It used to be that a good physician could assimilate, retain, and recall most of the known facts about medicine. Beginning with the years prior to World War II, it became evident that this was no longer possible. The systemizations to condense facts in other fields had not progressed as far as had the accumulation of facts in medicine. Thus, we have seen the emergence of the "medical specialist" and the "team approach" to disease. Even this multiple physician approach is beginning to fail before the exponentially increasing array of information about pathological processes.

I believe that there are just three things that can be done about this overwhelming wealth of information: (1) We can develop more specialists that are even more specialized, but already this approach is being hampered by problems of communication. (2) We can develop more encompassing theories of disease so as to reduce the large number of facts to a relatively few simple hypotheses. This is the goal of the model builders, perhaps the ideal approach, but we cannot afford to wait for this nirvana. And even if this were possible now we would only reach a temporary plateau upon which new mountains of data would pile. (3) The third approach is to utilize such mechanical and electronic slaves as are available to help us organize, retain, recall, and communicate those observations on disease worthy of record. I believe we
are forced to develop this third approach while evolving the best compromise for the first approach and pursuing the second with all possible vigor.

In order to utilize the latest engineering achievements we must be very clear about what instructions we give. If we are not, we may find ourselves in the position of the “sorcerer’s apprentice” who failed to learn how to turn off the water; only here we will have stacks and stacks of meaningless paper.

It is sensible to examine first how the human computer works when making medical diagnoses (or any kind of inductive or scientific inference) by discussing hypotheses which have attempted to describe this process. The machinery should be taught how to imitate the human diagnostician. Perhaps then we can find ways to improve processes when coupling the machine and human brain.


The discussion is divided into two sections. The first section deals with the problem of construction of a scheme of classification of disease entities. This may be termed the problem of classification. The second section deals with the credibility of a diagnosis after a patient has been assigned to a disease entity. The degree of credence attached to possible assignments may be used as the basis of assignment, and hence the basis of medical diagnosis itself. I will term this the problem of credence.

The Problem of Classification

It is convenient to refer to the state of a patient at a particular instant of time. Let us suppose that at such some instant a patient (or normal individual) may be completely characterized by the concomitant values of a sufficiently large number of variables. Some of these variables, such as sex, are constant throughout life; some, such as height and weight, change relatively slowly. Others, such as blood cholesterol, vary dramatically at different times. Some are nearly constant because of feedback control mechanisms. Some are periodic. Some strikingly reflect impacts from the environment. Some are random. Many are interrelated in complex fashions, and their nature is sought by the model builders. Whatever are the characteristics of the various variables, however, the set of values applying to a sufficiently large collection of variables uniquely characterizes the state of the individual at the particular point in time. It may be helpful to think in geometric terms. Suppose each variable is the axis of a geometrical space. If there are n variables there will be n axes for our space, and we will have an n-dimensional space. (It will not harm our concept to visualize a two-dimensional space with, for example, the first axis, \( x_1 \), equal to the weight of the individual, and the second axis, \( x_2 \), equal to the systolic blood pressure of the individual.) Now, at a particular instant of time, the set of values for the n variables will determine a single point in the state space. This single point will be called the state point of the individual at the particular instant. It is easy to visualize that in an instant the point can shift slightly from its original position. So, through life, from the moment of birth (or earlier) to the moment of death, the individual will be uniquely described by a succession of adjacent state points. Imagining these points strung together we have a line of state points twisting and bending through the state space from birth to death. This is indeed an abstract view of a patient, as a line in state space, a life line. One may immediately object to this “coldly mechanized” view. However, there is no reason in principle why some of the n variables cannot represent the emotional and affective states of the individual at each instant. Theoretically, every human experience and feeling can be represented as values on the axes, or on combinations of the axes, in the state space.

Suppose there is a line in state space for every person in the world, a bundle of more than three billion lines! As the course of life is somewhat similar for all of us, the life lines will have parallel tendencies although no two lines will be identical (possible but improbable). Typical individuals will lie toward the center of the bundle, atypical ones toward the outside. Clearly, life lines will not exist in all parts of the state space. The dictates of life are such that the living mechanism will not function in all possible states. Thus, very extreme life lines will not exist. Possible but still extreme life lines will occur rarely, whereas mild, atypical lines will occur much more frequently in this conception. (This central tendency of the life lines is predicted by the “central limit theorems” of mathematical probability, and is confirmed by common experience; the mathematical function most used to describe the density of lines at various distances from the center of the bundle is termed the normal or Gaussian distribution.)

Some of the lines may represent lives which at times have more than negligible malfunction. Then the individual is diseased. Satellite bundles of diseased life lines occur with new centers of density. If it is clear (sufficiently low density of lines between regions of high density) that these satellite clusters are not fortuitous irregularities in the tail of the normal density, then the satellites themselves are recognized as distinct disease entities and are appropriately named. Sometimes partial tails of the normal density function are taken to be disease entities (although not distinct) when the respective states represent some degree of malfunction or pathology. It is useful to distinguish these two patterns of disease.

In the past, recognition of disease pattern has been largely heuristic. Now, powerful quantitative tools exist for aiding this process. Of particular interest is the generalized measure of distance (squared) between cluster centers, due to Mahalanobis (1930, 1936) and known as Mahalanobis' D^2. The central idea in the use of D^2 is the measure of the distance (squared) between cluster centers, taking into account the functional dependencies between state variables. If the distance between centers is large, compared to the scatter about the centers, then distinct disease entities are recognized. Statistical tests of significance help to
distinguish real from accidental clustering. The method of "discriminant functions," due to Fisher (1938), and the "generalized T² test," due to Hotelling (1931), are mathematically equivalent procedures to D². These and some other procedures with similar objectives are frequently referred to as "cluster analysis." An over-all view of the rationale, mathematical derivation, and uses of these procedures is found in Rao (1952). The elementary discussions in Shephard and Turner (1959) and Hanna, Turner, and Hughes (1963) may be helpful.

Measurements of distances between cluster centers may be made for various fixed ages yielding a "distance function" of age. Alternatively, adjustments of the states for age may be made by replacing observed states with corresponding (sliding up or down the average life line) states at some age. The principles of "covariance analysis" are appropriate here.

There is one final consideration about choice of procedures for cluster analysis before we pass on to the problem of credence. The D²-T² discriminant function procedure is based upon one rather restrictive assumption about the equality of scatter, and interdependencies between variables, about two centers which we wish to measure the distance between. This assumption often is not even approximately true when comparing normal and diseased life line bundles. In this case, generalized procedures are available (Kendall, 1957), although they have not been used widely.

The Problem of Credence

Suppose we have divided the state space into a set of not necessarily mutually exclusive regions recognized as disease entities plus the "normal" region. It is immaterial whether informal or formal procedures were used in arriving at the regions. We will take the regions to be fixed for purposes of application of the ideas of this section. Let us realize, however, that these regions will be rearranged at times as information about the state space accumulates. Further suppose that a physician has observations corresponding to the values of some of the state variables. At this point he arrives at a provisional diagnosis (i.e., he assigns the "patient" to one of the regions in state space). But this diagnosis suffers from uncertainty due to

at least two causes: (1) his information is incomplete as he cannot measure all state variables, and (2) those measurements he has (signs, symptoms, tests, etc.) contain intrinsic errors of a random or systematic nature, due either to physiological variation or to measurement error. The physician now decides whether to take more measurements (new measurements or replications of old ones), to begin treatment based on his provisional diagnosis, or both. His behavior in these two important respects is predicated largely upon his belief in his own diagnosis. Thus there is the problem of how best to measure and reason about the subjective phenomenon, credibility.

We will relate credibility to probability by first examining some concepts of probability. The notion of mathematical probability first arose in the Italian Renaissance as a theory of repetitive happenings which was applied to games of chance and even to life insurance. The philosophical and mathematical bases of the theory of probability were subject to much dispute until the purely mathematical aspects of the theory were abstracted (cf. Kolmogorov, 1956). In this modern guise the essential ideas of the theory of probability can be simply stated. We consider the set of possible outcomes of an experiment. Call these results A₁, A₂, · · ·, Aₙ. Suppose B is another kind of result of the same experiment. We will let A,B stand for the event, "both A₁ and B happen." We will let A₁ ∪ B stand for the event, "either A₁ or B, or both A₁ and B happen." We will let S stand for the event which must happen, and O stand for the event which cannot happen. Then if we write A₁A₂ = O we imply that both A₁ and A₂ cannot both happen. Or if we write A₁ ∪ B = S we imply that either A₁ or B must happen. Now the theory of probability concerns certain real numbers which are assigned to each possible experimental result and are called "probabilities." Thus, we will write p(A₁) and read, "the probability that A₁ happens," or write p(A₁ ∪ B) and read "the probability that either A₁ or B happens," and so forth. It is important to realize that the theory of probability itself does not provide prescriptions for assigning the probabilities. These prescriptions must be obtained from other considerations. However, the probabilities must satisfy three restrictions (called the axioms of probability): (1) p(A₁) ≥ 0 where A₁ is any result, (2) p(S) = 1, and (3) if A₁, A₂ = O then p(A₁ ∪ A₂) = p(A₁) + p(A₂). This is all we need to establish from the theorems of the theory of probability. For example, we can derive that p(O) = 0, that 0 ≤ p(A₁) ≤ 1, that p(A₁ ∪ B) = p(A₁) + p(B) - p(A₁B), and many more. Before proceeding we will need to make one further definition. Let p(B | A₁) = p(A₁B) / p(A₁) and read p(B | A₁) as "the probability that B will happen given that A₁ has already happened," or "the probability of B given A₁," for short. Then we say that A₁ and B are independent if p(B | A₁) = p(B). If B and A₁ are independent then we see that p(A₁B) = p(A₁)p(B), the famous rule of multiplication for independent events.

It would be easy to demonstrate the truth of a very remarkable formula discovered by Thomas Bayes (1763) and now known as Bayes' Theorem. This formula can be written:

\[ p(A₁ | B) = \frac{p(A₁)p(B | A₁)}{p(B)} \]

where p(B) = p(A₁)p(B | A₁) + p(A₂)p(B | A₂) + · · · p(Aₙ)p(B | Aₙ), and supposing that A₁, A₂, · · ·, Aₙ are mutually exclusive events.

Probabilities have to do with the frequency of occurrence of possible outcomes of an experiment. Let us put aside all thoughts about probabilities and think about a set of possible hypotheses, H₁, H₂, · · ·, Hₙ, to explain some observed phenomenon. Suppose we would like to measure the credence we place in each hypothesis. What restrictions should we impose upon our measure? It has been suggested (cf. Polya, 1954) that rational humans behave as though their credences (write C(H₁), C(H₂), etc., for real measures) obeyed the following three restrictions: (1) C(H₁) ≥ 0 where H₁ is any hypothesis, (2) C(S) = 1 where S = H₁ ∪ H₂ ∪ · · · ∪ Hₙ, and (3) if H₁H₂ = O then C(H₁ ∪ H₂) = C(H₁) + C(H₂). Restriction (1) says that the measure of credibility which we will use is never negative. Restriction (2) says that the credence in at least one hypothesis is assigned the numerical quantity one. Finally, restriction (3) says that if two hypotheses cannot both be right then the degree of credence to be placed upon the compound hypothesis "either H₁ or H₂" is simply the sum of the respective individual credences.
The theory of credibility is identical in mathematical content to the theory of probability, although the purposes of the two theories are quite different. But since they are mathematically equivalent, any theorem of probability can be taken over for credence theory, and in fact, there is no logical reason why we cannot mix probabilities and credences in any valid formula derived from the axioms of probability. For example, the following mixed version of Bayes' Theorem is perfectly valid:

\[ C(H_1 \mid B) = C(H_1)p(B \mid H_1)/p(B). \]

This formula may be interpreted to say that if one wants to calculate the credence to be placed in hypothesis number 1, given the observations \( B \), then we need to know two things: (1) the credence placed in hypothesis number 1 before \( B \) was observed, and (2) the probability that \( B \) would be observed if hypothesis number 1 were true. Having similar information for all alternative hypotheses will allow computation of the denominator. This is a remarkable result because it provides a complete solution to the problem of assigning credences to various hypotheses or diagnoses in light of any given observations.

The key to using the mixed version of Bayes' Theorem for measuring or comparing credence in alternative diagnoses is in the source of the prior credences, \( C(H_i) \), \( C(H_2) \), etc. We consider four different situations.

1. **Prior credences estimated as relative frequencies of disease entity in a particular population.** Sometimes it is possible to estimate how often each disease entity occurs in a population from which a current patient was drawn at random. Such relative frequencies then may be used as proper measures of the prior credence.

2. **Prior credences locally uniform.** The posterior credence, \( C(H_i \mid B) \) will not be much affected by \( C(H_i) \), the prior credence, if there is sufficient information in the observations \( B \). This situation can be ensured by increasing the quality and quantity of the observations (more examinations, tests, etc.).

3. **Prior credences subjective.** The physician may not have formal information of the type encountered in situation 1 but may have strong, intuitively developed measures of prior
credence based upon experience. Numerical evaluation of these credences can be evoked but can lead to dangerous conclusions. Polya (1954) and others warn against attempting it; however, there is no doubt that all practicing physicians act as though they were making such an evaluation.

4. Minimax prior credences. Consider just two competing diagnoses, $H_1$ and $H_2$. Suppose the physician would take a certain action if the patient had disease entity number 1 and another action if the patient had disease entity number 2. What would be the loss to the patient if the wrong action were taken? We can choose prior credences so that we minimize the maximum loss to the patient. This approach necessitates very strong observational information before the physician will depart from the “conservative action.” The idea applies as well to more than two possible diagnoses.

Most of the current attempts to use Bayes’ Theorem with electronic computers to aid in making medical diagnoses involve situation 1 or 2. Thus, by situation 1 we replace $C(H_1)$, $C(H_3)$, etc. by observed relative frequencies of the respective disease entities, or by situation 2 we set the prior credences equal to each other; that is, $C(H_1) = C(H_3) = \cdots = C(H_n) = 1/k$. In either case we still need to know the second factors in the mixed Bayes’ Theorem, namely $p(B | H_1)$, $p(B | H_2)$, etc. We recall that these factors represent the probabilities of observing the set of signs, symptoms, and tests, given that a particular diagnosis is correct. In current applications these usually are empirically determined from the same population as are the prior credences of situation 1. That is, these probabilities are replaced by the relative frequencies of particular sign, symptom, and test configurations in the various diagnostic cluster groups.

Let us suppose we wish to compare two competing diagnoses. Let us form the ratio of $p(B | H_1)$ to $p(B | H_3)$. When $B$ has been observed, this ratio is termed the likelihood ratio (LR). By rearranging Bayes’ Theorem we have:

$$LR = \frac{p(B | H_1)}{p(B | H_3)} = \frac{C(H_1 | B)}{C(H_3 | B)} = \frac{C(H_1 | B)/C(H_1)}{C(H_3 | B)/C(H_3)}$$

We then see that the likelihood ratio amounts to a comparison of the posterior-to-prior credence ratios for the two diagnoses. If this LR is large we wish to favor $H_1$, or if it is small we might wish to favor $H_2$. The LR is the basis for the discriminant function techniques mentioned in the last section, and is the principal idea underlying the procedures adopted in medical diagnosis by Neyman (1947, 1950) and Collen et al. (1964).

Alternatively, one could employ directly the posterior credence ratio (CR) given by

$$CR = C(H_1 | B)/C(H_2 | B)$$

That is, if CR is large we would favor $H_1$, but if CR is small we would favor $H_2$. In order to compute CR we need to specify the prior credences as well as the likelihoods. This can be done by appeal to any one of the four situations enumerated. In particular, situation 1 has been considered. In the case of situation 2, the $LR = CR$ and this has often been used to justify the LR method.

Conclusions

Armed with an overwhelming accumulation of data about disease, how can we ensure that they will all be employed effectively to make a correct diagnosis in a particular patient? The use of electronic computers can be of some help in the collation, correlation, storage, and communication of the accumulated information, but we must be careful in instructing the machinery so we will not one day find a monster whose behavior is unpredictable. A reasonable procedure would be to analyse our own thought processes carefully to ascertain how the human diagnostician arrives at his conclusions. The matter is certainly not settled but the concepts of state spaces and the theory of credences seem to form a plausible “first model” of the human inference maker at work making medical diagnoses. It is hoped that a wider appreciation of these ideas will lead to the construction of better models that could enable the great potential of the “computer age” to have its full impact upon medical care.

References

Arnois, D. C., F. N. Silverman, and M. E. Turner. The radiographic evaluation of pulmonary vasculature in children with congenital cardio-


... What then is a good experiment? It is that which informs us of something besides an isolated fact; it is that which enables us to foresee, that is, that which enables us to generalize.

“For without generalization foreknowledge is impossible. The circumstances under which one has worked will never reproduce themselves all at once. The observed action then will never recur; the only thing that can be affirmed is that under analogous circumstances an analogous action will be produced. In order to foresee, then, it is necessary to invoke at least analogy, that is to say, already then to generalize....

“... Thus, thanks to generalization, each fact observed enables us to foresee a great many others; only we must not forget that the first alone is certain, that all others are merely probable. No matter how solidly founded a prediction may appear to us, we are never absolutely sure that experiment will not contradict it, if we undertake to verify it. The probability, however, is often so great that practically we may be content with it. It is far better to foresee even without certainty than not to foresee at all.”