STATISTICAL PROBLEMS IN MEDICAL DIAGNOSES

C. L. CHIANG, J. L. HODGES, JR., AND J. YERUSHALMY
UNIVERSITY OF CALIFORNIA

1. Introduction

The purpose of this report is to present some comments on a number of loosely related problems in the field of medical diagnosis which have some implications in statistical theory. The application of statistical theory to medical diagnosis is relatively new and only the simpler questions have been treated in a rigorous manner thus far. It was suggested that rather than select a single topic and provide a solution to a single well-specified problem, it would be more useful and interesting at such a symposium as this to touch on a number of matters and to formulate, at least partially, the statistical problems which they involve. Our report, therefore, will be concerned with questions and concepts rather than answers; it will be statistical in the broad sense, without being at all mathematical. We shall feel that our efforts are well repaid if some of these problems are found to be of sufficient interest to stimulate research or experimentation.

2. The uses of a diagnostic aid

It might be well first to delineate the area with which we are concerned. This does not, and in fact cannot, embrace the entire subject of clinical diagnosis. The latter is a complex operation and consists in identifying a disease process through a number of different operations, such as the obtaining of an individual history and a familial history, the results of physical examination and the findings of a number of different tests, including roentgenographic examinations and a host of different laboratory tests. The evaluation of diagnosis in this broad sense has not been, and possibly cannot be, the subject of a statistical analysis. Instead, we shall consider only that phase of the process of diagnosis which has been the subject of statistical evaluation and is related to what we shall term a *diagnostic aid:* that is, the findings obtained by the application of a certain diagnostic procedure as an aid in this complex process of diagnosis. Some examples are blood counts, the evaluation of a chest X-ray film, blood pressure measurement, reaction to a skin test, urinalysis, etc.

It may be well to consider first the different uses to which the findings of a given diagnostic aid are put since they would lead to different kinds of statistical problems. A diagnostic aid may be employed in at least four different ways:

(1) As an aid to the physician in the clinical diagnosis of a case. This relates primarily to the process of differential diagnosis, that is, to differentiate between several possible diagnostic categories, all of which may have common symp-

This investigation was supported (in part) by a research grant from the National Institutes of Health, Public Health Service.

tomatology but which can sometimes be distinguished by results of different tests. For example, cancer of the lung and pulmonary tuberculosis could have the same symptomatology and they can sometimes be distinguished by exfoliative cytologic methods.

(2) In case-finding programs where a diagnostic aid is used as a screening device of an apparently healthy population to select persons who are suspected of having a specific disease. For example, the population group may be subjected to an X-ray examination for the purpose of detecting those who may have tuberculosis. Recently population groups are being examined not only by one diagnostic aid, but simultaneously by a number of tests in a so-called multiphasic screening test.

(3) In the application to a population group for the purpose of estimating prevalence rates of a given disease or condition in the population. An example is the testing of a population group with tuberculin to determine the infection rate.

(4) In the periodic examination of a group of people for the purpose of early detection of deviations from normality in one or more characteristics. For example, workers in a factory exposed to special hazards of radiation may be subject to specific laboratory tests to determine the effect of radiation on their health.

In all of these, it will be noticed that we are interested in the results of the tests as one element which may be used in conjunction with other findings. The implication is, and that is essential, that there are other ways in which the disease can be diagnosed. Otherwise, there would be no means of validating the diagnostic aid, though its reliability would still be subject to study.[1] In general, there must exist another means of diagnosis which is considered more valid than is the aid itself. One may, therefore, inquire why the aid is used at all. There are several reasons:

(1) The aid may provide an additional and at least partly independent piece of evidence for the clinician to consider in making a differential diagnosis.

(2) A diagnostic aid may be employed in a mass survey for economic reasons: it may be much cheaper, and also may be more convenient, than the more valid diagnostic procedures. The cheap test can be applied on a wide and major scale to identify a certain number of suspects, to each of whom a more expensive and more certain test can be applied.

(3) In recent years, a much more important reason for using different diagnostic tests came to the fore. This has to do with recent trends in public health. As is known, great changes have occurred in the last few decades in the state of the public health and in the specific problems which it presents. As a result of extensive activity in the field, many of the diseases of an acute nature are greatly reduced. Consequently, many of the chronic diseases have assumed much greater importance because a larger proportion of the population attain the older ages to which these diseases are more common. The prevention of these diseases of long duration is much more complicated than prevention of the acute diseases, and it is generally assumed that it depends to a large extent on early detection. In addition, most of the chronic diseases have a silent period, that is, in the early stages they present

[1] The method of Neyman [7] for making inferences about the distribution in a population of the probability of positive response to a diagnostic aid is, in our terminology, concerned with the reliability of the aid in the sense that it is concerned with the ability of the aid to give consistent results when repeatedly applied to an individual. However, an aid may be highly reliable without there being any relation between its results and the actual disease it is supposed to diagnose; in which case it would not be valid. No examination restricted to the results of applying a diagnostic aid can establish its validity, as was pointed out by Neyman, p. 1456.

no symptoms and are not brought to the awareness of the person who has the disease. It is, therefore, a problem of detecting possible candidates for a disease where no other symptoms are present. For these reasons, there has been in recent years a search for diagnostic aids to detect specific diseases at a stage so early that more certain diagnosis (for example, by clinical symptoms) is impossible.

Incidentally, we note that while a program of mass screening with a diagnostic aid is largely justified by appeal to the plausible thesis that early detection is of value in curing or curbing the process of the disease in the affected individual, it is very difficult to find scientifically sound evidence to prove this thesis. In fact, there is rather heated controversy on the matter. (For an example, see the discussion of the value of early detection of cancer in [1].) In one sense, it is necessary that a diagnostic aid be put into mass use before its use can be proved of value, since it is only by mass screening that it is possible to detect a large number of "early cases." We are only now beginning to reach this stage of development, but two related questions which will arise can perhaps be foreseen:

Are individuals, free of symptoms but responsive to a diagnostic aid, really early cases of the disease in question, or do they merely represent a lack of specificity of the aid? Will early treatment really be successful in preventing the development of the symptomatic phase of the disease? The reader will see how these questions interact on each other. If an individual who responds to the aid is placed under treatment and then never shows symptoms of illness, how can we be sure whether (a) he was destined to become ill but was saved by the treatment or (b) he would not have become ill even if not treated? This question is at the heart of the criticism of the program for early detection of cancer: some critics maintaining that it is conceivable that many of the cures claimed for early surgery represent misdiagnosis of benign tumors. It seems clear that these questions will have to be faced if mass surveys are to be justified, and that many difficult statistical problems will arise in settling them. These problems are, however, not our concern here, and we shall postulate that early detection is of value in preventing the disease in question.

3. Selection of a critical value

Consider a diagnostic aid which produces a reading x with a continuous range. In the healthy population x will usually be characterized by a "bell-shaped" distribution, more or less normal in form. In a group of frankly sick persons, x will have a distribution offset from the healthy distribution—let us say to the right. Customarily there will be an overlap of the two distributions, representing values which may be possessed either by healthy persons with high normal readings, or by diseased persons with low sick readings. When an individual is found to have such an x-value, we are perplexed in classifying him.

In using a diagnostic aid in mass screening, we are faced with what is in essence a two-decision problem for each individual: he is either let through the screen or he is not (though of course in the latter case, a variety of actions may be available). It is customary to select a so-called critical value for x, say x_o, and to let through the screen those individuals for whom $x < x_o$. It has been suggested [2] that x_o should be determined by the requirement that the false positive rate α not exceed 5%, and that an aid is acceptable for screening use only if the resulting false negative rate is $\gamma \leqq 10\%$.

We believe that this type of suggestion reflects a tendency to apply to diagnostic problems habits of statistical thinking acquired in other, and quite different, areas. We can, as pointed out by Berkson [8], view the use of the diagnostic aid as a test of the hypothesis that the individual is healthy, so that a false positive diagnosis corresponds to the false rejection of the hypothesis being tested, or type I error. From this point of view, the use of $\alpha \leqq 5\%$ is sanctified by long custom. Similarly, a false negative is an instance of type II error, and $\gamma = 10\%$ corresponds to the familiar power level of 0.9. The introduction that the notion that two types of error are involved in a diagnosis, and the consideration of their probabilities, marked an essential step in the statistical analysis of diagnostic aids [8]. However, this approach does not take into account what we feel to be essential elements of the mass survey problem: the prevalence of the disease in the population, and the vital necessity of keeping within reasonable limits the ratio of false positives to true positives.

If we denote by π the proportion of affected individuals in the population, then π may be viewed as the *a priori* probability that an individual drawn from the population is diseased. As mentioned, the estimation of π may be a primary aim of a mass survey, but usually a guess for π is available in advance. The existence of an *a priori* probability that the hypothesis is false makes the testing problem one of the Bayes type. Using Bayes' formula, we calculate the *a posteriori* probability η that a positively diagnosed individual is actually diseased and we have

$$\eta = \frac{\pi(1 - \gamma)}{\pi(1 - \gamma) + (1 - \pi)\alpha} . \tag{1}$$

It is clear that for fixed α and γ, η tends to 0 as π tends to 0. For example, if we set $\alpha = 0.05$ and $\gamma = 0.1$, $\eta = 18\pi/(1 + 17\pi)$. Thus if $\pi = 0.01$ (1% prevalence of the disease), the aid will give us a positively diagnosed group of whom 15% are really diseased. But an aid with the same error probabilities applied to a rarer disease with $\pi = 0.001$ would produce a concentration of only 1.8% of truly diseased, or $56\frac{1}{2}$ false positives for each true positive.

Not enough attention has been paid to the unfortunate consequences of a low value of η. Suppose that a large number of people are bothered unnecessarily for each truly positive person discovered. Each person is subjected to considerable expense and mental trauma once he is notified that the results of the test are positive. If at the end of much expenditure of effort and money he is finally told that he really was negative all the time, his feeling of relief often expresses itself as indignation over the trouble to which he has been exposed, and results in loud criticism of the survey program. In many cases a high proportion of false positives among the positively diagnosed may eventually defeat an otherwise worthwhile program by engendering public opposition. It is for this reason that the value of η will be a very critical one, and certainly it is unrealistic to fix an arbitrary proportion of false positives without taking into account the value of π.

One might experiment with desired values for η. It is obvious that it must be a function of π. We cannot expect the same final concentration as a result of a screening test with a disease which is very rare as with a more common disease. We would of course require that η be a function of π lying between π and 1. For example,

one might postulate that the *a priori* concentration η equal $\sqrt{\pi}$. This means that on a logarithmic scale we shall have advanced by screening halfway from the original concentration toward certainty. Under such a requirement, it can be shown that

$$\eta = \sqrt{\pi} = \frac{\alpha}{1 - \gamma - \alpha}. \tag{2}$$

This condition would serve to determine the critical value in a not unreasonable way, though a realistic selection must involve a study of the costs (including the intangible of public opinion) associated with the various errors.

4. Multistage procedures

When the relative importance which is assigned to the expense and annoyance of a false positive is contrasted with the possibly greater damage resulting from a false negative, it becomes a question of how many false positives we are willing to tolerate as against the missing of one true positive. To answer this question one needs, among other things, information about the relative costs of the further tests which each of the false positives will have to undergo. For, as was pointed out above, in all of these operations, it is assumed that there are other means of determining who is positive and who is negative. The procedure, therefore, may be that as soon as a person has been identified by the diagnostic test as positive he is subjected to another test, and those who are retained as positive by the second test may be subjected to a third test. At each stage there is an added expense plus an added risk of further false negatives. We shall refer to a screening program of this kind as multistage.

It might be of interest to present the figures from a recent study here on the Berkeley campus in which 14,867 students were subjected to a 70-mm X-ray examination in a series of 10 radiological readings. Of these, 902 were identified as positives and subjected to the second test which consisted of 3 independent readings on a large 14″ by 17″ X-ray film. The second test identified 256 positives, and among them there were 159 who satisfied a more critical roentgenological examination in that all the three radiologists independently had called the case positive. Subsequently, on the basis of intensive clinical investigations, 12 of the 159 were found to have clinical evidence of the disease.

A whole circle of interesting statistical problems arises out of a multistage screening program, one of which we will discuss briefly. Suppose there are two stages only, the first producing a continuous reading x, large values of which are diagnostic, and a second stage whose verdict we accept as correct. A critical value x_o is selected for the first stage, and all individuals with $x > x_o$ are subjected to the second stage. Suppose n individuals are examined, P of whom receive a positive diagnosis at the first stage (that is, have $x > x_o$). Of these, say that S turn out to be sick (the true positives) while $H = P - S$ prove to be healthy (the false positives). Let us consider the problem of using data from a study of this kind to estimate the parameters α and γ.

The actual situation may be described by a two-by-two table, with each person on the one hand being healthy or sick, and on the other hand diagnosed as positive or negative. The expected proportions in the four cells, and the corresponding ob-

served proportions in the right-hand cells, are as shown below:

	−	+
Healthy	$(1-\pi)(1-\alpha)$	$H/n, (1-\pi)\alpha$
Sick	$\pi\gamma$	$S/n, \pi(1-\gamma)$.

The expectations possess three degrees of freedom for the cell entries—any four nonnegative values summing to one are possible—but we observe the frequencies in only two of the cells, since we do not find out how many of the negatives are sick. Thus, on the face of it we cannot hope to estimate any of the parameters.

Notice, however, that

$$\alpha = \frac{\alpha(1-\pi)}{1-\pi(1-\gamma)-\pi\gamma} = \frac{\alpha(1-\pi)}{1-\pi(1-\gamma)}\left[1 + \frac{\pi\gamma}{1-\pi(1-\gamma)} + \cdots\right] . \tag{3}$$

As γ is customarily small and π very small, $\pi\gamma$ is usually negligible compared to $1-\pi(1-\gamma)$, so that $[\alpha(1-\pi)]/[1-\pi(1-\gamma)]$ is usually a good approximation to α. Since H/n estimates $\alpha(1-\pi)$ and S/n estimates $\pi(1-\gamma)$, we are led to consider

$$\hat{\alpha} = \frac{H}{n-S} \tag{4}$$

as an estimate for α. Writing

$$\hat{\alpha} = \frac{H}{n} + \frac{HS}{n^2} + \frac{HS^2}{n^3} + \cdots \tag{5}$$

and using the multinomial nature of H and S, we have

$$E(\hat{\alpha}) = \alpha(1-\pi\gamma) + \text{terms in } \frac{\pi}{n} \text{ or } \pi^2 . \tag{6}$$

If n is large and π and γ are small, the estimate suggested will have a negligible bias. The variance of the estimate may be approximated by

$$\frac{\alpha(1-\pi-\alpha)}{n(1-\pi)^2} . \tag{7}$$

This quantity (7) reaches its maximum $1/4n$ when $\alpha = (1-\pi)/2$ and is symmetrical with respect to the maximum value. In practice, α will never be anywhere near $(1-\pi)/2$ and thus the variance of the estimate will be considerably smaller than $1/4n$. For $\alpha = 0.05$ and $\pi = 0.01$, the variance of $\hat{\alpha}$ is of the magnitude of $1/20n$.

It seems clear that no such estimate for γ is possible. Even if n were ∞, so that we could determine $\pi(1-\gamma)$ and $\alpha(1-\pi)$ exactly, we could not give a good estimate for γ without knowledge of π. The greater difficulty of estimating γ as compared with α is of course not peculiar to diagnostic aids. For example, college entrance boards find out their mistakes of commission, but not their mistakes of omission.

Several approaches to the problem of estimating γ may be considered. The direct experimental approach consists in admitting to the second stage a sample of indi-

viduals with $x < x_o$ on the first stage. An illustration is provided by the pilot selection program in World War II, in which a large number of applicants were tested and then all were accepted, in order to validate the tests. In our problem this approach is going to be very difficult to apply because of the smallness of $\gamma\pi$: the disease is usually rare, and most of the diseased have $x > x_o$, so a very large sample of those with $x < x_o$ would have to be subjected to the expensive second stage to produce a useful estimate of γ.

In some cases we may be able to get useful information about γ by examining the distribution of x-values among the diseased persons found in the $x > x_o$ group. We have in effect a truncated sample from the diseased population. If we are prepared to assume a form of distribution, we may be able to estimate γ from this truncated sample.

5. Longitudinal diagnosis

The approach of the two preceding sections is based on the key assumption that the population consists of two clearly-defined groups, the healthy and the diseased. This is the simplest approach; it is beyond doubt useful because of its simplicity, and it is historically the natural one to make, as each disease is first studied clinically in its advanced stages. Every diagnostic aid has as its basis the observation of a physiological difference between the general healthy population and a group of persons seriously ill with some disease. Thus, it is observed that hypertensive individuals characteristically have very high blood pressure. It is very natural then to assume that blood pressure may be used as a diagnostic aid for the detection of hypertension.

But this approach is essentially static, and fails to allow for the fact that (especially with a chronic disease) a diseased population forms a continuum of persons in various stages of the disease and progressing in time from one stage to another. The great gap in knowledge of the nature of the disease and particularly of its association with diagnostic procedure is on the large range between the beginning and the end. Here are the early stages, as well as the asymptomatic diseases and those who have been apparently cured. This part of the continuum is being uncovered by the application of diagnostic aids, and it is because of the meager knowledge and understanding of this large segment of the continuum that many of the problems of diagnosis emerge. An example may serve to clarify this point.

Tuberculosis occupies a large continuum from inception to the terminal stages. At very early stages of the disease it is often asymptomatic for a relatively long period, the symptoms are mild and therefore the patient often does not present himself for medical attention until greater symptoms have developed. Consequently, the knowledge that is available to the medical profession about this disease is based primarily on a study of the latter type of tuberculosis. The association with the appearance on the X ray was also inferred from observations in this tail end of the continuum. The development of a relatively inexpensive way of X-raying the population made it possible to obtain data on the middle and early parts of the continuum. At least, it is possible to obtain X rays of people who are apparently healthy and some of whom are in the process of developing the disease and progressing toward the other tail, the symptomatic part. It is no wonder that the evaluation of the results of X-ray findings was transferred from that part of the continuum which

was known to the medical profession to this new phase of the disease, often with very disturbing consequences. This trend is not confined to one disease or one diagnostic aid. One may look at it in general terms and state that the association between a disease process and the results of a certain diagnostic aid has been observed for one phase of the disease process, and that we are faced with the problem of utilizing it in another phase.

In section 3, we discussed the notion of two distributions of a continuous reading x, one in the healthy and the other in the diseased population. There we took a static view of the situation—now let us regard matters dynamically! Consider a healthy individual with a reading x, which may vary from time to time but typically within a relatively narrow range. At a certain point in time, the individual contracts

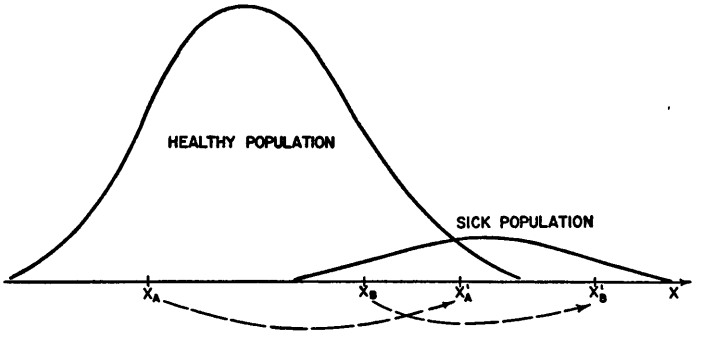

FIGURE 1

the disease, and as a consequence (if the diagnostic aid is in fact diagnostic) his typical value of x will increase. We may envisage two main types of increase: a relatively sudden saltus (as with the rupture of a blood vessel, or the contraction of an acute infection), or a gradual increase (as with most chronic diseases). In either case, we should expect to find his typical x-values to the right of their pre-illness positions.

It may be guessed (though only an observational program could confirm or refute this guess) that an individual A whose healthy value x_A is in the lower range of normal people, will on becoming ill, move to a value x'_A in the lower range of the sick, while an individual B who is high in health will also be high in sickness. Figure 1 illustrates this conjecture schematically. But in any case, there must be a motion to the right if the aid is meaningful.

We conclude that it would be of tremendous diagnostic value, when considering an individual now in the overlapping part, to know *where he used to be.* If for many years he has had x-values in the same range, we could conclude that he is a person for whom such high values are normal, and thus classify him in the healthy population. On the other hand, if we knew that his present value represents a recent and considerable increase from former values, it would be reasonable to conclude that he has become ill. In other words, a diagnostic aid is likely to be very much more valuable if it is used dynamically than if it is used statically. It is clear that dynamic use of the aid requires that it be used (and the results recorded and studied) over an extended period of time. We shall refer to this kind of use of an aid as "longitudinal."

It is helpful to envisage two distributions of a diagnostic aid measurement x. The

usual mass survey produces a distribution of x-values that may be termed "cross-sectional," which represents the variation from individual to individual at a fixed time. Contrasted with this, we may think of the "longitudinal" distribution of values which could be obtained by periodic examination of a single individual—representing variation from time to time for a fixed individual. In many cases, we believe, the spread of the longitudinal distribution of an individual whose health status is not changing will be very much less than the spread of the cross-sectional distribution. It is this fact which leads us to believe that the possibilities of early diagnosis may be radically improved if the diagnosis is based on comparison of recent measurements with earlier measurements *on the same person.*

A good example of the point is provided by the use of blood cholinesterase levels as a diagnostic aid in the detection of poisoning from organic phosphorus insecticides (see for example, [3]). It is known that the poisons tend to lower the readings. Normal levels vary considerably from person to person, displaying a coefficient of variation (for the plasma enzyme) of about 22%. The temporal coefficient of variation for a given healthy individual, however, is typically about 8%. It is obvious from these figures that one can detect the lowering of the level much more readily if the normal value for each exposed individual is on record.

In the future a most valuable possession of each of us would be a list of results of examinations obtained in different periods in our life for a number of specific tests. These results could be entered on a single card, perhaps in a graphic manner to suggest and facilitate comparisons. At any subsequent examination the importance of such a card would be incalculable for early detection of disease. It is apparent that for the moment this type of information is not generally available. As a forerunner, one might add parenthetically that a beginning in this direction can be made even now. For example, many industries are conducting periodic examinations of all their employees, so that annual or biennial measurements on a large number of employees are available. Unfortunately in many instances the main advantage of this periodic examination is lost, for each examination is considered on its own with very little reference to the results of previous examinations. But if methods of keeping the records were geared to a longitudinal basis, these examinations could become infinitely more valuable. Similarly, a number of us go periodically to our physicians for checkups. It is, however, an exceptional physician who keeps and consults the records on a healthy person (even if he does consult these records for the patients who have certain abnormalities). In short, if the continuing and changing values are of importance in diagnosis and if methods of analysis were available, it is not unlikely that a longitudinal record could become a more common practice in the future modes of diagnosis.

It can, of course, be questioned whether it is reasonable to treat a series of periodic readings on the same individual as a random sample. Indeed, the measurement of a diagnostic aid is usually a function of the age of the individual. Periodic readings need not be independent or identical variables even when the individual is healthy. The putative correlations between temporally adjacent readings would suggest the use of a stochastic process model. It may turn out that the series of periodic readings on a healthy person form a definite pattern which can be described by a stochastic model. The same type of model may also apply to other healthy individuals with values of the constants involved in the model adjusted for each individual. A

deviation from the pattern would then be an indication of illness. This point is to be taken up in the following section.

6. Analysis of longitudinal records

We shall conclude by formulating some of the statistical problems which the collection of longitudinal records will present. The problems arise in two phases:

(1) A large number of longitudinal records, gathered under fixed conditions, must be studied to reveal "normal" time patterns or paths of change of the diagnostic measurement, as well as "normal" degrees of variation about a normal pattern.
(2) We need to formulate a diagnostic rule for deciding when an individual has departed from his normal path.

Typically, the result of phase (1) could be the production of a grid (that is, a new coordinate system) in terms of which the normal path is a horizontal straight line, with variance about this line homoscedastic. The pediatric grid of Wetzel may serve as an example of phase (1) to a certain extent.

It was formerly customary for pediatricians, after measuring the height and weight of a child, to compare these figures with a standard height-weight chart. A child found to be a certain percentage away from the expected value for his age would be considered malnourished. Pediatricians realized the unsatisfactory nature of this procedure, as they found many perfectly healthy children who could not meet the criterion, and conversely. Nor was the unaided judgment of the pediatrician reliable, as shown by comparing several judgments on the same child.

An important conceptional advance was made when Wetzel [4] suggested that each child possess a pattern of growth which is normal *for him*. The problem then is to discover for each child his own normal growth pattern, departures from which indicate malnutrition. Wetzel produced a grid, now in wide clinical use. Essentially, Wetzel's grid consists in the statements that in normal growth the point whose coordinates are the logarithms of height and weight moves in a straight line, and that the lines for all healthy children are parallel. We were not able to find in Wetzel's writings any indication that his paths were derived from the consideration of large numbers of actual longitudinal records, or a convincing demonstration that there is in fact a normal path, departure from which is associated with malnutrition as judged by another criterion.

Generally, it is clear that phase (1) cannot be satisfactorily attacked from cross-sectional data, that a large number of longitudinal records both in health and with the onset of disease will be required. One way of attacking the problem is to study the displacement of the periodic readings of a diagnostic aid. Let x_k be the kth reading in a series. The idea is to study the distribution of the difference $x_k - x_{k-1}$, for all possible values of k. Methods of analysis will depend on whether the displacements can be treated as independent random variables. The problem will be relatively simpler when they are independent. In case they are not independent, one may consider the application of the established theory of stochastic processes, such as Uhlenbeck-Ornstein process [5]. If study is to be made on the readings themselves, presumably an attack on such data will involve search for transformations of variables under which healthy patterns become stable in time, with constant

variance about the constant mean; while the onset of disease is represented by a change of mean—often in a known direction.

Fortunately, many problems are already in this simple form to start with. For example, in the insecticide poisoning example introduced above, the cholinesterase level tends to be temporally stable in health, becoming depressed with poisoning. Two modes for this change may occur:

(1) A sudden saltus to a new level (for example, ingestion of a single, large dose of the poison).

(2) A gradual drift to a new level (for example, the slow accumulation of poison).

The statistical methods useful in detecting the two types of change may well differ. The first is simpler and we shall conclude by examining it. For still greater simplicity, let us assume

(i) Observations are made at regular intervals.

(ii) Observations are normal, independent, and have known variance (which we take to be 1).

(iii) In health, observations have a known mean (which we take to be 0).

(iv) The onset of illness is marked by the saltus of the mean to a new value $\mu > 0$, where it remains.

Of these assumptions, (iii) is least realistic, as the whole idea of our longitudinal approach requires the existence of person-to-person variation. In practice, we should estimate the individual's mean from a number of initial observations, which we would have to assume were made in health.

Superficially, our problem is very like that of industrial quality control charts "with known standard." But the resemblance is only superficial. In industrial work, the measurements are frequent, forcing the analysis to be very simple, and the consequences of a slight delay in detecting loss of control are not too serious. In the medical problem, there is usually an extended period between observations, and early detection is the essence of the whole program. Our problem is thus to devise a statistically efficient use of the control chart.

A decision rule for the control chart is essentially a sequential stopping rule—positive diagnosis being equivalent to stopping the process. Under any reasonable rule, we are certain to stop eventually, even if health is maintained. Therefore, we cannot formulate the problem in terms of the usual probabilities of rejection under hypothesis and alternative. A reasonable criterion seems to be the expected value of the number N of observations required to stop. We may state the problem explicitly as follows: *Find that sequential stopping rule, having a fixed (large) value of $E(N)$ for an individual who remains healthy, which minimizes the expected value of the number N of observations required to stop the process after the onset of disease.*

Let us first examine the classical control chart from this point of view. We fix an "upper control limit," say C, and stop the process as soon as any measurement exceeds C. If the kth measurement is denoted by X_k, we see that for a healthy individual, the conditional probability of stopping on the kth observation, given that we have not stopped before, is

$$\pi_k = Pr\{X_k > C\} = 1 - \Phi(C) \tag{8}$$

where Φ denotes the normal distribution function. Thus N has a geometric distri-

bution, with $E(N) = 1/\{1 - \Phi(C)\}$. Now suppose an individual becomes ill ($E(X) = \mu$); N is also geometric, with

$$\pi_k' = Pr\{X_k > C\} = 1 - \Phi(C - \mu) \tag{9}$$

and $E(N) = 1/\{1 - \Phi(C - \mu)\}$.

It seems intuitively likely that the classical control chart rule can be improved, since it restricts itself to looking each time at just the most recent observation. There is always the chance that we do not detect loss of control at once (a good chance, unless μ is very large). Correspondingly, if we are contemplating the possibility after observing X_k that the process is out of control, we must consider the possibility that X_{k-1} was also out of control. Therefore, a rule which examines each time only the last observation may be quite inefficient.

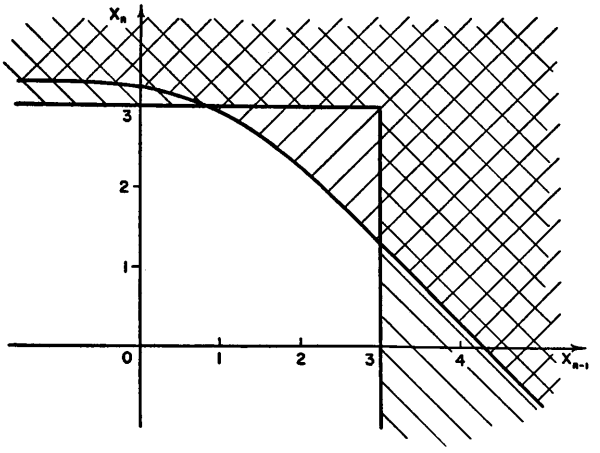

FIGURE 2

A suggestion of a possibly better rule arises easily from a Bayes model (in this connection, see Girshick and Rubin [6]). Suppose there is in each interobservational period a small probability ϵ that the process goes out of control. To simplify, suppose the process cannot have been out of control for more than K observations. The *a priori* probability that the process is still in control, after observing X_n, is

$$\frac{1}{1 + \frac{\epsilon}{1 - \epsilon}\left[L(x_n) + \frac{L(x_n)L(x_{n-1})}{1 - \epsilon} + \cdots\right]}, \tag{10}$$

where $L(x)$ is the likelihood ratio for a single observation. Since ϵ is small, this suggests the following rule.

Stop as soon as

$$L(x_n) + L(x_n)L(x_{n-1}) + \cdots + L(x_n) \cdots L(x_{n-K+1}) \tag{11}$$

exceeds C'. The constant C' could be adjusted so as to give to $E(N)$ the desired value.

It is interesting to compare this rule with the control chart rule for $K = 2$. The figure shows the stopping regions for the rules

(1) Stop when any $X_n > 3$

(2) Stop when $L(X_n) + L(X_n)L(X_{n-1}) > 100$, $\mu = 2$.

The comparison of the performance characteristics of such rules involves rather heavy computations. A numerical investigation is in progress at the present time.

REFERENCES

[1] M. M. BLACK and F. D. SPEER, "Biologic variability of breast carcinoma in relation to diagnosis and therapy," *New York State Jour. of Medicine*, Vol. 53 (1953), pp. 1560–1563.

[2] JOHN E. DUNN and SAMUEL W. GREENHOUSE, *Cancer Diagnostic Tests*, Public Health Service Publication No. 9, U. S. Government Printing Office, Washington, D. C., 1950.

[3] J. M. BARNES and D. R. DAVIES, "Blood cholinesterase levels in workers exposed to organophosphorus insecticides," *British Medical Jour.*, October 6, 1951, pp. 816–819.

[4] N. C. WETZEL, "Growth III. Measurement analysis and evaluation of growth, development and metabolism in children by the grid technique," *Medical Physics*, Vol. 1 (1944), pp. 535–567.

[5] G. E. UHLENBECK and L. S. ORNSTEIN, "On the theory of the Brownian motion," *Physical Review*, Series 2, Vol. 36 (1930), pp. 823–841.

[6] M. A. GIRSHICK and HERMAN RUBIN, "A Bayes approach to a quality control model," *Annals of Math. Stat.*, Vol. 23 (1952), pp. 114–125.

[7] J. NEYMAN, "Outline of statistical treatment of the problem of diagnosis," *Public Health Reports*, Vol. 62 (1947), pp. 1419–1456.

[8] JOSEPH BERKSON, "Cost-utility as a measure of the efficiency of a test," *Jour. Amer. Stat. Assoc.*, Vol. 42 (1947), pp. 246–255.