Expert Opinion
About Uncertainty

Barbara Ann Heinrich

April 1, 1971

71-1-12

This research was supported by the Personnel and Training Research Programs Office of the Office of Naval Research under ONR Contract No. N00014-67-A-0103-0011, Contract Authorization No. NR 151-313. Reproduction in whole or in part is permitted for any purpose of the United States government. This document has been approved for public release and sale; its distribution is unlimited.

52

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
University of Washington
Department of Psychology
Seattle, Washington

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE

EXPERT OPINION ABOUT UNCERTAINTY

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Scientific Interim

5. AUTHOR(S) *(First name, middle initial, last name)*

Barbara Ann Heinrich

6. REPORT DATE
April 1, 1971

7a. TOTAL NO. OF PAGES
51

7b. NO. OF REFS
77

8a. CONTRACT OR GRANT NO.
N00014-67-A-0103-0011 NR151-313

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
None 71-1-12

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
None

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; it's distribution is unlimited.

11. SUPPLEMENTARY NOTES
TECH, OTHER

12. SPONSORING MILITARY ACTIVITY
Personnel and Training Research Programs Office, Office of Naval Research

13. ABSTRACT

The paper reviews decision theoretic approaches to the use of experts' verbal estimates of the probabilities of events in military decision making, weather forecasting, and medical diagnosis. The basic concept is Bayes' Theorem, and scoring rules and the data on verbal estimates are reviewed.

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Decision Making						
Intuitive Statistics						
Subjective Judgments						

EXPERT OPINION ABOUT UNCERTAINTY

Barbara Ann Heinrich

In order to make an optimal decision in the face of uncertainty, most decision-making models require that the decision-maker have two types of information: estimates of the probabilities of the alternatives involved and estimates of the worths of the alternatives. The more accurate these estimates are, the greater the probability that an optimal decision will be made. Calling on experts to give their opinions is one way of obtaining more accurate estimates of the probabilities and worths involved. The decision-maker can then use the opinions of the experts to arrive at a final decision, or, in fact, can have the expert make the decision for him.

PROBABILITY, WORTH, AND EXPECTED VALUE

The estimates given by the expert can be derived from either objective or subjective information. Probability estimates based on relative frequencies or the logical constraints of a situation are considered to be objective. Probability estimates based on the personal opinion of an individual are considered to be subjective. The same may be said for estimates of worth. Some estimates of the worths of alternatives can be expressed in terms of dollars and cents (i.e. monetary value), while other estimates can only be expressed in terms of subjective feelings (i.e. utility). In many cases an optimal decision simply requires the decision-maker to appropriately combine probability and worth and choose the most favorable alternative. Mathematically the most favorable alternative is that alternative which has the highest expected value. Since each alternative has to have two or more possible outcomes, the expected value for any given alternative can be calculated by simply multiplying the probabilities of each possible outcome of that alternative by their worths and summing these values. (For a discussion of the various types of expected value models, i.e. EV, EU, SEV, and SEU, see Edwards, 1955). In the opinion of some theorists, maximizing expected value is the fundamental principle of rational behavior (e.g. Good, 1952). This principle has actually been used in many areas of applied decision-making such as medical diagnosis and weather forecasting. Thus the expert, knowing the probabilities and worths of the alternatives involved can make an optimal decision in the face of uncertainty by maximizing his expected value.

If an expert must have knowledge of the probabilities and worths involved in order to maximize his expected value, then these entities must be quantifiable, especially in cases where the expert gives his knowledge to the decision-maker rather than making the final decision himself. Of course objective probabilities and monetary values are readily quantifiable. But experts usually do not have objective probabilities and values available to them. Thus they must use their own subjective estimates. It will be shown shortly that subjective probabilities can be measured and quantified and are often quite accurate. Although I realize that value and utility estimates are also important to the expert in making a final decision, they will not be discussed in this paper. For an extended discussion of utility, etc., the reader can consult Parts I and II of Edwards and T versky (1967).

MEASUREMENT OF SUBJECTIVE PROBABILITIES

Two methods have been used to infer an individual's subjective probabilities. The first method involves having an individual make choices among bets. His subjective probabilities are then inferred from his betting behavior on the assumption that he is trying to maximize his expected value (e.g. Preston and Baratta, 1948; Beach and Phillips, 1967; Beach and Wise, 1969). The second, less involved method consists of simply asking the individual for verbal estimates of his subjective probabilities (e.g. Attneave, 1953; Beach, 1966; Beach and Wise, 1969). Accuracy of subjective probabilities can be determined by comparing an individual's subjective estimates or inferred probabilities with the objective probabilities of the events which he experienced. Both methods, bets and verbal estimates, have shown that individuals frequently are quite accurate at estimating probabilities, depending upon their experience with the events in question. They tend to be slightly conservative in their estimates (i.e. they over-estimate low probabilities and under-estimate high probabilities).

Beach and Phillips (1967) and Beach and Wise (1969) have also shown that subjective probabilities inferred from choices among bets and those estimated by individuals are practically equivalent to each other. However, estimated subjective probabilities appear to correlate slightly higher with objective probabilities than inferred subjective probabilities. This is fortunate since in most applied decision-making situations the expert is simply asked to give probability estimates. Choosing among bets can be a time-consuming and costly venture for any expert and therefore is the less preferred method for obtaining subjective probabilities in applied decision-making.

While all of this research shows that individuals are accurate in estimating subjective probabilities, there is still the problem of determining just who is an expert or good probability assessor. Winkler (1967a) and Winkler and Murphy (1968) suggest two standards of "goodness" which could be used to evaluate a probability assessor: normative and substantive. Normative means that the expert can make his probability assessment correspond to his judgements. Substantive means that the expert's probability assessments correspond to something in reality. For example, suppose individual A predicts a .20 chance of rain tomorrow while individual B predicts a .90 chance of rain. Tomorrow comes and it rains. Who is the better probability assessor, A or B? Substantively speaking, B would be called the better probability assessor. However, if both A and B's assessments correspond to their own respective judgements, then normatively speaking, both are good probability assessors. In fact, if both A and B feel that the probability of rain is about .20, then A would be the better assessor in terms of normative evaluation.

Basically this paper will be concerned with the substantive aspect of probability assessment since it will be assumed that in most cases the expert's assessments will correspond to his judgements, i.e. he is a good normative probability assessor. It should be kept in mind that if an individual is not a good substantive probability assessor, it may be because his assessments fail to correspond to his judgements, in which case training in statistics, etc., might be helpful.

CONCEPT OF DIAGNOSIS AND SCOPE OF PAPER

In a broad sense we might think of an expert as an individual who attempts to diagnose an event or situation and determine how likely it is that his diagnosis is true. Any expert in a diagnostic situation must obtain as much information as possible to aid him in arriving at his final diagnosis. Of course, the diagnosticity of this information will vary. Highly diagnostic information is data which is highly valuable to the decision-maker. That is, it has strong implications for the problem being considered. Mildly or weakly diagnostic information, while not too valuable as a single datum, can aid the decision-maker when several data are considered simultaneously. In this paper the final diagnosis will often be referred to as the expert's final subjective probability estimate. Just how the expert uses or combines this information and arrives at his final subjective probability estimate is going to be the focus of this paper. Decision-making by experts in three different types of diagnostic situations will be explored: Military decision-making, weather forecasting, and medical diagnosis. However, before looking into these areas I would like to discuss the optimal way for the expert to arrive at his final subjective probability or diagnosis. In order to do so, I will first discuss probability theory and Bayes' theorem.

SUBJECTIVE PROBABILITY AND PROBABILITY THEORY

Since the logic of probability theory is going to be used to aid experts in making optimal decisions, it is necessary to show that subjective probabilities satisfy the basic axioms of the mathematical theory of probability if they are to be used along with the logic of probability theory. The basic axioms of the mathematical theory of probability are as follows:

1. A probability is a number which lies between 0 and 1.
2. The sum of an exhaustive set of mutually exclusive events is equal to one.
3. The probability of either of two mutually exclusive events occurring is the sum of their individual probabilities.
4. The probability of two independent events both occurring is the product of their individual probabilities.

Do subjective probabilities sum to one? The studies are few and the results contradictory. In some cases the sum has been found to be greater than one; in other cases, less than one. Methodological problems often arise in trying to determine if subjective probabilities sum to one (Peterson and Beach, 1967). However, in applied decision-making this inconsistency can easily be removed by normalizing the probabilities or having the expert reassess his probabilities in some other manner so that they do sum to one. As Winkler (1967a, p. 1114) states, "...assessors can be taught, to a certain degree, to identify and reconcile inconsistencies. Because of this, the fact that people do violate the postulates of coherence [i.e.consistency] should not create a serious problem." Thus the fact that subjective probabilities may not sum to one need not necessarily create serious difficulty for applied decision-makers.

People are not as inconsistent as one might think. Even if they are inaccurate with respect to assessing the true probability of the occurrence of a single event or having their probabilities sum to one, they combine subjective probabilities of two or more events in accordance with axioms 3 and 4. Beach and Peterson (1966) have shown that estimates of unions of events are equal to the sums of the estimates for the component events. Likewise, Shuford (1959) and Peterson, Ulehla, Miller, Bourne, and Stilson (1965) have shown that estimates of joint probabilities closely approximate the products of the component events.

Bayes' Theorem

Since individuals use the logic of probability theory in arriving at their subjective probability estimates, one might also ask if they use the logic of probability theory to modify their judgements in the light of new information. Probability theory states that:

$$P(A \cap B) = P(A/B)\ P(B) = P(B/A)\ P(A)$$

That is, the probability of the joint occurence of both events A and B is equal to the conditional probability of A, given that B has occurred, times the probability of B, and so on. From this it follows that:

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B/A)\ P(A)}{P(B)} \qquad (1)$$

This is called Bayes' theorem and is considered the optimal way to revise one's opinion in the light of new information. If we let A stand for any hypothesis (H) and B stand for any datum (D), we can rewrite equation (1) in the form:

$$P(H/D) = \frac{P(D/H)\ P(H)}{P(D)} \qquad (2)$$

How Bayes' Theorem Works. An illustration given by Morris (1968, pp. 31-32) will illustrate how Bayes' theorem works. Suppose your friend has an ordinary coin and a special die, four sides of which are labeled "heads" and the remaining two, "tails". Out of your sight, your friend flips the coin and rolls the die and then covers one of the objects with a cup. Your task is to guess which object is uncovered, the coin or the die. Let H_c stand for the hypothesis "the coin is uncovered" and H_d stand for the hypothesis "the die is uncovered". Since you would most likely be indifferent about either hypothesis, your initial probabilities for H_c and H_d would be 1/2. That is, $H_c = H_d = 1/2$. Now suppose your friend tells you whether the top of the uncovered object is heads or tails. How should this influence your opinion about the uncovered object? Let's say he says, "Heads is showing." Intuitively you would change your opinion to favor the die. According to Bayes' theorem, you would also change your opinion to favor the die. Why? If we let "h" stand for the report "heads is showing", then it follows that:

$$P(h/H_c) = 1/2 \quad \text{and} \quad P(h/H_d) = 2/3$$

Bayes' theorem prescribes that your revised opinion should be

$$P(H_d/h) = \frac{P(h/H_d)\ P(H_d)}{P(h)}$$

where $P(h) = P(h/H_c)\ P(H_c) + P(h/H_d)\ P(H_d)$. Substituting the values given above we get:

$$P(H_d/h) = \frac{(2/3)\ (1/2)}{(1/2)(1/2) + (2/3)\ (1/2)} = 4/7$$

Thus Bayes' theorem tells you your probability of H_d should rise from 1/2 to 4/7. Note Bayes' theorem not only gives you the direction of your revision, but also the amount of revision you should make concerning a particular hypothesis.

So that it will be easier to talk about Bayes' theorem in non-mathematical terms, I would like to simplify its description and discuss the use of Bayes' theorem in ratio form. In equation (2), P(H) is called the prior or a priori probability. It is the initial probability or opinion of an event given no other information. P(D/H) is the probability of a given sample result, observation, or item of information (i.e. a piece of data), under the assumption that some particular hypothesis is true. It is called the likelihood. P(H/D) is called the posterior probability and is the prior probability revised after receiving additional information. Note that a posterior probability may become a prior probability if additional data becomes available.

Bayes' Theorem and the Likelihood Ratio. It has been suggested (Edwards, Lindman, and Savage, 1963; Lusted, 1968) that a particularly convenient version of Bayes' theorem for some applications in decision-making is the odds-likelihood ratio form. That is, suppose you are entertaining two hypotheses, H_1 and H_2. Bayes' theorem can be written for each one:

$$P(H_1/D) = \frac{P(D/H_1)\ P(H_1)}{P(D)} \tag{3}$$

$$P(H_2/D) = \frac{P(D/H_2)\ P(H_2)}{P(D)} \tag{4}$$

Dividing equation (3) by equation (4) we obtain:

$$\frac{P(H_1/D)}{P(H_2/D)} = \frac{P(D/H_1)}{P(D/H_2)} \times \frac{P(H_1)}{P(H_2)} \tag{5}$$

which is often written as:

$$\Omega_1 = L\,\Omega_0 \tag{6}$$

where Ω_1 = the posterior probability of H_1 and H_2 in odds, L = the likelihood ratio, and Ω_0 = the prior probability of H_1 and H_2 in odds. According to Edwards (1966a) estimating odds or likelihood ratios, $\frac{P(D/H_1)}{P(D/H_2)}$, appears to be easier for most individuals than estimating conditional probabilities, $P(D/H_1)$ or $P(D/H_2)$, and for this reason, equation (6) rather than equation (2) is often used in applied decision-making.

Note that if the likelihood ratio is 1.00, it will have no effect on one's opinion and the posterior probability will be the same as the prior probability. The more the likelihood ratio differs from one, the greater the difference between the prior and posterior probability. Notice also that as long as $P(D/H_1)$ and $P(D/H_2)$ are multiplied by a constant, the final odds are not affected. This is called the likelihood principle, and it states that all observations leading to the same likelihood ratio should lead to the same conclusion. Or, as Edwards, Lindman, and Savage (1963) put it, two potential data, D_1 and D_2, are of the same import if they lead to the same likelihood ratio.

Bayes' Theorem as a Model for Human Decision-Making. Do people revise their opinions according to Bayes' theorem? The data indicate that in some cases they do (e.g. Peterson, Ulehla, Miller, Bourne, and Stilson, 1965; Beach, 1966; Phillips and Edwards, 1966). While people's behavior resembles Bayes' theorem, their revised estimates are usually much more conservative than Bayes' theorem would predict (Edwards, 1968). It should be pointed out that if an expert's subjective probabilities do not always conform to the basic axioms and logic of probability theory, including Bayes' theorem, it may be because he displays inconsistent behavior in arriving at his final probability estimates. Because individuals are inconsistent at times, this is no reason for rejecting Bayes' theorem and probability theory as a model for describing human decision-making. As de Finetti states (1965, p. 88): "Although it is known that people often do not exhibit logical or probabilistic coherence [i.e. consistency], this only makes it more important to use probability theory to show them how to avoid unnecessary losses due to such inconsistency."

DECISION-MAKING IN THE REAL WORLD

Are subjective probabilities and Bayes' theorem adequate models of how experts form and revise their opinions? To answer this question, three areas of applied decision-making will be examined. Each of these areas, military decisior-making, weather forecasting, and medical diagnosis, require decisions to be made in the face of uncertainty. Information about the states of the world (i.e. the probabilities of events) are obtained from experts trained in each of these areas. The question then becomes one of how reliable and "expert" their opinions are. Before reviewing each of these areas, some general questions that can be asked about the quantification and evaluation of expert opinion will be discussed, and laboratory data will be summarized. These questions will provide a framework for the subsequent discussion of applied decision-making.

The Questions

Basically there are six questions one can ask about the formation and use of expert opinion. These are:

1. Is there any economic value in using probability estimates, as opposed to categorical statements, in the decision-making process?
2. Given that probabilistic statements have some value, should probability estimates be objective or subjective?
3. Should individual or group (i.e. consensus) probability estimates be used?
4. What role should computers play in applied decision-making?
5. What kinds of data should the expert base his opinion upon, and does he use the available data appropriately?
6. How can individuals be trained to become experts, and how can experts improve their own performance?

The Evidence

Questions 1, 2, and 3. Questions 1 and 2 cannot be appropriately discussed without reference to a specific area of applied decision-making, so no laboratory evidence will be presented here. As far as individual versus consensus estimates are concerned (question 3), Winkler (1967b), in an experiment involving the assessment of probabilities for the outcomes of collegiate and NFL football games by college students, found consensus estimates to be better than individual estimates. Given that consensus probability estimates appear to be valuable, question 3 raises such problems as how one can best combine the opinion of several experts in the same field (e.g. five radiologists) or different fields (e.g. a diagnostic group consisting of a radiologist, a surgeon, and an internist). Winkler

(1968) has suggested several methods for arriving at a consensus of subjective probability distributions or estimates. He divides these methods into two general categories: Mathematical approaches and behavioral approaches.

Mathematical approaches. These approaches involve using either a weighted-average or combining expert opinions using Bayes' theorem. The difficulty with using the weighted-average method lies in determining how to weight the experts' opinions. Winkler (1968) suggests several ways: Assign equal weights; assign weights according to where the expert lies on a ranking scale of "expertness"; assign weights according to experts' self-ratings of "expertness"; or assign weights based primarily on previous performance of the experts. In one study, Winkler (1967b) found little difference between three different weighted-average consensus systems.

The second mathematical method essentially involves combining a group of experts' estimates by having a final decision-maker treat each expert's estimate as a datum from a sample, and revising his own opinions using Bayes' theorem.

Behavioral approaches. These approaches involve simply letting a group of experts come up with a final probability estimate. This can be done using either of two methods. The first method involves allowing each expert to see the opinions of the remaining experts without actually meeting them. Of course this system will work only if repeated reassessments by the experts lead to some sort of convergence of opinion.

The second method allows experts to discuss the issues with each other in order to arrive at a final probability estimate. One of the difficulties with this method is that the experts may falsify their own estimates in the hope of swaying other experts towards their own point of view. Once a group probability estimate has been arrived at, the experts might want to reassess their own individual estimates. These revised estimates might then be combined mathematically, this final estimate being used rather than the group's non-mathematically derived opinion.

Winkler (1968) investigated the use of several of these methods. Since he had no "correct" opinion, he could not determine which method was most accurate. He did find, however, that different methods produced different results. He also found that when the behavioral approach is used, convergence of opinion does indeed occur. That is, the difference between the experts' individual opinions decreased after group feedback, with or without contact.

Question 4. Computers can play a valuable role in applied decision-making. A good example of a general man-machine system called Probabilistic Information Processing, or PIP for short, is given by Edwards (1966b) and Edwards et al. (1968). This system appears to be fairly successful and will be described in greater detail when military decision-making is discussed. Yntema and Torgerson (1961) have also shown that computers can aid man in making decisions. Their system takes both probability and worth into account and arrives at a final decision by maximizing expected value.

Question 5. Many issues are implicit in the question of the expert's use of data. Some types of data may be more valuable than others in helping the expert form accurate subjective probabilities. Do experts give these data more weight? And do experts agree on such weightings? Stated a little differently, are experts able to discriminate between highly diagnostic and mildly or weakly diagnostic data? One way of evaluating this is by comparison of experts' estimated likelihood ratios for different data. Alternatively, a multiple regression model (e.g. Beach, 1967) or a non-linear configurality model (Hoffman, 1968) could be used to analyze how the "better" experts utilize the information available to them.

A second issue is how information should be presented to an expert for analysis. Data can be presented all at once (i.e. simultaneously) or piece-by-piece (i.e. sequentially). Peterson and DuCharme (1967) have found a primacy effect when data are given sequentially. Too much data can also create problems. It is well known that individuals have difficulty aggregating information when the amount of data is increased (e.g. Peterson, Schneider, and Miller, 1965; Peterson and Swensson, 1968). As we shall see later, in applied decision-making, where the amount of data may be considerable, use of a PIP-type system (e.g. Edwards, 1966b) is a possible solution to this problem.

A third issue is how much data is needed in order for an expert to form his opinion. In some applied decision-making situations, data can be expensive and therefore prohibitive. Thus part of the skill of the expert lies in knowing when it is appropriate to stop obtaining additional data and quantify his opinion. While there is some evidence that individuals "purchase" information optimally, (e.g. Edwards and Slovic, 1965), other research has shown that individuals, instead of purchasing an optimal amount of information, tend to purchase too much or too little information (e.g. Pitz, 1968; Pitz, Reinhold, and Geller, 1969), depending upon the exact nature of the task presented to them. Fried and Peterson (1969) have also found that while individuals do a near optimal job of purchasing information in a fixed stopping condition (i.e. where the individual has to decide prior to purchasing how much information he wishes to buy), they tend to purchase too little information in an optional stopping condition (i.e. where the individual decides after receiving a piece of information if he wants to buy more information).

Question 6. The last question concerns the training of naive individuals to become experts and the improvement of the performance of existing experts, given that they are performing sub-optimally. This is intimately tied up with questions 3, 4, and 5, and leads us to a whole series of issues. There is the issue of whether an expert should be an expert in statistics, in his field of interest, or both. There is some evidence that training in statistical concepts leads to better performance (e.g. Wheeler and Beach, 1968; Peterson, DuCharme, and Edwards, 1968). Winkler (1967b) showed that being an expert in one's own field helps. He found that sportswriters were better than college students at predicting scores of football games. Evaluation scores and giving payoffs for good performance may be used to aid the expert in giving good unbiased estimates (e.g. Phillips and Edwards, 1966; Winkler and Murphy, 1968).

Another issue lies in whether there is a need for developing prior probabilities in experts and experts-to-be. Prior probabilities are important if very little new data is forthcoming. However, if there is a great deal of new data, prior probabilities become relatively unimportant. Edwards, Lindman, and Savage (1963) have mentioned that if several individuals were to start with completely different prior probabilities about the same hypothesis and were all given a great deal of additional data, they would all end up with very similar final posterior probabilities. However, no psychological data on this argument seem to exist. With respect to training potential experts, the recent development of computer simulation training techniques, which can teach students prior and likelihood probabilities, may prove valuable, but little has been done in this direction.

Laboratory Versus Real World Decision-Making

Now that the questions have been raised, discussed, and answered to the extent that there is relevant laboratory evidence, let us look at what is going on in the real world. It should be kept in mind that laboratory studies on Bayesian decision-making have been concerned with modifying Bayes' theorem, the optimal model for making decisions, to make it descriptive of how human beings actually make decisions. Laboratory experimenters have been primarily interested in finding out if information processing by human beings resembles Bayesian logic. They have relied upon very abstract tasks to test their hypotheses. These tasks typically involve having an individual try to determine from which of two well-defined populations (e.g. two urns filled with poker chips) a given sample of events (i.e. poker chips) has been selected. For example, urn A may contain 70 red chips and 30 blue chips while urn B may contain 70 blue chips and 30 red chips. An urn is randomly selected by the experimenter and a sample of "x" chips is drawn. The individual, ignorant of which urn has been selected, is shown the sample and is asked to state the probability that the sample was drawn from urn A or urn B.

In applied situations, such as the ones we are about to discuss, the "experimenters" are concerned more with producing optimal results; that is, in using Bayesian logic to help people make optimal decisions. If an expert does not perform optimally according to Bayes' theorem, the experimenter will do what he can to help the expert perform optimally. It should also be remembered that optimal performance does not mean "always being right" but rather maximizing one's expected value over a long period of time. Not all of the questions that have been raised above have answers in the real world. Part of this is due to the fact that many companies, professions, etc., are not strongly pressed to do research and report their techniques in existing journals. Fortunately there are three areas in the real world for which there are considerable data available. These three areas will be reviewed in the following order: Military decision-making, weather forecasting, and medical diagnosis.

MILITARY DECISION-MAKING AND PIP SYSTEMS

Economic Value (Question 1)

While the exact procedures used in much military decision-making are unavailable because they are classified, there is a rather large body of literature concerning military decision-making from simulation studies performed in the laboratory. There is no doubt that there is much to be gained from optimal military decision-making. Money and lives can be saved by determining which of several strategies the enemy may be using, or determining if a radar signal received symbolizes a friend, an enemy, or something else. Given a certain number of fighter jets, one may want to determine how to schedule their missions so as to get maximum benefit from them in the shortest amount of time.

Basic PIP System (Question 4)

Since men are conservative information processors (Edwards, Lindman, and Phillips, 1965; Phillips and Edwards, 1966; Schum and Martin, 1968), the problem arises of how to aid them in making more nearly optimal decisions. Edwards (1966b) has suggested having experts estimate P(D/H) rather than P(H/D), and letting a computer aggregate the P(D/H) estimates to come up with a final posterior probability estimate. This type of a system is called a Probabilistic Information Processing system, or PIP for short. A general diagram of how such a system might work for the military is given in figure 1. Briefly, a list of all the available hypotheses (e.g. Russia is about to attack

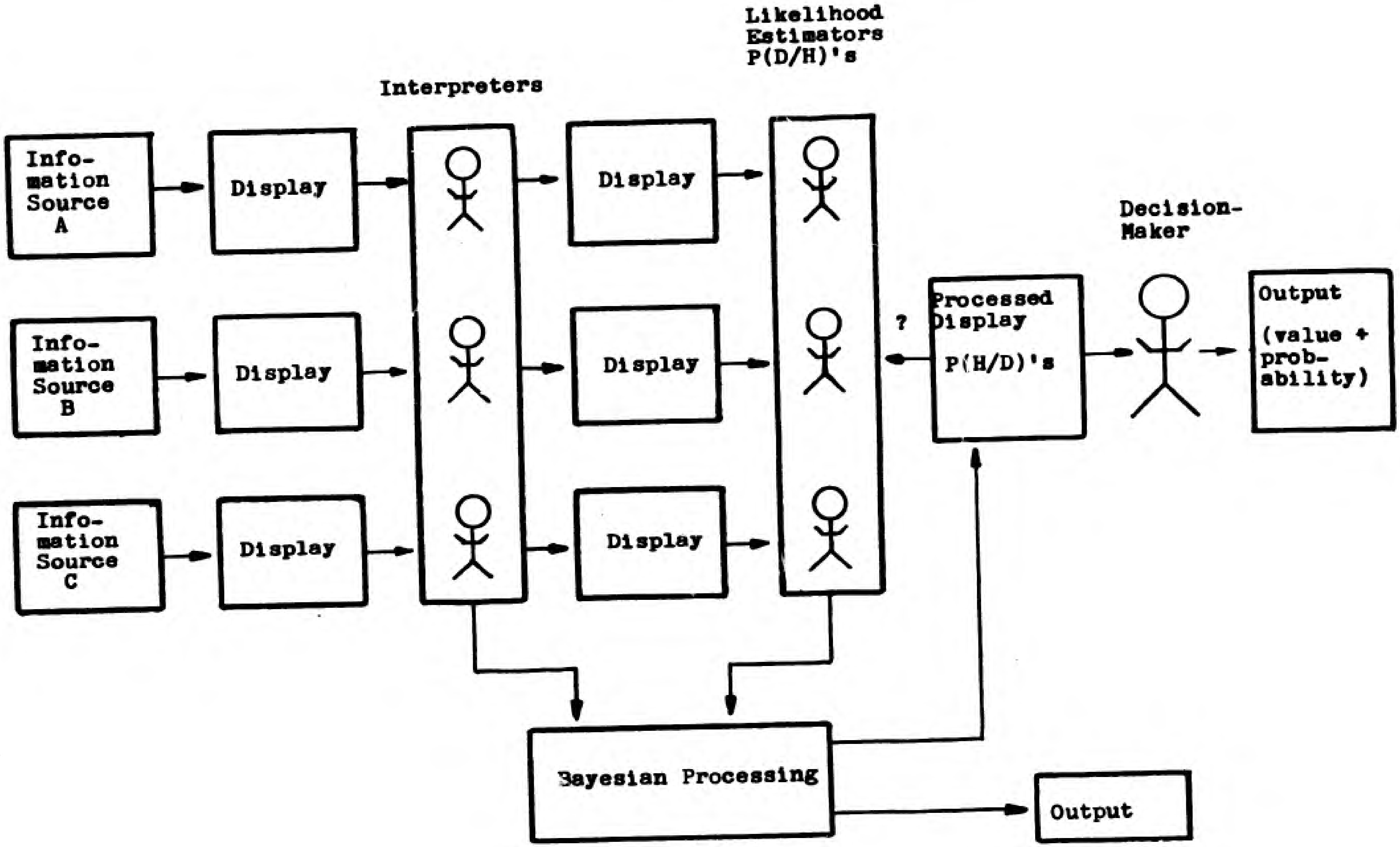

Figure 1. A Probabilistic Information Processing system, PIP for short (taken from Edwards et al., 1968, p. 253).

China) and their prior probabilities, P(H), are established. Information coming into the system is then filtered and interpreted by experts specifically trained for this task. This data is then passed on to experts who have been trained to make likelihood estimations, P(D/H), for these data. These P(D/H)'s are then processed by a computer using Bayes' theorem. A display of the final posterior probabilities, P(H/D) over all given hypotheses (H), is then made available to the decision-maker to aid him in making his final decision.

Comparison of Various Decision-Making Systems (Questions 4 and 5)

Independent Data. Most simulation studies of military decision-making, using independent data, have compared a PIP system with other decision-making systems. Edwards compared four information processing systems in a large scale simulation study with a political-military setting (Edwards, 1966b; Edwards et al., 1968). Subjects well-trained in the history of a "world of 1975", which supposedly consisted of six major nations, were asked to make inferences about six predetermined hypotheses (e.g. Russia is about to attack the UAR; Peace will prevail) on the basis of information received from three sources: radar, reconnaissance satellites, and intelligence. These "experts" were given sixty items of information, one at a time. After seeing each item, the experts were asked to make one of a number of types of probability inference, depending upon which type of information processing system was being investigated:

1. PIP experts were asked to estimate likelihood ratios, always comparing one of five war hypetheses with the sixth hypothesis which was always "peace will prevail".

2. POP experts were asked to estimate posterior odds, always comparing one of five war hypotheses with the sixth hypothesis which was always "peace will prevail".

3. PEP experts, who would be penalized if the data turned out to favor a war hypothesis, were asked to decide on a fair price for an insurance policy that would pay the penalty for them. There was no penalty for peace. This system was assumed to resemble decision-making in the military as it exists today.

4. PUP experts simply were asked to estimate posterior probabilities for each hypothesis, rather than posterior odds.

Results showed that while all four systems gave very similar qualitative results (i.e. all favored the same hypothesis, r = .85), quantitatively, the PIP system was superior to POP, PEP, and PUP. That is, PIP extracted much more certainty from the data and favored the appropriate hypothesis earlier in the data sequence than the other systems did. For example, when PUP final odds were 1:5, PIP's were 1:32.8; when PEP final odds were 1:5, PIP's were 1:35.3; and when POP final odds were 1:5, PIP's were 1:12.4. Edwards also found that if PIP experts are given feedback about the current posterior probabilities for each hypothesis, their performance was impaired, possibly because the experts were being swayed in the direction of estimating P(H/D) and not P(D/H). Thus feedback may not be desirable in a PIP system, as the question mark in figure 1 indicates.

Kaplan and Newman (1966) also compared a PIP system [i.e. men estimate P(D/H)] with a POP system [i.e. men estimate P(H/D)] by having "experts" try to detect an enemy's strategy by observing bombs falling on certain targets. They performed three different experiments. In the first experiment, they varied the certainty of the data and no feedback was given. They found the PIP system to be more efficient in that it gave higher posterior probabilities for the correct hypothesis, and reached an asymptote at a faster rate. They also found that as data became less diagnostic (i.e. the difficulty of the task increased), POP performance was reduced while PIP performance remained unaffected. In their second experiment, they added both an information purchasing system and a feedback system. The result of this change was to show no difference in PIP or POP performance. Both systems performed very poorly. That there was no difference may be due to the fact that the task was too difficult; also the feedback system may have turned the PIP task into a POP task. Their final experiment again showed PIP to be superior to POP, but this difference in performance decreased as the amount of information given to the experts increased. In all three experiments both PIP and POP were inferior to ideal performance (i.e. these experimenters had objective P(D/H)'s which, when put into Bayes' theorem, gave optimal posterior probabilities).

Schum and his associates have done a considerable amount of research in military decision-making. Briggs and Schum (1965) found a PIP system to be better than a POP system. However, as they decreased the fidelity (i.e. diagnosticity) of the data, they found the POP system to be better except when fidelity was very low, in which case the PIP system did better. This is contrary to the findings of Kaplan and Newman mentioned earlier. However, feedback was always given by Briggs and Schum, so this may be part of the reason why their PIP system showed poor performance.

Dependent Data. In all of the experiments discussed so far, all items of data received by the experts were independent of each other. Schum was also interested in seeing if subjects would be sensitive to items of data that were dependent, or what he calls conditionally nonindependent data. For example, suppose one has two items of data, D_1 and D_2. The problem is to determine the probability of a hypothesis (H) given D_1 and D_2. If the two data are dependent, the estimate of P(H) may be greater than it would be if the two data were independent of each other. By having subjects estimate P(H/D)'s in a military diagnostic situation when given information about conditional nonindependencies, Schum (1965) has shown men are capable of taking these dependencies into account. In fact, when compared to optimal performance (i.e. Bayes' theorem for conditionally nonindependent data), Schum found little conservatism. Schum (1966) has also shown that when the amount of available information increases, whether dependent or independent, prior probabilities become relatively unimportant.

Different Evaluation Techniques. Schum (1967a) has discussed several ways of evaluating probabilistic information processing systems. Four evaluative measures have generally been used by investigators in this area and these measures all have their good and bad points (Schum, 1967a). One can compare different information processing systems in the following ways:

1. Number of times the correct hypothesis is chosen.
2. Magnitude of the final posterior probabilities.
3. Accuracy Ratio (AR) - ratio of the system's log likelihood estimates to optimal log likelihood estimates as given by Bayes' theorem.
4. Difference Measure (DM) - difference between the system's final posterior probability estimates and optimal posterior probability estimates as given by Bayes' theorem.

Using these various measures, Schum, Southard, and Womboldt (1968) also evaluated a semi-PIP and POP system in a military diagnostic situation. (For details of the basic experimental method see Schum, 1967b). In a semi-PIP system experts are asked to estimate one P(D/H) for several items of information given simultaneously. Pairs of conditional nonindependence existed among the items of information, and all subjects had access to information about these conditional nonindependencies. Three specific experiments were performed. In the first, experts were given six items of information either one at a time, three at a time, or all six at once. Thus "sample size" was varied. Results showed POP to be unaffected by sample size while PIP performance got worse as sample size increased. PIP performed better than POP when sample size equaled one. Again the results are contrary to those of Kaplan and Newman who found differences between PIP and POP to decrease when the amount of information increases. Schum et al. (1968) feel the POP superiority with the six item samples may be due to the fact that their subjects had had considerable experience with probability estimation. Thus they may have been doing six quick revisions rather than treating all the items as one datum. In a second experiment, varying the amount of information, semi-PIP always showed smaller DM scores than POP, although DM scores for both systems increased with the amount of information. In a final experiment where the diagnosticity of the information and the amount of information were varied independently, it was found that a semi-PIP system is always superior to a POP system, especially when the information is highly diagnostic, abundant, or both. In one case, semi-PIP produced final posterior odds four times greater than those produced by POP.

Problems Involved With Use of PIP Systems in the Real World (Questions 5 and 6)

The evidence reviewed here appears to favor a PIP system, at least in a military decision-making context. However, there are some problems with diagnostic tasks in the real world that have not been studied extensively in a "laboratory PIP" system, so applying PIP systems elsewhere should be done with caution (e.g. Schum, 1968). There are problems in the real world of defining an exhaustive set of mutually exclusive hypotheses, nonindependence of data (either pairs or higher orders), reliability of data, and nonstationarity of data (i.e. P(D/H) may change with time). Specific sequencing of data may also be important. Edwards (1966b, p. 76) has found that, at least when estimating posterior probabilities, P(H/D), early data exerts more influence than later data. However, Kaplan and Newman (1966) have found that when subjects estimate p(D/H), they do not appear to be influenced by prior information, so it is possible that specific sequencing is not important in a PIP system but it may be in a POP system. Schum (1968) has also suggested a semi-PIP system may be more successful in making use of conditional nonindependencies and weakly diagnostic data. Both Edwards and Schum have repeatedly pointed out the importance of training individuals to use a PIP-type system. Most "experts" in the above mentioned experiments would not be considered similar to experts in the real world who may have been on the job for 20 years or longer. Real world experts have much more experience and possibly a better feeling for the impact of the information. Because of this, there may be less of a difference between PIP and POP systems in the real world. The crucial part of training experts to function in a PIP system lies in teaching them what a P(D/H) is. Even if experts understand likelihoods, it is very easy for them to slip back into estimating the probability of the hypothesis, P(H/D). Also even if experts are denied access to the current state of affairs among the hypotheses, one might assume that they might keep track in their head.

WEATHER FORECASTING

Economic Value (Question 1)

Most meteorologists and users of weather forecasts will agree that there are sound economic reasons for reporting weather forecasts in terms of probability statements (e.g. Thompson and Brier, 1955; Malone, 1957). A hypothetical but descriptive example given by Malone (1957, pp. 156-157) serves to illustrate this: Imagine a construction company faced with the problem of deciding whether or not to pour concrete each day. If the concrete is poured and .15 of an inch of rain falls in the subsequent 36 hours, $5,000 damage will result. If, on the other hand, the newly poured concrete were protected from such rainfall, the cost of these protective measures would be $400. If no protective measures were taken, say, for one winter season, the cost would average around $85,000. If protective measures were taken everyday, the cost would be approximately $72,800. If protective measures were taken only when the probability of rain was .50 (the ordinary type of forecast), the cost would average $32,600. If, however, one had a probability forecast of rain, the cost could be reduced even further. The problem is to select the probability level that will minimize the total expense. According to Malone, this can be done using the principle of the calculated risk, which prescribes that protective measures should be taken only when $P > C/L$, where P is some probability of the critical amount of rain falling within 36 hours of the time the concrete was poured; C is the cost of protective measures; and L is the contingent loss. Thus in the illustration just given, C = $400; L = $5,000, therefore P = 400/5000 = .08. If protective measures are taken only when the probability of rain exceeds .08, the total cost would average only $24,400. The same sort of paradigm can be applied to dispatching and cancelling commercial aircraft, evacuation of aircraft from a military base when a tornado or hurricane threatens, and scheduling stand-bys or overtime crews for telephone line maintenance if a thunderstorm should occur (Malone, 1957).

Can meteorologists provide realistic probability statements about meteorological events? Some investigators doubt that they can (e.g. Dexter, 1962), but most existing research shows that they are indeed capable of providing realistic statements (Root, 1962; Sanders, 1963). In fact, Epstein (1962) has suggested using Bayes' theorem in obtaining information about future weather events.

Objective Versus Subjective Probability Estimates (Question 2)

Which is best, objective or subjective probability forecasts, where the former are usually based on climatological expectancies? These climatological expectancies are really relative frequencies, i.e. given the initial state of the atmosphere, what has been the occurence of each of the available meteorological events in the past for this state? Subjective forecasts, on the other hand, utilize the forecaster's skill in arriving at a probability estimate for a meteorological event. Often the forecaster uses the climatological expectancy as an initial probability estimate (i.e. his prior probability) and "sharpens" this estimate by using any additional information available to him. Both Root (1962) and Sanders (1963) found subjective probability estimates to be better than objective probability estimates for a variety of meteorological events (e.g. occurence of precipitation, wind speed, visibility, temperature, etc.). However, Sanders (1963) has found that forecasters have little or no skill in making probability statements about meteorological events which have extremely high or low climatological expectancies (i.e. probability below .10 or above .90).

Individual Versus Consensus Probability Estimates (Question 3)

Sanders (1963) has given some data to show that group-mean probability estimates, his group consisting of 12 meteorological students, are more accurate than single estimates of any one particular student, even the best one. Even when two instructors were asked to give estimates, the mean probability estimates for the two were better than the estimates of either instructor alone. Therefore it appears that there is some value in "consensus forecasting".

Computers and Meteorological Data (Questions 4 and 5)

I have been unable to find any reference to the use of a PIP-type system in meteorology. Computers aid mostly in data collection. Given that forecasters have an abundance of meteorological data, a PIP system might easily help them aggregate all the data more efficiently. There also appears to be no data comparing Bayesian-produced probability estimates with those of the meteorologist to see if conservatism exists in weather forecasting.

Very little reference is made to the types of data used by meteorologists. It is not known, for example, whether more optimal predictions are obtained with data displayed simultaneously or with data displayed sequentially. No description of how meteorologists weigh various types of data in arriving at their final probability estimates is given in the studies reviewed. Malone (1957) discusses some of the problems involved in the prediction of meteorological events. For example, data used is often incomplete in that there are many technical difficulties involved in obtaining information about the initial state of the atmosphere. Also there are problems in that the atmosphere is inherently unstable at times. Sanders (1963) has found that forecasts for a particular area are more accurate when more valid synoptic (i.e. immediate weather conditions for a broad area) information about that area is made available (e.g. weather ships on the North Atlantic Ocean). Forecasts are more accurate for some meteorological events than others because of the information available (e.g. wind direction is easier to forecast than wind speed).

Sanders (1963) has also found spot forecasts (i.e. forecasts for a specific instant of time) to be less accurate than period forecasts (i.e. forecasts for a longer period of time), and both Sanders (1963) and Root (1962) have found a decrease in forecast ability as the range of the forecast period is extended (e.g. from a 24-hour to a 48-hour period). H$_o$wever, Root reports that in spite of this, subjective forecasts are still better than climatological expectancies for any given length of forecast period.

Expert Forecasting (Question 6)

How can one become an expert forecaster? Perhaps by using Bayes' theorem as suggested by Epstein (1962). Another way would be by using evaluative methods such as scoring rules, which encourage a meteorologist to be honest, i.e. to make his assessments correspond to his judgements (e.g. Murphy and Epstein, 1967; Winkler and Murphy, 1968). These scoring rules have the effect of motivating the forecaster to minimize or maximize his score (depending upon which rule is used), especially if a system of payoffs is attached to the score. Scoring rules also aid forecasters in becoming better assessors because they allow for comparison of scores among different forecasters, and, as a result, can point out unreasonable biases held by forecasters which may be hindering their performance.

Does a great deal of experience in probability forecasting lead to better performance? Evidently not, according to Sanders (1963), who found that students and instructors in meteorology perform about equally well, although students tend to overforecast the probability of occurence of an event, possibly because of their previous academic experience in which the results of laboratory experiments were usually positive. Sanders also found, however, that when synoptic information offers little concrete guidance, forecasters tend to remain close to the climatological expectancy and use this value as their probability estimate. If the synoptic information offers a gain over prediction based solely on climatological expectancies, some forecasters grasp the implications more firmly than others. Thus, experience of some sort may help. There appears to be very little information about what kinds of experience aid the forecaster. Nor is there much information about how to train forecasters to weigh data differentially, etc.

Finally it might be mentioned that Epstein (1962) has shown that the different types of prior distributions held by forecasters are essentially unimportant as long as subsequent data used to revise prior opinions is available. Therefore training experts so that they have accurate prior probability distributions may not be necessary.

MEDICAL DIAGNOSIS

Economic Value (Question 1)

Sir William Osler once said, "Medicine is a science of uncertainty and an art of probability."[1] This statement is more descriptive of medical diagnosis than prognosis. There is no doubt that use of probabilistic statements in medical diagnosis has economic value. In some cases, waiting for absolute certainty about a diagnosis (if it exists at all) before initiating treatment may lead to the death of a patient. In other cases, if probabilistic statements are taken into account, one might reduce the number of laboratory tests, etc., needed for diagnosis, thereby reducing time, cost, and discomfort to both the patient and the doctor. In fact, probability theory can act as the basis for setting up an efficient screening program where the more probable cases can be referred to a specialist for further examination before a final diagnosis is made.

[1] From Bean, W. B. (ed.) Aphorisms From His Beadside Teaching and Writings. N. Y.: Schuman, 1950, p.125.

Public Models and Subjective Probabilities (Question 2)

Are subjective or objective probabilities more effective in medical diagnosis? It should be pointed out that often all the medical diagnostician has to go on is his own subjective probabilities. For rare or uncommon diseases, it is very difficult to determine any kind of reliable objective probabilities (i.e. relative frequencies). Lusted (1968) refers to the objective probabilities which exist as "public models". Using terms we have already discussed, objective prior probabilities, P(H), would be the incidence of each disease in a given sample. P(D/H)'s would be the incidence of a symptom (D) given a particular disease (H). These public models are simply relative frequencies and can be derived from a sample of medical records of previous cases for any particular disease. These relative fraquencies can then be arranged into a symptom-disease matrix, which is essentially the public model Lusted is talking about. For illustrative purposes, a symptom-disease matrix for primary bone tumors is given in figure 2. Often problems arise because public models do not exist, or because there is disagreement among diagnosticians as to the usefulness of public models since the sample from which the model was developed may not be random or representative.

In spite of these disagreements, Winkler, Reichertz, and Kloss (1967) have shown that there can be agreement between objective P(D/H)'s. These investigators found great similarity between P(D/H) estimates for symptoms of hyperthyroidism derived from a sample taken in West Germany and those derived from a sample taken in Florida. Lodwick (1966) has shown that when probability values taken from public models are modified by personal experience (i.e. so that they now become subjective probabilities), there is an increase of approx-

Tumor Type	P(H)	Age			Location						Size		Matrix		Grade				
(Hypotheses)	Incidence	P(D/H)			P(D/H)		P(D/H)				P(D/H)		P(D/H)		P(D/H)				
		up to 21	21 - 30	31 +	small & flat	long	epiphysis	plate	metaphysis	shaft	1-6 cm.	6 + cm.	bone	cartilage	I - A	I - B	I - C	II	III
Giant cell tumor	15	20	35	45	20	80	99	1	100	20	40	60	1	0	15	35	50	0	0
Chondroblastoma	5	75	20	5	20	80	100	50	75	0	90	10	0	30	50	35	15	0	0
Chondromyxoidfibroma	3	50	35	15	30	70	30	20	100	25	85	15	0	2	85	15	1	0	0
Chondrosarcoma	17	25	25	50	35	65	40	1	85	65	20	80	5	65	15	20	25	25	15
Fibrosarcoma	10	20	20	60	5	95	55	0	90	65	20	80	25	2	0	10	40	30	20
Osteosarcoma	25	65	25	10	10	90	30	5	95	75	15	85	98	5	0	0	10	30	60
Parostealsarcoma	5	20	35	45	0	100	30	1	100	50	25	75	100	5	15	25	55	5	0
Ewing's tumor	15	70	25	5	35	65	20	5	85	90	15	85	0	0	0	5	10	20	65
Reticulum cell	5	10	25	65	20	80	50	1	85	80	15	85	0	0	0	0	20	30	50

Figure 2. Symptom-disease matrix for primary bone tumors. Each number reflects the probable number of cases in which a finding would be present in a sample of 100 cases. Taken from Lodwick, G. S., "A probabilistic approach to the diagnosis of bone tumors," *Radiol*. *Clin*. *N*. *Amer*., 1963, *3*, 487-497; modified slightly.

imately 5% in correct diagnosis of bone tumors. This suggests that the best estimates to use in Bayesian decision-making may be the objective P(D/H)'s, modified by the individual diagnostician who takes his own specialized knowledge about local or special conditions into account. As Overall and Williams (1963) point out, "Too much time is required for the individual to acquire experience which is an adequate basis for reliable subjective probability estimates." Thus, development of public models may help reduce some of the problems faced by medical diagnosticians.

Individual Versus Consensus Diagnosis (Question 3)

There appears to be no literature on whether individual or group probability estimates are more profitable. Part of the reason for this may be that it is impractical and too costly to have five specialists diagnose the same patient. Edwards, Cox, and Garland (1962) have suggested that determination of whether or not exploratory thoracotomy should be undertaken on patients suspected of having a malignant coin lesion should be decided by a conference of three types of specialists (i.e. radiologists, surgeons, and internists). This procedure should help even out biases held by each of the specialists. However, no suggestions are made as to how this conference should arrive at a final decision.

Use of Computers and Bayes' Theorem in the Diagnostic Process (Question 4)

As far as I can tell, Ledley and Lusted (1959) introduced Bayesian decision-making to medical diagnosis. They also suggested that computers should be used to make the long, tedious calculations required by Bayes' theorem. Of course, computers can aid medical diagnosis in many other ways such as interviewing patients, storing medical histories, analyzing radiographic film, monitoring physiological signs of patients at the bedside or in the operating room, and so on. For illustrations of these uses the reader is referred to

Tolles (1964) and Earle (1966). However, this section will be concerned only with the role of computers in helping the diagnostician form his opinion once he has all the available information. Diagnosis of quite a few diseases with the help of Bayes' theorem and a computer has been attempted. Several of these studies will be summarized briefly.

Heart Disease. Several investigators have studied diagnosis of heart disease using Bayes' theorem to aid them (e.g. Bruce, 1963; Warner et al., 1964; Bruce and Yarnall, 1966; Templeton et al., 1966). Warner et al. (1964), using a matrix of 53 symptoms and 35 diseases, compared physician and computer diagnosis of congenital heart disease. In their study, physicians were asked to fill out a symptom check-list made up of mutually exclusive symptoms. They did this by indicating whether a given symptom was present, absent, or uncertain. Then the physician was asked to make a diagnosis. The check-list information of each physician was also input to a computer which already had a symptom-disease matrix, i.e. P(H)'s and P(D/H)'s, stored in its memory. The computer then made a diagnosis on the basis of all this information using Bayes' theorem. On the average, the computer diagnoses were better than the physicians'. However, Warner et al. also found that the physicians improved in accuracy with time. They attributed this greater accuracy to the physician's gradually increasing experience in preparing data for the computer and in receiving feedback from the computer in the form of differential diagnosis.

Bruce (1963) and Bruce and Yarnall (1966) suggest that Bayes' theorem and computer diagnosis may be less effective than the Warner et al. (1964) results imply. For example, in diagnosis of valvular heart disease, Bayes' theorem came up with the correct diagnosis in only 45% of the cases; in diagnosis of congenital heart disease, 86%

of the cases were diagnosed correctly. However, these investigators also suggested some possible reasons for this poor performance and these will be discussed later.

Templeton et al. (1966), in comparing diagnosis of congenital heart disease by radiologists and by computers, obtained results similar to Warner et al. (1964). Using 231 cases whose correct diagnoses were confirmed by autopsy and surgery, and a technique similar to that of Warner et al., they found that radiologists correctly diagnosed approximately 79% of the cases,^ while the computer correctly diagnosed 78% of the cases. These investigators also pointed out that since a radiologist must accurately identify radiographic information so it can be input to a computer, he may reassess and reorganize his pattern recognition process so he becomes more precise at identifying relevant data. This in turn, will aid both the physician and the computer in making a more accurate diagnosis.

<u>Bone Tumors</u>. Lodwick and his co-workers (1965) have developed a general classification of bone tumors, using descriptions and photographs for illustrative and diagnostic purposes. On the basis of this classification system, they have set up a symptom-disease matrix which includes both P(H) and P(D/H) for nine types of bone tumors. When check-list data (e.g. age, radiographic information, etc.) was input to a computer, which already had stored in its memory information from the symptom-disease matrix, Lodwick et al. (1966) found that out of 76 cases, the computer correctly diagnosed 85.5% of them while the diagnostician, without the aid of the computer, diagnosed only 80% of the cases correctly.

<u>Thyroid Disease</u>. Overall and Williams (1963) and Fitzgerald et al. (1966) have investigated computer diagnosis of thyroid disease. They have set up a computer program which revises the computer's symptom-disease matrix whenever a new case is diagnosed by the computer.

Thus the computer "learns" as the physician would, modifying its P(H)'s and P(D/H)'s with "experience". Overall and Williams (1963), using a 21 symptom - 3 disease matrix, found the compuber to diagnose correctly 258 out of 268 cases (96%) of thyroid diseases. If analysis is made using only subjective symptoms (i.e. laboratory data not used) the computer then diagnosed 88% of the cases correctly.

Cushing's Disease. Nugent (1964) has used Bayes' theorem to diagnose Cushing's disease. Using a total of 11 symptoms which differentiated between individuals having Cushing's disease and those not having it, he had a computer diagnose 211 cases, 52 of these cases having Cushing's disease and the remaining 159 not having Cushing's disease. Ninety-five of the non-disease cases were given a posterior probability of .01 of having the disease, while 19 of the disease cases were given a posterior probability of .99 or higher. The 11 symptoms used were based on simple clinical data. For the cases where the posterior probabilities were less differentiating, i.e. .3 to .6, one might want to give more detailed, expensive tests before making a final diagnosis.

Epigastric Pain. Rinaldo, Scheinck, and Rupe (1963) input a computer with an 8 symptom - 6 disease matrix based on 204 cases of epigastric pain. Again the symptoms were all subjective, i.e. no laboratory data was used. The computer then analyzed 96 cases of epigastric pain. The percentage of correct predictions for the six diseases were 73, 69, 27, 75, 38, and 33%. For some diseases this is not bad given that diagnosis was made only on the basis of 8 subjective symptoms. The investigators suggested that one of the reasons for the low percentage of correct diagnoses was due to the variability of the data received from the patients, thus making it difficult to construct a reliable symptom-disease matrix.

Computer Versus Physician Diagnosis. From the results of the studies mentioned above, it appears that computers can function as effectively as the physician in many diagnostic situations. Some studies did not make comparisons between computer diagnosis and physician diagnosis so it is not known, for example, if the general physician does a better job of diagnosis than the computer or if the specialist does a better job of diagnosis than the computer. Given that the computer can do an effective job, the physician might spend more time trying to obtain more precise and reliable data, etc., which would then increase the accuracy of computer diagnosis.

Apparently no studies have been made using a PIP system in medical diagnosis where the diagnostician estimates P(D/H) or P(H/D). In most of the studies reviewed, posterior probabilities were derived by comparing the probability of one disease to the probability of all other diseases considered or no disease. Likelihood ratios in the form of odds for one disease to another disease were rarely estimated. Naturally if one has 35 diseases and 53 symptoms like Warner et al. (1964), estimation of likelihood ratios in this manner could be a very time-consuming process. Lusted (1968, pp. 163-168) has suggested experimental PIP programs for diagnosis of primary bone tumors and congenital heart diseases, as well as for selection of optimum treatment for any given disease.

It should be pointed out that most of the studies cited were concerned with having the physician diagnose the correct disease rather than having him give a posterior probability distribution for the entire set of diseases being considered. Therefore, although physicians and computers do equally well in diagnosing the correct disease, it is impossible to tell whether the physician is more conservative than the computer. Lodwick et al. (1966) point out that the computer will give posterior probability estimates as high as .99 or as low as .01, and it is doubtful that physicians would come up with such extreme estimates. It is interesting to note that Edwards, Cox, and Garland (1962) found that when specialists showed a high degree of preoperative diagnostic ability in predicting whether a coin lesion was benign or malignant, they appeared to be unaware of their ability to diagnose correctly, because they usually recommended surgical removal of the tumor for further diagnostic purposes. Thus we might consider them conservative information processors. However, it should be remembered that the important task of medical diagnosis is that the correct diagnosis be reached rather than attaching any sort of magnitude estimation to the diagnosis.

Problems Involved in Use of Medical Data (Question 5)

Independence of Symptoms. Consider the data used in medical diagnosis. Data fed to the computer has to be obtained from the physician who usually records it on a symptom check-list. This data can be obtained both by interviewing the patient (e.g. age, subjective symptoms such as headaches, shortness of breath, etc.) and by performing certain laboratory tests. One problem that arises is whether the data, or what might be called symptoms, used in diagnosis are independent. In several of the studies mentioned, the investigators used only those symptoms that were independent. Use of Bayes' theorem requires that symptoms be independent, unless, for example, conditional nonindependencies (e.g. Schum, 1965) are taken into account when estimating likelihoods. In spite of this problem, Overall and Williams (1963) have found that Bayes' theorem appears to work well, even when some dependent symptoms are treated as independent symptoms.

Defining a Set of Diseases. Another problem with using Bayes' theorem is in establishing an exhaustive set of mutually exclusive diseases. Here it may be advisable to have a "residual" category where rare or unknown diseases are all lumped together. If two diseases are not completely independent, one can merely regard them as a third disease. This may be necessary, for example, when one particular symptom is absent when an individual has either one disease (D_1) or another disease (D_2) but not both. However, if an individual has both diseases ($D_1 + D_2$) at the same time, then the symptom will be present.

Complete Versus Partial Information in the Diagnostic Process. Obtaining medical data can be a costly venture for the patient and a time-consuming process for the diagnostician. Gorry and Barnett (1968) investigated the possibility of balancing the risk of making a diagnosis against the cost of further testing and the value of the evidence which is obtained. Using data from Warner et al. (1964), they put P(H)'s and P(D/H)'s into a computer along with the costs of tests and misdiagnoses, the latter being rather difficult to establish. They then compared sequential and complete diagnostic accuracy. They found no difference in accuracy between diagnosis made on the basis of complete information and diagnosis made on the basis of information received sequentially until it became too expensive to obtain more information. This is quite remarkable considering that 31 pieces of information were used in the complete diagnosis, while only an average of 6.9 pieces of information were needed for sequential diagnosis.

Other Factors Influencing Diagnostic Accuracy. Mount and Evans (1963) simulated a medical diagnostic situation using Bayes' theorem. Two of their findings are relevant here. First, they found that there is an improvement in the percentage of correct diagnoses as the number statistically independent symptoms used in the diagnostic process is increased. Second, they found that as the sample size from which the symptom-disease matrix, P(H) and P(D/H), is constructed is increased, accuracy of diagnosis increases up to a point and then begins to decrease slightly. One problem with increasing sample size, however, is that dependencies among symptoms may begin to appear. If so, one can simply reanalyze the data into a new symptom-disease matrix, either eliminating dependent symptoms or taking these dependent relations into account when estimating P(D/H)'s.

Establishing a reliable symptom-disease matrix can be difficult. Bruce (1963), for example, prepared two symptom-disease matrices, one for acquired valvular heart disease based on 170 cases and one for congenital heart disease based on 124 cases. When the computer diagnosed these same cases using Bayes' theorem and the established symptom-disease matrix, it correctly diagnosed 97% of the valvular heart disease cases and 98% of the congenital heart disease cases. However, when the same symptom-disease matrix was used to analyze cases which were not part of the sample used to construct the matrix, the computer did rather poorly. In diagnosing 119 new cases of valvular heart disease, the computer was correct only 45% of the time; for 76 new cases of congenital heart disease, it was correct 86% of the time. Bruce feels that part of this poor performance was due to the paucity of specific symptoms for the new cases.

<u>Expertise and the Education of Medical Students (Question 6)</u>

Finally we come to the problem of training a doctor to become a good diagnostician. Expertise in a special area appears to increase accuracy in diagnosis. For example, Gustafson (1963) compared the diagnosis of congenital heart disease made by a computer, pediatric cardiologists, and non-specialized physicians. The pediatric cardiologists and the computer appeared to be about equally accurate, correctly diagnosing 63-74% of the cases, while the non-specialized physicians were less accurate, correctly diagnosing only 36-52% of the cases. Gustafson suggests that computers can be valuable aids to physicians with diagnostic problems outside their area of special training.

The basic problem in training the diagnostician appears to be in making him aware of P(D/H). Developing prior probabilities does not appear to be that important (e.g. Lodwick et al., 1966), especially if the amount of available information is large. One way of "teaching" the to-be diagnostician P(D/H) would be to give him a symptom-disease matrix to study or memorize. Using this symptom-disease matrix as a starting point, he can modify it once he has had some diagnostic experience. Another way of teaching P(D/H) to the to-be diagnostician would be through use of computer instruction techniques. These techniques get the student personally involved in the diagnostic process and the experiences the student obtains from them more nearly resemble those which he would obtain out in the real world. Of course, another major problem of training the diagnostician is in teaching him to recognize symptoms, etc., but this problem is beyond the scope of this paper. A very interesting example of a hypothetical medical case presented to a student by a computer is reproduced in Lusted (1968, pp. 80-87).

Several investigators have suggested use of computer diagnosis using Bayes' theorem as a teaching or learning device, even for specialists (e.g. Bruce, 1963; Warner et al., 1964; Lodwick et al., 1966). That is, by having specialists fill out a check-list of symptoms and signs, they are forced to focus more systematically on what they are doing. Receiving feedback from the computer then helps the specialists reanalyze their diagnostic process so they may become even more accurate.

CONCLUDING REMARKS

The literature reviewed here favors the use of subjective probabilities in making diagnoses and decisions, whether it be a military decision, a weather forecast, or a medical diagnosis. Subjective probabilities can be measured, and individuals are quite able to make accurate estimates of the probabilities of events that they have experienced, although they do not always revise their estimates in an unbiased manner. Moreover, when the implications of a decision are great, they require more than the objectively required amount of information before they are willing to make a decision. In this sense, men are conservative information processors.

It is true that in most everyday decisions conservatism may not exert much influence and can therefore be considered negligible. However in some areas of applied decision-making, such as those that have been reviewed in this paper, elimination of this conservatism may have great value in terms of saving both lives and money.

Systems like PIP, that apply Bayes' theorem to quantified expert opinion, clearly show that this approach is profitable. In some cases it was shown that Bayesian techniques were superior to currently used techniques. Even when there is no difference, use of Bayesian techniques in a man-machine system would be valuable in that it would free the expert from many routine tasks. He could then spend more time trying to develop more reliable data, which would, in turn, result in a more optimal decision-making system.

While a great deal of experimentation using Bayesian techniques has gone on in the laboratory, much more research needs to be done out in the "real world". PIP systems could be developed for weather forecasting and medical diagnosis. These systems could then be tried out in the real world and their usefulness evaluated. In areas where several expert opinions are available, decision-makers could experiment with Winkler's methods for combining the opinions of several individuals (Winkler, 1968). In doing so they might find a method which would aid them in arriving at more nearly optimal decisions in that area. Bayesian decision-making techniques can also be applied to other areas where decisions must be made on the basis of probabilistic information such as in business and the buying and selling of stock (e.g. Schlaifer, 1961; Green, Halbert, & Minas, 1964; Green, 1967; Riter, 1967). In fact something similar to the climatological expectancies used by meteorologists or the public models used by physicians could be developed for business organizations.

In conclusion, the application of Bayesian techniques to areas which require decisions to be made in the face of uncertainty can be beneficial. Using Bayesian techniques, one can set up a structured decision-making program for any specific area; the results of this program will then provide feedback that can be used to evaluate the program. Experience and information gained from this evaluation can then be used to re-design certain aspects of the program to make it function as efficiently and as near to optimal as possible.

Glossary

For those who are interested in the meanings of some of the technical terms mentioned in this paper, the following definitions are provided.

Climatological expectancies - relative frequencies or averages of weather events for a given period of time.

Climatology - a science which deals with climates and their phenomena. Note: Climate = the average course or condition of the weather at a place over a period of time.

Coin lesion - a rounded coin-like tumor .

Cushing's disease - a disease characterized by obesity, especially of head, neck, and trunk, brownish streaks on the abdominal wall, and muscular weakness; associated with dysfunction of the adrenal cortex or the anterior lobe of the pituitary gland.

Epigastric pain - pain in the upper middle portion of the abdomen, over or in front of the stomach.

Exploratory thoracotomy - surgical incision of the wall of the chest for exploratory purposes.

Hyperthyroidism - excessive activity of the thyroid gland characterized by increased basal metabolism, protrusion of the eyeball, and disturbances in the vegetative nervous system.

Meteorology - the science which deals with the atmosphere and its phenomena; more specifically it deals with weather and weather forecasting.

Pediatric cardiologist - a physician who specializes in the study and treatment of heart disease in children.

Radiograph - a photograph taken with roentgen rays, i.e. x-rays or gamma-rays.

Synoptic - relating to or displaying atmospheric and weather conditions as they exist simultaneously over a broad area.

References

Attneave, F., "Psychological probabilities as a function of experienced frequency," J. Expt'l. Psych., 1953, 46, 81-86.

Beach, L. R., "Accuracy and consistency in the revision of subjective probabilities," IEEE Trans. on Human Factors in Electronics, 1966, HFE-7, 29-37.

Beach, L. R., "Multiple regression as a model for human information utilization," Organ. Behav. and Human Perf., 1967, 2, 276-289.

Beach, L. R. and Peterson, C. R., "Subjective probabilities for unions of events," Psychon. Sci., 1966, 5, 307-308.

Beach, L. R. and Phillips, L. D., "Subjective probabilities inferred from estimates and bets," J. Expt'l. Psych., 1967, 75, 354-359.

Beach, L. R. and Wise, J. A., "Subjective probability and decision strategy," J. Expt'l. Psych., 1969, 79, 133-138.

Briggs, G. E. and Schum, D. A., "Automated Bayesian hypothesis selection in a simulated threat-diagnosis system," in J. Spiegel and D. Walker (eds.) Second Congress of Information System Sciences, Washington D. C.: Spartan Books, 1965, 169-176.

Bruce, R. A., "Computer diagnosis of heart disease," Proc. 5th IBM Med. Symp., 1963, 77-98.

Bruce, R. A. and Yarnall, S. P., "Computer-aided diagnosis of cardiovascular disorders," J. Chronic Diseases, 1966, 19, 473-484.

deFinetti, B., "Methods for discriminating levels of partial knowledge concerning a test item," British J. of Math. and Stat. Psych., 1965, 18, 87-123.

Dexter, R., "Confidence factors are fictional," Weather,1962, 17, 132-135.

Earle, D. P. (ed), "Symposium: Computers in Medicine", J. Chronic Diseases, 1966, 19, 1-539.

Edwards, W., "The prediction of decisions among bets," J. Expt'l. Psych., 1955, 50, 201-214.

Edwards, W., "Introduction to special issue on revision of opinions by men and man-machine systems," IEEE Trans. on Human Factors in Electronics, 1966a, HFE-7, 1 - 6.

Edwards, W., Nonconservative Probabilistic Information Processing Systems, (ESD-TR-66-404), Institute of Science and Technology, University of Michigan, Ann Arbor, December, 1966b.

Edwards, W., "Conservatism in human information processing," in B. Kleinmuntz (ed.), Formal Representation of Human Judgement, N. Y.: Wiley, 1968, 17-52.

Edwards, W. and Slovic, P., "Seeking information to reduce the risk of decisions," Amer. J. Psych., 1965, 78, 188-197.

Edwards, W. and Tyversky, A., Decision-Making: Selected Readings, Harmondsworth, Middlesex, England: Penguin Books, 1967.

Edwards, W., Lindman, H., and Phillips, L. D., "Emerging technologies making decisions," in New Directions in Psychology II, N. Y.: Holt, Rinehart, and Winston, 1965, 265-325.

Edwards, W., Lindman, H., and Savage, L. J., "Bayesian statistical inference for psychological research," Psych. Rev., 1963, 70, 193-242.

Edwards, W., Phillips, L. D., Hays, W. L., and Goodman, B. C., "Probabilistic information processing systems: design and evaluation," IEEE Trans. on Systems Science and Cybernetics, 1968, SSC-4, 248-265.

Edwards, W. M., Cox, R. S., Jr., and Garland, L. H., "The solitary nodule (coin lesion) of the lung," Amer. J. Roentgen., 1962, 88, 1020-1042.

Epstein, E. S., "A Bayesian approach to decision making in applied meteorology," J. Applied Meteor., 1962, 1, 169-177.

Fitzgerald, L. T., Overall, J. E., and Williams, C. M., "A computer program for diagnosis of thyroid disease," Amer. J. Roentgen., 1966, 97, 901-905.

Fried, L. S. and Peterson, C. R., "Information seeking: optional versus fixed stopping," J. Expt'l. Psych., 1969, 80, 525-529.

Good, I. J., "Rational decisions," J. Royal Stat. Soc.- Series B, 1952, 14, 107-114.

Gorry, G. A. and Barnett, G. O., "Sequential diagnosis by computer," J. Amer. Med. Assoc., 1968, 205, 849-854.

Green, P. E., "A behavioral experiment in the economics of information," in G. Fisk (ed.), The Psychology of Management Decision, Sweden: CWK Gleerup Publishers, 1967, 170-185.

Green, P. E., Halbert, M. H., and Minas, J. S., "An experiment in information buying," J. of Advertising Research, 1964, 4, 17-23.

Gustafson, J. E., "The computer for use in private practice," Proc. 5th IBM Med. Symp., 1963, 101-111.

Hoffman, P. J., "Cue-consistency and configurality in human judgement," in B. Kleinmuntz (ed.), *Formal Representation of Human Judgement*, N. Y.: Wiley, 1968, 53-90.

Kaplan, R. J. and Newman, J. R., "Studies in probabilistic information processing," *IEEE Trans. on Human Factors in Electronics*, 1966, *HFE-7*, 49-63.

Ledley, R. S. and Lusted, L. B., "Reasoning foundations of medical diagnosis," *Science*, 1959, *130*, 9-21.

Lodwick, G. S., "A probabilistic approach to the diagnosis of bone tumors, " *The Radiologic Clinics of N. Amer.*, 1965, *3*, 487-497.

Lodwick, G. S., "Computer-aided diagnosis in radiology. A research plan," *Invest. Radiol.*, 1966, *1*, 72-80.

Lodwick, G. S., Turner, A. H., Jr., Lusted, L. B., and Templeton, A. W., "Computer-aided analysis of radiographic images," *J. Chronic Diseases*, 1966, *19*, 485-496.

Lusted, L. B., *Introduction to Medical Decision Making*, Springfield, Ill.: Thomas, 1968.

Malone, T. F., "Applied meteorology," *Meteor. Research Reviews: Meteor. Monographs*, 1957, *3*, 152-159.

Morris, W. T., *Management Science: A Bayesian Introduction*, Englewood Cliffs: Prentice Hall, 1968.

Mount, J. F. and Evans, J. W., "Computer-aided diagnosis - a simulation study," *Proc. 5th IBM Med. Symp.*, 1963, 113-127.

Murphy, A. H. and Epstein, E. S., "Verification of probabilistic predictions: a brief review," *J. Applied Meteor.*, 1967, *6*, 748-755.

Nugent, C., "The diagnosis of Cushing's disease," in J. A. Jacquez (ed.), *The Diagnostic Process*, Ann Arbor: Malloy Lithographing, Inc., 1964, 185-197.

Overall, J. E. and Williams, C. M., "Conditional probability program for diagnosis of thyroid function," *J. Amer. Med. Assoc.*, 1963, *183*, 307-313.

Peterson, C. R. and Beach, L. R., "Man as an intuitive statistician," *Psych. Bull.*, 1967, *68*, 29-46.

Peterson, C. R. and DuCharme, W. M., "A primacy effect in subjective probability revision," *J. Expt'l. Psych.*, 1967, *73*, 61-65.

Peterson, C. R. and Swensson, R. G., "Intuitive statistical inferences about diffuse hypotheses," *Organ. Behav. and Human Perf.*, 1968, *3*, 1-11.

Peterson, C. R., DuCharme, W. M., and Edwards, W., "Sampling distributions and probability revisions," *J. Expt'l. Psych.*, 1968, *76*, 236-243.

Peterson, C. R., Schneider, R. J., and Miller, A. J., "Sample size and the revision of subjective probabilities," *J. Expt'l. Psych.*, 1965, *69*, 522-527.

Peterson, C. R., Ulehla, Z. J., Miller, A. J., Bourne, L. E., Jr., and Stilson, D. W., "Internal consistency of subjective probabilities," *J. Expt'l. Psych.*, 1965, *70*, 526-533.

Phillips, L., and Edwards, W., "Conservatism in a simple probability inference task," *J. Expt'l. Psych.*, 1966, *72*, 346-354.

Pitz, G. F., "Information seeking when available information is limited," *J. Expt'l Psych.*, 1968, *76*, 25-34.

Pitz, G. F., Reinhold, H., and Geller, E. S., "Strategies of information seeking in deferred decision making," *Organ. Behav. and Human Perf.*, 1969, *4*, 1-19.

Preston, M. G. and Baratta, P., "An experimental study of the auction-value of an uncertain outcome," *Amer. J. Psych.*, 1948, *61*, 183-193.

Rinaldo, J. A., Jr., Scheinok, P., and Rupe, C. E., "Symptom diagnosis: a mathematical analysis of epigastric pain," *Annals of Internal Med.*, 1963, *59*, 145-154.

Riter, C. B., "The merchandising decision under uncertainty," *J. of Marketing*, 1967, *31*, 44-47.

Root, H. E., "Probability statements in weather forecasting," *J. Applied Meteor.*, 1962, *1*, 163-167.

Sanders, F., "On subjective probability forecasting," *J. Applied Meteor.*, 1963, *2*, 191-201.

Schlaifer, R., *Introduction to Statistics for Business Decisions*, N. Y.: McGraw-Hill, 1961.

Schum, D. A., *Inferences on the Basis of Conditionally Nonindependent Data*, AMRL Technical Report 65-161, Aerospace Medical Research Lab, Wright Patterson Air Force Base, Ohio, December, 1965.

Schum, D. A., "Prior uncertainty and amount of diagnostic evidence as variables in a probabilistic inference task," *Organ. Behav. and Human Perf.*, 1966, *1*, 31-54.

Schum, D. A., "Concerning the evaluation and aggregation of probabilistic evidence by man-machine systems," in D. E. Walker (ed.) *Information System Science and Technology*, Washington D. C.: Thompson Books, 1967a, 337-347.

Schum, D. A., Concerning the Simulation of Diagnostic Systems Which Process Complex Probabilistic Evidence Sets, Rice University Interdisciplinary Program in Applied Mathematics and Systems Theory, report 46-1, September 12, 1967b.

Schum, D. A., Behavioral Decision Theory and Man-Machine Systems, Rice University Systems 46-4, Interdisciplinary Program in Applied Mathematics and Systems Theory, July 15, 1968.

Schum, D. A. and Martin, D. W., "Human processing of inconclusive evidence from multinomial probability distributions," Organ. Behav. and Human Perf., 1968, 3, 353-365.

Schum, D. A., Southard, J. F., and Womboldt, L., Aided Human Processing of Inconclusive Evidence in Diagnostic Systems: A Summary of Experimental Evaluations, Rice University Interdisciplinary Program in Applied Mathematics and Systems Theory, Systems 46-3, February 1, 1968.

Shuford, E. H., A Comparison of Subjective Probabilities for Elementary and Compound Events, Report # 20, University of North Carolina Psychometric Lab., 1959.

Templeton, A. W., Lehr, J. L., and Simmons, C., "The computer evaluation and diagnosis of congenital heart disease, using roentgenographic findings," Radiology, 1966, 87, 658-670.

Thompson, J. C. and Brier, G. W., "The economic utility of weather forecasts," Monthly Weather Rev., 1955, 83, 249-254.

Tolles, W. E. (ed.), "Computers in medicine and biology," Annals of the N. Y. Acad. of Sci., 1964, 115, 543-1140.

Warner, H. R., Toronto, A. F., and Veasy, L. G., "Experience with Bayes' theorem for computer diagnosis of congenital heart disease," Annals of the N. Y. Acad. of Sci., 1964, 115, 558-567.

Wheeler, G. and Beach, L. R., "Subjective sampling distributions and conservatism," Organ. Behav. and Human Perf., 1968, 3, 36-46.

Winkler, C., Reichertz, P., and Kloss, G., "Computer diagnosis of thyroid disease: comparison of incidence and consideration of the problem of data collection," Amer. J. Med. Sci.,1967, 253, 27-34.

Winkler, R. L., "The quantification of judgement: some methodological suggestions," J. Amer. Stat. Assoc. , 1967a, 62, 1105-1120.

Winkler, R. L., "The quantification of judgement: some experimental results," Proc. of the Amer. Stat. Assoc., 1967b, 386-395.

Winkler, R. L., "The consensus of subjective probability distributions," Management Science, 1968, 15, B61-B75.

Winkler, R. L. and Murphy, A. H., "Good probability assessors," *J. Applied Meteor.*, 1968, *7*, 751-758.

Yntema, D. B. and Torgerson, W. S., "Man-computer cooperation in decisions requiring common sense," *IRE Trans. on Human Factors in Electronics*, 1961, *HFE-2*, 20-26.