned} \text{CF}[HE,\text{E}] &= \text{MB}[HE,E] - \text{MD}[HE,E] \\ &= \text{MB}[HE,E] - 0 \\ &= \frac{P(HE|E) - P(HE)}{1 - P(HE)} = \frac{P(H|E) - P(HE)}{1 - P(HE)} \end{aligned}$$

But

$$\begin{aligned} \text{CF}[H,E] &= \text{MB}[H,E] - \text{MD}[H,E] \\ &= \text{MB}[H,E] - 0 \\ &= \frac{P(H|E) - P(H)}{1 - P(H)} \end{aligned}$$

Thus CF[HE,E] = CF[H,E] if and only if:

$$P(H) = P(HE) = P(E|H)\,P(H)$$
i.e., $P(E|H) = 1$

would be calculated in accordance with the definitions of MB and MD if the requisite probabilities were known.

In § 4.2, when discussing Bayes' Theorem, I explained that I would like to devise a method that allows us to approximate the value for $P(D_i|E)$ solely from the $P(D_i|S_k)$, where D_i is the ith possible diagnosis, S_k is the kth clinical observation, and E is the composite of all the observed S_k. I have explained why probabilities are inadequate representations of the decision rules with which we wish to deal. Thus our goal should be rephrased in terms of certainty factors as follows:

Suppose that $\mathrm{MB}[D_i,S_k]$ is known for each S_k, $\mathrm{MD}[D_i,S_k]$ is known for each S_k, and E represents the conjunction of all the S_k. Then our goal is to calculate $\mathrm{CF}[D_i,E]$ from the MB's and MD's known for the individual S_k's.

Suppose that $E = S_1 \& S_2$, and that E confirms D_i. Then:

$$\mathrm{CF}[D_i,E] = \mathrm{MB}[D_i,E] - 0 = \frac{P(D_i|E) - P(D_i)}{1 - P(D_i)}$$

$$= \frac{P(D_i|S_1 \& S_2) - P(D_i)}{1 - P(D_i)}.$$

Clearly there is no exact representation of $\mathrm{CF}[D_i,S_1 \& S_2]$ purely in terms of $\mathrm{CF}[D_i,S_1]$ and $\mathrm{CF}[D_i,S_2]$. As was true for the discussion of Bayes' Theorem in § 4.2, the relationship of S_1 to S_2, within D_i and all other diagnoses, needs to be known in order to calculate $P(D_i|S_1 \& S_2)$. Furthermore, the CF scheme adds one complexity not present with Bayes' Theorem because we are forced to keep MB's and MD's isolated from one another.† I shall therefore introduce an approximation technique for handling the net evidential strength of incrementally acquired observations. The combining convention must satisfy the following criteria (where E_+ represents all confirming evidence acquired to date, and E_- represents all disconfirming evidence acquired to date):

†Suppose S_1 confirms D_i (MB>0) but S_2 disconfirms D_i (MD>0). Then consider $\mathrm{CF}[D_i,S_1 \& S_2]$. In this case, $\mathrm{CF}[D_i,S_2 \& S_2]$ must reflect both the disconfirming nature of S_2 and the confirming nature of S_1. Although these measures are reflected in the component CF's (it is intuitive in this case, for example, that $\mathrm{CF}[D_i,S_2] \leq \mathrm{CF}[D_i,S_1 \& S_2] \leq \mathrm{CF}[D_i,S_1]$), we shall demonstrate that it is important to handle component MB's and MD's separately in order to preserve commutativity (see item (3) of Defining Criteria).

Defining Criteria

(1) Limits:
   (a) $MB[h,E_+]$ increases towards 1 as confirming evidence is found, equalling 1 only if a piece of evidence logically implies h with absolute certainty.
   (b) $MD[h,E_-]$ increases toward 1 as disconfirming evidence is found, equalling 1 only if a piece of evidence logically implies not.h with certainty.
   (c) $CF[h,E_-] \leqslant CH[h,E_-\&E_+] \leqslant CF[h,E_+]$.
   These criteria reflect our desire to have the measure of Belief approach certainty asymptotically as partially confirming evidence is acquired, and to have the measure of Disbelief approach certainty asymptotically as partially disconfirming evidence is acquired.

(2) Absolute confirmation or disconfirmation:
   (a) If $MB[h,E_+]=1$, then $MD[h,E_-]=0$ regardless of the disconfirming evidence in E_-; i.e., $CF[h,E_+]=1$.
   (b) If $MD[h,E_-]=1$, then $MB[h,E_+]=0$ regardless of the confirming evidence in E_+; i.e., $CF[h,E_-]=-1$.
   (c) The case where $MB[h,E_+]=MD[h,E_-]=1$ is contradictory and hence the CF is undefined.

(3) Commutativity:
   If $S_1\&S_2$ indicates an ordered observation of evidence, first S_1 and then S_2:
   (a) $MB[h,S_1\&S_2]=MB[h,S_2\&S_1]$.
   (b) $MD[h,S_1\&S_2]=MD[h,S_2\&S_1]$.
   (c) $CF[h,S_1\&S_2]=CF[h,S_2\&S_1]$.
   The order in which pieces of evidence are discovered should not affect the level of Belief or Disbelief in a hypothesis. This criterion assures that the order of discovery will not matter.

(4) Missing information:
   If $S_?$ denotes a piece of potential evidence, the truth or falsity of which is unknown:
   (a) $MB[h,S_1\&S_?]=MB[h,S_1]$.
   (b) $MD[h,S_1\&S_?]=MD[h,S_1]$.
   (c) $CF[h,S_1\&S_?]=CF[h,S_1]$.
   The decision model should function by simply disregarding rules of the form $CF[h,S_2]=X$ if the truth or falsity of S_2 cannot be determined.

There are a number of observations to be made on the basis of these criteria. For example, items (1) and (2) indicate that the MB of a hypothesis never decreases unless its MD goes to 1. Similarly the

MD never decreases unless the MB goes to 1. In § 4.5, where it was always true that MB=0 or MD=0, it was always the case that either CF=MB−0 or CF=0−MD. As evidence is acquired sequentially, however, both the MB and MD may become nonzero. Thus CF=MB−MD is an important indicator of the *net* Belief in a hypothesis in light of current evidence. Furthermore, a certainty factor of zero may indicate either absence of both confirming and disconfirming evidence (as discussed in § 4.5), or the observation of pieces of evidence that are equally confirming and disconfirming. In effect $CF[h,e]=0$ is the "don't know more than I did before" value (i.e., equally confirmed and disconfirmed). Negative CF's indicate that there is more reason to disbelieve the hypothesis than to believe it. Positive CF's indicate that the hypothesis is more strongly confirmed than disconfirmed.

It is important also to note that, if $E=E_+\&E_-$, then $CF[h,E]$ represents the certainty factor for a complex new rule that could be given us by an expert. $CF[h,E]$, however, would be a highly specific rule customized for the few patients satisfying *all* the conditions specified in E_+ and E_-. Since the expert gives us only the component rules, we seek to devise a mechanism whereby a calculated cumulative $CF[h,E]$, based upon $MB[h,E_+]$ and $MD[h,E_-]$, gives a number close to the $CF[h,E]$ that would be calculated if all the necessary conditional probabilities were known.

With these comments in mind, I therefore present the following four combining functions, the first of which satisfies the criteria that I have outlined. The other three functions are necessary conventions for implementation of the model.

Combining Functions

(1) Incrementally acquired evidence:†

$$\text{(a) } MB[h,S_1\&S_2]=\begin{cases} 0 \text{ if } MD[h,S_1\&S_2]=1, \\ MB[h,S_1)+MB[h,S_2](1-MB[h,S_1]) \text{ otherwise.} \end{cases}$$

$$\text{(b) } MD[h,S_1\&S_2]=\begin{cases} 0 \text{ if } MB[h,S_1\&S_2]=1, \\ MD[h,S_1]+MD[h,S_2](1-MD[h,S_1]) \text{ otherwise.} \end{cases}$$

†It has been pointed out that the first of these functions is equivalent to:

$$MB[h,S_2]=\frac{MB[h,S_1\&S_2]-MB[h,S_1]}{1-MB[h,S_1]}.$$

Thus this combining function parallels our original definition of an MB, but with MB's

(2) Conjunctions of hypotheses:
(a) $MB[h_1 \& h_2, E] = \min(MB[h_1, E], MB[h_2, E])$.
(b) $MD[h_1 \& h_2, E] = \max(MD[h_1, E], MD[h_2, E])$.

(3) Disjunctions of hypotheses:
(a) $MB[h_1 \vee h_2, E] = \max(MB[h_1, E], MB[h_2, E])$.
(b) $MD[h_1 \vee h_2, E] = \min(MD[h_1, E], MD[h_2, E])$.

(4) Strength of evidence:
If the truth of falsity of a piece of evidence S_1 is not known with certainty, but a CF (based upon prior evidence E) is known reflecting the degree of Belief in S_1, then if $MB'[h, S_1]$ and $MD'[h, S_1]$ are the degrees of Belief and Disbelief in h when S_1 is known to be true with certainty (i.e., these are the decision rules acquired from the expert) then the actual degrees of Belief and Disbelief are given by:
(a) $MB[h, S_1] = MB'[h, S_1] \cdot \max(0, CF[S_1, E])$.
(b) $MD[h, S_1] = MD'[h, S_1] \cdot \max(0, CF[S_1, E])$.
This criterion relates to our statement early in § 4.5 that evidence in favor of a hypothesis may itself be an hypothesis subject to confirmation. Suppose, for instance, you are in a darkened room when testing the generalization that all ravens are black. Then the observation of a raven that you think is black, but that may be navy blue or purple, is less strong evidence in favor of the hypothesis that all ravens are black than if the sampled raven were known with certainty to be black. Here the hypothesis being tested is "*All* ravens are black" and the evidence is itself an hypothesis, namely the uncertain observation that "*This* raven is black."

Function (1) simply states that, since an MB (or MD) represents a proportionate decrease in Disbelief (or Belief), the MB (or MD) of a newly acquired piece of evidence should be applied proportionately to the Disbelief (or Belief) still remaining. Function (2)(a) indicates that the measure of Belief in the conjunction of two hypotheses is only as good as the Belief in the hypothesis that is believed less strongly, whereas Function (2)(b) indicates that the measure of Disbelief in such a conjunction is as strong as the Disbelief in the most strongly disconfirmed. Function (3) yields complementary results for disjunctions of hypotheses. The corresponding CF's are

substituted for the probability measures that we lack. Note also that this formula bears the same relationship to our MB definition as the sequential diagnosis form of Bayes' Theorem does to the simple Bayes formula (§ 4.2).

merely calculated using the definition CF=MB–MD. The reader is left to satisfy himself that Function (1) satisfies the Defining Criteria. (Note that $MB[h,S_?]=MD[h,S_?]=0$ when examining Criterion (4).)

Functions (2) and (3) are needed in the use of Function (4). Consider, for example, a rule such as:

$$CF'[h,S_1 \& S_2 \& (S_3 \vee S_4)] = X;$$

i.e., in our format, a rule such as:

IF: 1) THE STAIN OF THE ORGANISM IS GRAM NEGATIVE, AND
2) THE MORPHOLOGY OF THE ORGANISM IS ROD, AND
3) [A - THE AEROBICITY OF THE ORGANISM IS AEROBIC, OR
B - THE AEROBICITY OF THE ORGANISM IS UNKNOWN
THEN: THERE IS SUGGESTIVE EVIDENCE (.6) THAT THE CLASS OF THE ORGANISM IS ENTEROBACTERIACEAE

Then, by Function (4):

$$\begin{aligned} CF[h,S_1 \& S_2 \& (S_3 \vee S_4)] &= X \cdot \max(0, CF[S_1 \& S_2 \& (S_3 \vee S_4), E]) \\ &= X \cdot \max(0, MB[S_1 \& S_2 \& (S_3 \vee S_4), E] \\ &\quad - MD[S_1 \& S_2 \& (S_3 \vee S_4), E]). \end{aligned}$$

Thus, we use Functions (2) and (3) to calculate:

$$\begin{aligned} MB[S_1 \& S_2 \& (S_3 \vee S_4), E] &= \min(MB[S_1, E], MB[S_2, E], MB[S_3 \vee S_4, E]) \\ &= \min(MB[S_1, E], MB[S_2, E], \\ &\quad \max(MB[S_3, E], MB[S_4, E])). \end{aligned}$$

$MD[S_1 \& S_2 \& (S_3 \vee S_4), E]$ is calculated similarly.

It is also worth noting that Function (2) gives, for H confirmed by E:

$$\begin{aligned} CF[HE,E] &= MB[HE,E] - MD[HE,E] \\ &= \min(MB[H,E], MB[E,E]) - \max(MD[H,E], MD[E,E]) \\ &= \min(MB[H,E], 1) - \max(MD[H,E], 0) \\ &= MB[H,E] - MD[H,E] \\ &= CF[H,E] \end{aligned}$$

Thus the use of an approximation via Function (2) allows us to satisfy Dc3 of Törnebohm's criteria (see end of § 4.5) and hence to satisfy all his requirements for a quantitative approach to confirmation.

An analysis of Function (1) in light of the probabilistic definitions of MB and MD does not prove to be particularly enlightening. The assumptions implicit in this function include more than an acceptance of the independence of S_1 and S_2. The function was conceived purely on intuitive grounds in that it satisfied the four Defining Criteria I have listed. However, some obvious problesm are present. For example, the function always causes the MB or MD to increase, regardless of the relationship between new and prior evidence. Yet Salmon has discussed an example from subparticle physics [Salmon, 1973] in which either of two observations taken alone confirm a given hypothesis, but their conjunction disproves the hypothesis absolutely! Our model assumes the absence of such aberrant situations in the field of application for which it is designed. The problem of formulating a more general quantitative system for measuring confirmation is well recognized and referred to by Harré [Harré, 1970]: "The syntax of confirmation has nothing to do with the logic of probability in the numerical sense, and it seems very doubtful if any single, general notion of confirmation can be found which can be used in all or even most scientific contexts." Although we have suggested that perhaps there *is* a numerical relationship between confirmation and probability, we agree that the challenge for a confirmation quantification scheme is to demonstrate its usefulness within a given context, preferably without sacrificing human intuition regarding what the quantitative nature of confirmation should be.

Our challenge with Function (1), then, is to demonstrate that it is a close enough approximation for our purposes. We have attempted to do so in two ways. First we have implemented the function as part of the MYCIN system (§ 4.7) and have demonstrated that the technique models the conclusions of the expert from whom the rules were acquired. Second, we have written a program that allows us to compare CF's computed both from simulated real data and by using Function (1). Our notation for the following discussion will be as follows:

CF*[h,E] = the computed CF using the definition of CF from § 4.5 (i.e., "perfect knowledge" since $P(h|E)$ and $P(h)$ are known)

CF[h,E] = the computed CF using Function (1) and the known MB's and MD's for each S_k where E is the composite of the S_k's (i.e., $P(h|E)$ not known but $P(h|S_k)$ and $P(h)$ known for calculation of MB[h,S_k] and MD[h,S_k])

The program was run on sample data simulating several hundred "patients." Clearly the question to be asked was whether CF[h,E] is a good approximation of CF* [h,E]. Figure 4-1 shows a graph summarizing our results. For the vast majority of cases, the approxima-

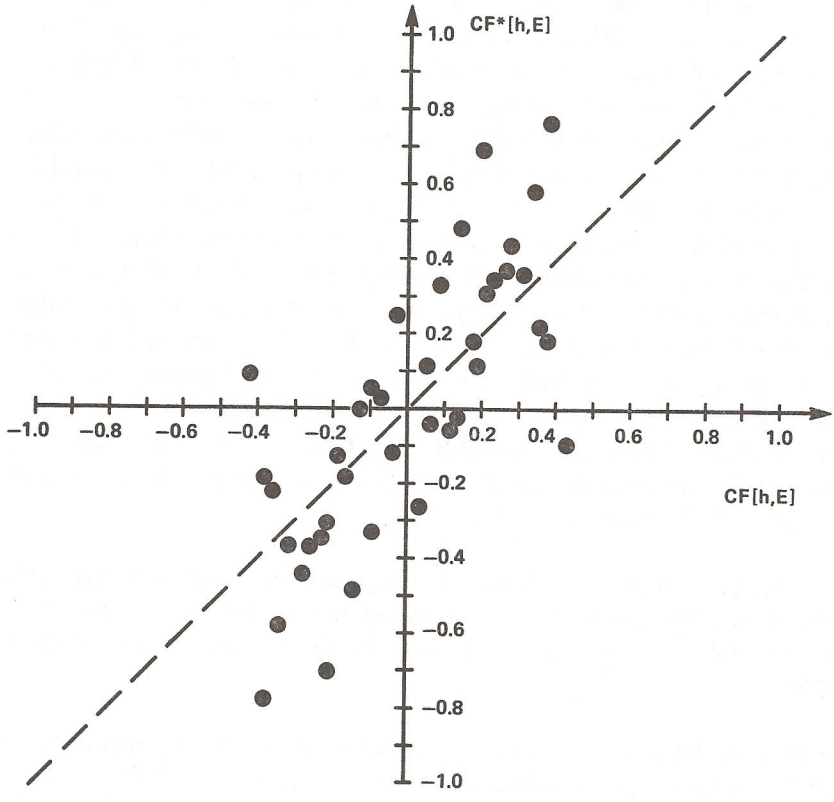

Figure 4-1: Chart demonstrating the degree of agreement between CF and CF* for a sample data base. CF is an approximation to CF*. The terms are defined in the text.

tion does not produce a CF[h,E] radically different from the true CF*[h,E]. In general, the discrepancy is greatest when Function (1) has been applied several times (i.e., several pieces of evidence have been combined). This result is in keeping with Zadeh's observation from fuzzy logic that "the more steps there are in the proof, the fuzzier the result" [Zadeh, 1974]. The most aberrant points, however, are those that represent cases in which pieces of evidence were strongly interrelated for the hypothesis under consideration (termed "conditional nonindependence"). This result is expected because it reflects precisely the issue that makes it difficult to use Bayes' Theorem for our purposes.

Thus I should make it clear that we have not avoided many of the problems inherent with the use of Bayes' Theorem in its exact form. We have introduced a new quantification scheme which, although it makes many assumptions similar to those made by subjective Bayesian analysis, permits us to utilize criteria as rules and to manipulate them to the advantages described in § 4.3. In particular, the quantification scheme also allows us to consider confirmation separately from probability and thus to overcome some of the inherent problems that accompany an attempt to put judgmental knowledge into a probabilistic format. Just as Bayesians who use their theory wisely must insist that events be chosen so that they are independent (unless the requisite conditional probabilities are known), we must insist that dependent pieces of evidence be grouped into single rather than multiple rules. As Edwards has pointed out [W. Edwards, 1972], a similar strategy must be used by Bayesians who are unable to acquire all the necessary data:

> . . . [An approximation] technique is the one now most commonly used. It is simply to combine conditionally non-independent symptoms into one grand symptom, and obtain [quantitative] estimates for that larger more complex symptom.

The system therefore becomes unworkable for applications in which large numbers of observations must be grouped in the PREMISE of a single rule in order to insure independence of the decision criteria. In addition, we must recognize logical subsumption when examining or acquiring rules and thus avoid counting evidence more than once. For example, if S_1 implies S_2, then CF[h,S_1&S_2]=CF[h,S_1] regard-

less of the value of CF[h,S_2]. Function (1) does not "know" this. Rules must therefore be acquired and utilized with care (see § 6.3).

The justification for our approach therefore rests not with a claim of improving on Bayes' Theorem but rather with the development of a mechanism whereby judgmental knowledge can be efficiently represented and utilized for the modeling of medical decision making, especially in contexts where (a) statistical data are lacking, (b) inverse probabilities are not known, and (c) conditional independence can be assumed in most cases.

4.7 Mycin's Use of Model

Formal quantification of the probabilities associated with medical decision making can become so frustrating that some investigators have looked for ways to dispense with probabilistic information altogether [Ledley, 1973]. Diagnosis is not a deterministic process, however, and we believe that it should be possible to develop a quantification technique that approximates probability and Bayesian analysis and that is appropriate for use in those cases where formal analysis is difficult to achieve. The certainty factor model that we have introduced is such a scheme. It has been implemented as a central component of the MYCIN system. The program uses certainty factors to accumulate evidence and to decide upon likely identities for organisms causing disease in patients with bacterial infections. A therapeutic regimen is then determined—one that is appropriate to cover for the organisms requiring therapy.

All of the program's knowledge is stored in decision rules such as those described in § 4.2 and 4.3. Each rule has an associated certainty factor that reflects the measure of increased Belief or Disbelief of the expert who suggested the rule. The capturing of such quantitative medical intuitions has been the subject of recent investigations by others [Card, 1970b] but, as we have noted, our approach has been simply to ask the expert to rate the strength of the inference on a scale from 1 to 10 (see § 4.3).

MYCIN remembers the alternate hypotheses that are confirmed or disconfirmed by the rules for inferring an organism's identity. With each hypothesis is stored its MB and MD, both of which are initially zero. When a rule for inferring identity is found to be true for the patient under consideration, the ACTION portion of the rule allows

either the MB or the MD of the relevant hypothesis to be updated using the first Combining Function (§4.6). When all applicable rules have been executed, the final CF may be calculated, for each hypothesis, using the definition CF=MB−MD. These alternate hypotheses may then be compared on the basis of their cumulative certainty factors. Hypotheses that are most highly confirmed thus become the basis of the program's therapeutic recommendation.

Suppose, for example, that the hypothesis H_1 that the organism is a streptococcus has been confirmed by a single rule with a CF=.3. Then, if E represents all evidence to date, MB[H_1,E]=.3 and MD[H_1,E]=0. If a new rule is now encountered which has CF=.2 in support of H_1, and if E is updated to include the evidence in the PREMISE of the rule, we now have MB[H_1,E]=.44 and MD[H_1,E]=0. Suppose a final rule is encountered for which CF=−.1. Then if E is once again updated to include all current evidence, we use Function (1) to obtain MB[H_1,E]=.44 and MD[H_1,E]=.1. If no further system knowledge allows conclusions to be made regarding the possibility that the organism is a streptococcus, we calculate a final result that CF[H_1,E]=.44−.1=.34. This number becomes the basis for comparison between H_1 and all the other possible hypotheses regarding the identity of the organism.

It should be emphasized that this same mechanism is used for evaluating *all* knowledge about the patient, not just the identity of pathogens. When the user answers a system-generated question, the associated certainty factor is assumed to be +1 unless he explicitly modifies his response with a CF (multiplied by 10) enclosed in parentheses. Thus, for example, the following interaction might occur (MYCIN's prompt is in lower-case letters):

14) Did the organism grow in clumps, chains, or pairs?
**CHAINS (6) PAIRS (3) CLUMPS (−8)

This capability allows the system automatically to incorporate the user's uncertainties into its decision processes. A rule that referenced the growth conformation of the organism would in this case find:

MB[chains, E]=0.6, MD[chains, E]=0,
MB[pairs, E]=0.3, MD[pairs, E]=0,
MB[clumps, E]=0, MD[clumps, E]=0.8.

Consider, then, the sample rule we introduced in § 4.2:

$$\mathrm{CF}[H_1, S_1 \& S_2 \& S_3] = 0.7,$$

where H_1 is the hypothesis that the organism is a streptococcus, S_1 is the observation that the organism is gram positive, S_2 that it is a coccus, and S_3 that it grows in chains. Suppose gram stain and morphology were known to the user with certainty so that MYCIN has recorded:

$$\mathrm{CF}[S_1, E] = 1, \qquad \mathrm{CF}[S_2, E] = 1.$$

In the case above, however, MYCIN would find that:

$$\mathrm{CF}[S_3, E] = 0.6 - 0 = 0.6.$$

Thus, it is no longer appropriate to use the rule in question with its full confirmatory strength of .7. That CF was assigned by the expert on the assumption that all three conditions in the PREMISE would be true with certainty. The modified CF is calculated using the fourth Combining Function (§4.6):

$$\begin{aligned}\mathrm{CF}[H_1, S_1 \& S_2 \& S_3] &= \mathrm{MB}[H_1, S_1 \& S_2 \& S_3] - \mathrm{MD}[H_1, S_1 \& S_2 \& S_3] \\ &= 0.7 \cdot \max(0, \mathrm{CF}[S_1 \& S_2 \& S_3, E]) - 0.\end{aligned}$$

Calculating $\mathrm{CF}[S_1 \& S_2 \& S_3, E]$ using the second Combining Function, this gives:

$$\begin{aligned}\mathrm{CF}[H_1, S_1 \& S_2 \& S_3] &= 0.7 \cdot 0.6 - 0 \\ &= 0.42 - 0,\end{aligned}$$

i.e.,

$$\mathrm{MB}[H_1, S_1 \& S_2 \& S_3] = 0.42$$

and

$$\mathrm{MD}[H_1, S_1 \& S_2 \& S_3] = 0.$$

Thus, the strength of the rule is reduced to reflect the uncertainty regarding S_3. Function (1) is now used to combine .42 (i.e.,

MB[H_1,S_1&S_2&S_3]) with the previous MB for the hypothesis that the organism is a streptococcus.

I have shown that the numbers thus calculated are approximations at best. Hence, it does not seem justifiable simply to accept as correct the hypothesis with the highest CF after all relevant rules have been tried. Therapy is therefore chosen to cover for all identities of organisms that account for a sufficiently high proportion of the possible hypotheses on the basis of their CF's. This is accomplished by ordering them from highest to lowest and selecting all those on the list until the sum of their CF's exceeds z (where z is equal to .9 times the sum of the CF's for *all* confirmed hypotheses).

Finally, it should be noted that our definition of CF's allows us to validate those of our rules for which frequency data become available. This will become increasingly important as the program becomes a working tool in the clinical setting where it can actually be used to gather the statistical data needed for its own validation. In the meantime, validation will necessarily involve the comments of recognized infectious disease experts who will be asked to evaluate the program's decisions and advice. Early experience with a limited set of rules has provided suggestive evidence that MYCIN will someday give advice similar to that suggested by infectious disease experts [Shortliffe, 1974b]. We are therefore gaining confidence that the certainty factor approach will continue to prove itself as the number of decision rules increases and we acquire rules from additional infectious disease experts.

Appendices

APPENDIX 4.A PARADOX OF RAVENS

In order to examine the Paradox of the Ravens (§ 4.4), I introduce the following informal notation:

- *i*RB - the hypothesis that exactly *i* ravens are black
- ARB - the hypothesis that all ravens are black (i.e., *y*RB, where *y* = the number of ravens)
- *i*nBnR - the hypothesis that exactly *i* nonblack objects are nonravens
- AnBnR - the hypothesis that all nonblack objects are nonravens (i.e., *z*nBnR where *z* = the number of nonblack objects)

BR - the observation of a raven that is found to be black
nBnR - the observation of a nonblack object that is found to be a nonraven

The paradox, then, is based on the observation that it is counter-intuitive to assert that CF[ARB,nBnR]=CF[AnBnR,nBnR]. Yet our definition of a CF quickly leads to the conclusion that the equality *does* hold since ARB is logically equivalent to AnBnR and thus P(ARB|nBnR)=P(AnBnR|nBnR). It may therefore be tempting to assert that the certainty factor model of confirmation has failed to provide insight into the paradox.

However, as Suppes has pointed out [Suppes, 1966a], the reason the paradox occurs is because we are convinced that "we are right in our intuitive assumption that we should look at randomly selected ravens and not randomly selected nonblack things in testing the generalization that all ravens are black." Expressed in terms of certainty factors, our intuition is that CF[ARB,BR] $>>$ CF[ARB,nBnR] and, in fact, that CF[ARB,nBnR]=0. Thus we prefer to sample ravens rather than nonblack objects in testing the hypothesis ARB, i.e., we feel that a black raven is significantly greater evidence in favor of the hypothesis than is a green vase.

Let us use our definition of CF, then, to calculate both CF[ARB,BR] and CF[ARB,nBnR]. We define:

y = the number of ravens in the universe
z = the number of nonblack objects in the universe

We then make the following two assumptions:

(1) $z \geqslant y$

This assumption, although clearly true for the example at hand, may seem bothersome as a requirement for the analysis. However, it can be shown that, in fact, the paradox is reversed for $z<y$. Consider, for example, a universe of 100 ravens and 5 nonblack objects that may or may not be ravens. In this case observation of a green vase is clearly *better* evidence in favor of the hypothesis that all ravens (in this limited universe) are black than is the observation of a black raven.

Suppes uses another example to make this point [Suppes, 1966a]. Suppose we want to test the generalization that all voters in a specific district are literate. We can either sample voters and see whether they are literate or else sample illiterate individuals and check to be sure they are nonvoters. The preferable strategy seems intuitively to depend upon

whether there are more voters than illiterate individuals, i.e., on the relationship between z and y from our example.

(2) We initially have no knowledge regarding either colors of ravens nor distributions of colors in the universe.
This assumption allows us to state that, before observing any ravens, we believe all the hypotheses iRB to be equally likely. This amounts to the assumption of a uniform distribution of the $P(i\text{RB})$ before sampling begins. The analysis proceeds more easily with this assumption, but it should be clear that another prior distribution will not alter the qualitative nature of our final result. Thus:

$$P(i\text{RB}) = 1/(y+1) \qquad \text{for } 0 \leqslant i \leqslant y$$

which leads to the conclusion that $P(\text{ARB})=P(y\text{RB})=1/(y+1)$.

Using assumptions (1) and (2) we can also show that:

$$P(i\text{nBnR}) = \begin{cases} 0 & \text{for } 0 \leqslant i < z-y \\ 1/(y+1) & \text{for } z-y \leqslant i \leqslant z \end{cases}$$

The proof is left for you to complete (note that there can be no fewer than $z-y$ nonravens among the z nonblack objects). It leads to the conclusion that $P(\text{AnBnR})=P(z\text{nBnR})=1/(y+1)$. This is an important result since ARB and AnBnR are logically equivalent and we therefore must require that P(ARB)=P(AnBnR).

From our definitions of certainty factors, we now note that:

$$\begin{aligned} \text{CF[ARB,BR]} &= \text{MB[ARB,BR]} - \text{MD[ARB,BR]} = \text{MB[ARB,BR]} - 0 \\ &= \frac{P(\text{ARB}|\text{BR}) - P(\text{ARB})}{1 - P(\text{ARB})} = \frac{P(\text{ARB}|\text{BR}) - [1/(y+1)]}{1 - [1/(y+1)]} \end{aligned}$$

and:

$$\begin{aligned} \text{CF[ARB,nBnR]} &= \text{MB[ARB,nBnR]} - \text{MD[ARB,nBnR]} = \text{MB[ARB,nBnR]} - 0 \\ &= \frac{\text{P(ARB|nBnR)} - \text{P(ARB)}}{1 - \text{P(ARB)}} \\ &= \frac{P(\text{AnBnR}|\text{nBnR}) - P(\text{AnBnR})}{1 - \text{P(AnBnR)}} = \frac{P(\text{AnBnR}|\text{nBnR}) - [1/(y+1)]}{1 - [1/(\text{y+1})]} \end{aligned}$$

Thus we can calculate CF[ARB,BR] if we can derive $P(\text{ARB}|\text{BR})$ and can calculate CF[ARB,nBnR] if we can derive $P(\text{AnBnR}|\text{nBnR})$. Both of the requisite conditional probabilities can be found using Bayes' Theorem:

$$P(\text{ARB}|\text{BR}) = \frac{P(\text{BR}|\text{ARB})\,P(\text{ARB})}{\sum_{1}^{y} P(\text{BR}|i\text{RB})P(i\text{RB})} \quad \frac{1\ [1/(y+1)]}{\sum_{1}^{y} [i/y]\ [1/(y+1)]} = \frac{y}{\sum_{1}^{y} i}$$

$$= 2/(y+1) \qquad \text{since } \sum_{1}^{y} i = y(y+1)/2$$

$$P(\text{AnBnR}|\text{nBnR}) = \frac{P(\text{nBnR}|\text{AnBnR})\,P(\text{AnBnR})}{\sum_{1}^{z} P(\text{nBnR}|i\text{nBnR})\,P(i\text{nBnR})}$$

$$= \frac{1\ [1/(y+1)]}{\sum_{z-y}^{z} [i/z]\ [1/(y+1)]}$$

$$= \frac{z}{\sum_{1}^{z} i - \sum_{1}^{z-y-1} i} = \frac{2z}{z(z+1) - (z-y-1)(z-y)}$$

$$= \frac{2z}{2z + 2zy - y - y^2} = \frac{2z}{(y+1)\,(2z-y)}$$

$$= \frac{2}{y+1} \cdot \frac{z}{2z-y} = (2z)/[(y+1)(2z-y)]$$

Note that $P(\text{ARB}|\text{BR}) = P(\text{ARB}|\text{nBnR})$ if $z=y$!

Thus:

$$\text{CF[ARB,BR]} = \frac{[2/(y+1)] - [1/(y+1)]}{1 - [1/(y+1)]} = 1/y$$

and:

$$\text{CF[ARB,nBnR]} = \frac{(2z)/[(y+1)(2z-y)] - [1/(y+1)]}{1 - [1/(y+1)]} = \frac{1}{(2z-y)}$$

Note that CF[ARB,BR]$\geqslant$CF[ARB,nBnR] and that the equality only holds when $z=y$. Thus, if there are fewer ravens than nonblack objects, observing a black raven confirms the hypothesis ARB more strongly than a green vase confirms that all ravens are black.

But we wished to show that our intuition is correct in suggesting that CF[ARB,BR] >> CF[ARB,nBnR] and that CF[ARB,nBnR]=0. As mentioned in the discussion of assumption (1) above, our intuition is tainted by our knowledge of real work. For instance, we may be willing to accept estimates of y and z such that $y = 10^7$ and $z=10^{15}$. Actually z is undoubtedly larger, but these numbers will suffice for current purposes. Then:

$$\begin{aligned} \text{CF[ARB,BR]} &= 1/(10^7) = .0000001 \\ \text{CF[ARB,nBnR]} &= 1/(2\cdot 10^{15} - 10^{17}) \approx 1/(2\cdot 10^{15}) \\ &\approx .0000000000000005 \end{aligned}$$

Clearly CF[ARB,nBnR] is essentially zero, and CF[ARB,BR] is significantly greater than CF[ARB,nBnR]. Note, however, that these results are obtained only because we are willing to accept the original estimates for x and y.

APPENDIX 4.B PROOF OF UPPER LIMIT

I include here a proof of the assertion that the sum of the CF's of confirmed but mutually exclusive hypotheses cannot exceed 1. Since MD[h,e]=0 for a hypothesis that is confirmed by e, CF[h,e]=MB[h,e] when e confirms h. Suppose there are n mutally exclusive hypotheses h_i confirmed by evidence e. Then we wish to identify the upper limit on Σ_1^n CF[h_i,e], i.e., on Σ_1^n MB[h_i,e]. To simplify the manipulation of symbols, let:

$a_i = P(h_i|e)$ such that $\sum^n a_i \leqslant 1$,

$b_i = P(h_i)$ such that $\sum^n b_i < 1$ and $0 < b_i < 1$ for all i.

Then:

$a_i > b_i$ for all i since the h_i are confirmed by e

We wish to find the upper limit, if any, on:

$$\sum^n \text{MB}[h_i,e] = \sum^n \frac{a_i - b_i}{1 - b_i}$$

Proof: We first note that, for $n=1$:

$$\sum^{n} \frac{a_i - b_i}{1 - b_i} = \frac{a_i - b_i}{1 - b_i} \leqslant 1 \quad \text{since} \quad a_i \leqslant 1.$$

For $n>1$, however:

$$\sum^{n} \frac{a_i - b_i}{1 - b_i} < \sum^{n} \frac{a_i - b_i}{(1 - b_i) \prod_{j \neq i}^{n} (1 - b_j)} \qquad \left(\text{since } \prod_{1}^{n} (1 - b_j) < 1\right)$$

$$< \frac{\sum^{n} (a_i - b_i)}{\prod^{n} (1 - b_j)} = \frac{\sum^{n} a_i - \sum^{n} b_i}{\prod^{n} (1 - b_j)}.$$

But:

$$\prod^{n} (1 - b_j) = 1 - \sum_{i=1}^{n} b_i + \sum_{i=1}^{n} \sum_{j \neq i}^{n} b_i b_j - \sum_{i=1}^{n} \sum_{j \neq i}^{n} \sum_{k \neq j \neq i}^{n} b_i b_j b_k + \cdots$$

$$= 1 - \sum_{i=1}^{n} b_i + \sum_{i=1}^{n} \sum_{j \neq i}^{n} b_i b_j \left(1 - \sum_{k \neq j \neq i}^{n} b_k\right)$$

$$+ \sum_{i=1}^{n} \sum_{j \neq i}^{n} \sum_{k \neq j \neq i}^{n} \sum_{l \neq k \neq j \neq i}^{n} b_i b_j b_k b_l \left(1 - \sum_{m \neq l \neq k \neq j \neq i}^{n} b_m\right) + \cdots.$$

And since $\sum^{n} b_i < 1$, $1 - \sum^{n} b_i > 0$ in all terms above. Thus $\prod^{n} (1 - b_j) > 1 - \Sigma^{n} b_i$. Therefore:

$$\sum^{n} \frac{a_i - b_i}{1 - b_i} < \frac{\sum^{n} a_i - \sum^{n} b_i}{\prod^{n} (1 - b_j)}$$

$$< \frac{\sum^{n} a_i - \sum^{n} b_i}{1 - \sum^{n} b_i} \leqslant \frac{1 - \sum^{n} b_i}{1 - \sum^{n} b_i} \qquad \left(\text{since } \sum^{n} a_i \leqslant 1\right)$$

$$< 1.$$

Thus we have demonstrated that 1 is the upper limit for the sum of the CF's of confirmed mutually exclusive hypotheses.

The rather weak inequality we have shown is better understood, however, if we examine a special case. Suppose there are m mutually exclusive hypotheses such that $\Sigma_1^m P(h_i)=1$. We assume that each is initially equally likely, i.e., $P(h_i)=1/m$. Suppose now that first n of the m hypotheses are confirmed by the evidence e. Then:

$$\sum^{n} \mathrm{CF}[h_i,e] = \sum^{n} \mathrm{MB}[h_i,e] - \sum^{n} \mathrm{MD}[h_i,e]$$

$$= \sum^{n} \frac{P(h_i|e)-P(h_i)}{1-P(h_i)} - 0 = \sum^{n} \frac{P(h_i|e)-1/m}{1-1/m}$$

$$= \sum^{n} \frac{mP(h_i|e)-1}{m-1} = \frac{1}{m-1}\left[m\sum^{n} P(h_i/e)-n\right]$$

$$= \frac{m\sum^{n} P(h_i|e)-n}{m-1} \leqslant 1.$$

This interesting result shows that the sum is equal to 1 only if h_1 is taken to be certain on the basis of e *and* when $n=1$. If only two hypotheses remain possible after e has been observed and all the others have been ruled out with certainty, $\Sigma_1^n P(h_i|e)=1$ but $\Sigma_1^n \mathrm{CF}[h_i,e]=(m-2)/(m-1)$ and is therefore less than 1.