Applications of Computer Produced Frequency Distribution Curves

I. Quality Control

N. Ressler and L. S. Whitlock

Applications to quality control of a laboratory's own individual frequency distribution curves are described. The method has recently become feasible for routine use, since the curves can now be compiled automatically by means of an electronic data-processing system. Applications described include verification of the validity of results generated by a given method in terms of the frequency distributions obtained for each type of diagnostic classification, and the determination of the constancy of these distributions. The relative value of different test methods can be compared by the success which is achieved in distinguishing diagnostic classifications (as measured by the overlaps of different frequency distribution curves). Individual frequency distributions can also be used to determine the precision required for given test, in order to avoid any loss in the discrimination which can be achieved.

As the variety and number of tests done in clinical laboratories increases, the role of quality control becomes more important. With increasing work loads, an inaccurate procedure can lead to a larger number of answers which are incorrect or of little value. Effective quality control is difficult to achieve, however, for the following reasons (1). Limitations of time or personnel may restrict the selection of methods to those which are simpler to perform, or which lend themselves to automation. Accurate standards are not always available (2), and some test results may be affected by other substances in serum in addi-
tion to the component being determined (1). The use of individual frequency distribution curves, along with standards and controls, for the control of accuracy and precision is discussed in this paper. The term "individual" frequency distribution curves refers to a separate frequency distribution for normal and for each abnormal diagnostic classification. Applications to quality control of a single frequency distribution curve, composed of all samples received, both normal and abnormal, have been discussed by Waid and Hoffmann (3). Further applications of frequency distributions to quality control are made possible by the correlations between test values and the particular diagnostic classifications represented by individual curves for each diagnostic classification.

It has not previously been feasible for a routine clinical laboratory to compile individual frequency distributions, since the manual work involved would require an unreasonable amount of time. Methods have recently been described for the direct incorporation of patient case history, chart information (4, 5), and test results (6) into data-processing systems. The frequency distributions can thus be compiled automatically, thereby making the applications described feasible for routine use in each laboratory.

Individual frequency distribution curves for different diagnostic classifications have previously been utilized very little on a routine basis in clinical chemistry. One example of their use is the application of a method which has been developed by Carlstrom et al. (7) for the evaluation of hepatic tests in the differential diagnosis of jaundice.

Some of the calculations and compilations of frequency distribution curves representing the laboratory data used to illustrate various steps of the applications described here were done manually, while others were done with the aid of an IBM 7090 computer.

**Frequency Distribution Curves**

**Interpretation of Overlap**

If the concentrations of a particular constituent in the serum of normal individuals and of individuals with a specific disease are plotted, any degree of overlap between the two frequency distribution curves is possible. This is illustrated by the hypothetical curves in Fig. 1. In Fig. 1A, each curve is discrete, with no overlap between the normal and abnormal values. Any concentration less than $x^\prime$ is normal, any concentration greater than $x^\prime$ is abnormal, and each population can be distinguished by the test. In Fig. 1C there is complete overlap, and the test has no diagnostic value for this disease. In Fig. 1B there is a partial
overlap. Any value less than $x'$ is normal, and any value greater than $x''$ is abnormal. Between $x'$ and $x''$ the interpretation is uncertain.

The probability of a particular value in the overlap region being abnormal can easily be plotted, as has been discussed by Hoffmann (8) and as is shown in Fig. 2. At any given concentration, the probability is determined by the relative heights of the two curves in the overlap area of Fig. 1B.

Fig. 1. Normal (solid line) and abnormal (dashed line) frequency distribution curves. Ordinate is number of samples, and abscissa is concentration. A. Discrete curves—no overlap. Populations can be distinguished by test. B. Partial overlap. Distinguishment uncertain. C. Complete overlap. Test has no diagnostic value.

Fig. 2. Probability of test results (in overlap zone of Fig. 1B) corresponding to an abnormal individual.

The determination in this manner of the probability of an unknown patient, with a given test value being normal or abnormal, is nonpara-
metric and independent of the shape of the frequency distribution (whether normal, skewed, irregular, etc.). When both high and low abnormal distributions occur on either side of the normal one (as with glucose concentrations, for example), the probability can be determined over each overlap region in the manner described.

Obviously, the proportion of answers which cannot be definitely distinguished as normal or abnormal increases with an increasing degree of overlap. The sum of the areas of the two curves in the overlap region (between \( x' \) and \( x'' \) in Fig. 1B) divided by the sum of the total areas under both curves is a measure of the proportion of values which cannot be distinguished by the method.

Although tests may result in complete discrimination between normal and abnormal populations in some instances, i.e., as in Fig. 1A (9, 10), normal and abnormal curves overlap in the vast majority of the cases (as in Fig. 1B). Composite frequency distributions for a number of the more commonly determined constituents, such as cholesterol, glucose, urea nitrogen, etc., do not exhibit separated curves for normal and abnormal values (8, 11, 12). Despite the fact that an undiagnosed patient cannot be classified with certainty by a test value which falls within the overlap region of the curves, such a value is nevertheless frequently labeled normal or abnormal, depending on whether or not it exceeds a fixed limit. Such labels must be incorrect in a certain percentage of the cases, depending on the overlap. It is difficult to avoid this incorrect labeling unless both normal and abnormal frequency distributions are available, since both are necessary in order to determine the overlap and the discrimination actually achieved (13).

Data for frequency distribution curves can be classified in a number of useful ways. In the diagnosis of a given disease, for example, one curve may consist of all patients with that disease, and a second of all other individuals without that disease, either normal or abnormal. Other comparisons can be made between normal individuals and those with a particular pathologic process, such as myocardial infarctions or hepatitis.

**Use for Quality Control**

The shape of a distribution curve is related to both the population being tested and to laboratory accuracy. The compilation of individual distribution curves permits a comparison of the results obtained in a given laboratory to those published or obtained in other laboratories. It is important to compare abnormal as well as normal ranges, since different methods for determining the same constituent may agree in the normal range but not the abnormal range. Colorimetric and ultra-
violet methods for the determination of serum transaminase generally give the same answers with normal specimens, but not when the values are elevated (14). Comparisons of individual frequency distributions can be made in terms of range, means, modes, and the proportion of patients with a given disease which can be distinguished by a particular test (i.e., the proportion of the areas of the curves outside the overlap region). Standard methods for the comparison of any two frequency distributions, such as the t test, in which the standard error of the difference between two means is a meaningful statistic, can be utilized. When any of the ranges and their mean values or the degrees of discrimination differ from those obtained in other laboratories, the accuracy of the method can be checked. This can be done by the use of an alternate method (15), additional standards, exchange of samples, etc.

A comparison of some normal and abnormal ranges is presented in Fig. 3. Figure 3A shows the ranges for serum potassium levels considered normal at 9 of the larger hospitals in the Detroit, Mich., area,
and the range at our medical center (range c). Five of the hospitals had
the same normal range (b), but the ranges of all of the others differed.
The laboratories of some of these hospitals are comparing results ob-
tained on the same samples, which may help to determine if differences
in those ranges are due to methodology, or to other causes.

Figure 3B presents frequency distributions for the levels of serum
lactic dehydrogenase among patients with megaloblastic anemia
(16-18). The results show that these patients were distinguished from
normal individuals by means of the lactic dehydrogenase levels in two
of the laboratories (II and III), but that there was a significant overlap
between the levels of the patients and normal individuals in the other
laboratory (I). A comparison of the results obtained by these labora-
tories on the same samples or on standards of known concentrations
could help to determine if this difference is due to the methodology, or
to such factors as the etiology or severity of the megaloblastic anemia,
criteria used for diagnosis, etc.

The compilation of a laboratory's own frequency distributions is
important in relating test answers to normal or abnormal populations.
If an answer which corresponds to a normal result is considered ab-
normal it is of little value, even though the accuracy may be perfect.
Since the range of potassium values in which a patient would be con-
sidered normal in one laboratory and abnormal in another (I and III at
the bottom of Fig. 3A) is over twice as great as the range in which
there is agreement (II), it is apparent that the verification of normal
or abnormal ranges can be important in the use of the potassium results
and in the diagnosis of patients. The danger of an uncritical acceptance
of literature values for normal and abnormal ranges has been discussed
by Zieve (19).

In addition to outside comparisons, the ranges, means, modes, and
discriminations can be important in regard to their constancy over a
period of time. Changes in these values in a given laboratory may indi-
cate that a testing procedure is not producing consistent results. The
use of this method for quality control, with a single frequency distribu-
tion which includes all diagnostic classifications, has been proposed by
Hoffmann and Waid (20, 21). The value of their method has been ques-
tioned, however (22). Changes in frequency distributions may be due
to variation in the types of hospital patients, rather than to the test
methodology. This may be especially relevant in regard to diseases
related to seasons, or in the case of epidemics. This limitation, as well
as any assumptions concerning the shapes of the frequency distribution
curves, is avoided by the nonparametric use of individual frequency
distributions, since each diagnostic classification is considered independently.

Data for the excretion of hippuric acid by normal individuals and by patients with liver disease can be used to illustrate how the validity of control methods based upon composite curves, but not those based upon individual curves, depends on assumptions concerning the proportion of abnormal cases and on the shapes of the frequency distributions. The data (23) are presented in the form of individual frequency distributions for normals and for abnormals in Fig. 4A. A single composite curve representing the data obtained on both groups is shown in Fig. 4B. It is apparent from a comparison of Fig. 4A with Fig. 4B that the normal range cannot be determined from the composite curve alone, and that an attempt to do so by fitting a normal curve about the mode (the dotted line in Fig. 4B) may not lead to correct results.

If the hippuric acid determinations are well controlled, and the number of patients with liver disease doubles (while the number of normals remains the same), the individual and composite frequency distributions which would be expected to result are shown in Fig. 4C and 4D, respectively. The abnormal curve in Fig. 4C is the same as in Fig. 4A except that the ordinate values are doubled. The mode of the composite curve in Fig. 4D, however, has shifted to lower values and the proportion of the total area over the lower values has increased. Consequently, the "number-plus" (20) and "average-of-normals" (21) methods of quality control, based upon composite curves, would both tend to indi-

![Fig. 4. Urinary excretion of hippuric acid (gm. excreted 1 hr. after I.V. injection of 1.77 gm. sodium benzoate), with variations in method of representation, proportion of abnormal values, and accuracy. In upper curves separate frequency distributions used for normals (solid lines), and for patients with liver disease (dashed lines). Lower curves, all values combined in single composite distribution. A and B, direct representation of laboratory data; others were calculated. C and D, changes which result when proportion of abnormal cases doubles. E and F, curves are altered when doubling is accompanied by increase of 0.11 gm. in all hippuric acid values. Dotted line in B is apparent normal distribution, and is also shown in D and F in same position for purposes of comparison.](image-url)
Fig. 5. Representation of variations of Fig. 4 by calculated quality control charts. Charts I and II based on apparently normal range and mode of composite Curve B in Fig. 4. In control Charts III and IV, based on separate frequency distributions for normals and abnormal, points represent modes, and vertical arrowed lines represent ranges. Time period A,B (abscissa) represents experimentally found distributions; C,D, a doubling of the abnormal distribution; and E,F, represents time during which doubling of abnormal population is also accompanied by increase of 0.11 gm. in all hippuric acid values. See text for methods of calculation.
cate a lapse in quality control, although none has occurred. The number-plus method of control is related to the constancy of the proportion of total area on one side of the mode (of the estimated normal range) in a composite curve of Fig. 4B. The average-of-normals method is based upon the constancy of the mean of all the values within the estimated normal range of the composite curve. A composite frequency curve such as in Fig. 4B, compiled with a large number of samples, is used as a basis for the comparison of subsequent changes. The above shifts, caused by the increase in abnormals, would indicate a change in the values from the original distribution.

If the abnormal population remains doubled, but all values obtained are 0.11 gm. too high, the frequency distribution curves would be shifted along the abscissa as shown in Fig. 4E and F. The composite curve in Fig. 4F now has the same normal mode as the curve in Fig. 4B. The percentage of the total area to the left of this mode is also similar to that of the curve in Fig. 4B (58.5 vs. 55.0, respectively). Thus, neither the number-plus nor the average-of-normals method would detect the 0.11-gm. error which has occurred.

Each of the above situations is represented in terms of quality control charts in Fig. 5. Since the points were calculated from the above curves, there are no deviations from expected distributions. The 95% confidence limits were calculated in the manner described by Hoffmann and Waid (20, 21) and were made on the condition that each point represents 20 samples. When the number of abnormal cases doubles, the number of plus tests, i.e., the percentage of the total area to the left of the mode value of 0.87 gm. times the number of samples (0.705 × 20), and the normal mode both shift to values outside the 95% confidence limits. Thus, a loss in control is falsely indicated. The control charts based upon the individual normal and abnormal distributions do not indicate any error. Modes have been used in the above example instead of the mean values employed in the average-of-normals method, since they are easier to follow in the different frequency distributions. The increase in abnormals causes a similar decrease in the mean value. It decreases less than the mode, however (to 0.80 gm.). The error indicated in the above charts by the modes is thus greater than would be indicated by mean values.

When an increase in the abnormal curve is also accompanied by a 0.11-gm. error in the determinations, the above methods, based on the composite curve, fail to detect the error. The error is apparent, however, from the ranges or modes of the individual normal and abnormal control charts.
These examples demonstrate how quality control charts, based upon individual frequency distributions for each diagnostic classification, can be of value for the maintenance of quality control. Use of these charts does not, however, obviate the need for the use of standards and for the additional quality control charts, which indicate standard deviations obtained with the controls that are processed along with the unknown samples.

In addition to compilation of the individual frequency distributions and quality control charts, a data-processing system can automatically provide the laboratory with warnings when any of the ranges, modes, means, degrees of discrimination, and standard deviations obtained with controls do not compare to outside results, or when they change beyond acceptable limits.

Use for Diagnostic Discrimination

The ability of a test method to distinguish between the different diagnostic classifications for which it is used, as measured by the degree of overlap of the individual frequency distributions, indicates the value of that method. The discrimination achieved by different tests can be compared in this manner. The selection of the most discriminating tests, and elimination of the less effective ones, can thereby be facilitated.

Individual frequency distribution curves can also be of value in the determination of acceptable limits of standard deviations. Standard deviations are generally determined in clinical laboratories, but it has not always been clear how much precision is required, or what standard-deviation limits should be considered allowable. The effect of a given deviation upon the discrimination achieved by a particular test can be considered in terms of frequency distribution curves. Figure 6 illustrates how lack of precision decreases the diagnostic specificity of a test when the normal and abnormal frequency distribution curves overlap. Figure 6 is similar to the curve shown in Fig. 2, but an allowance is included for the effect of lack of precision.

If a concentration \( b \) is obtained from an unknown sample (Fig. 6), the true concentration will be within the range of two standard deviations, i.e., \( a - c \), in 95% of the cases. The probability of concentration \( b \) being from a clinically abnormal individual would be 50% with absolute precision. If, in the above illustration, the true value were \( a \), the probability would be 25%, while if it were \( c \), the probability would be 75%. Since the standard deviation makes the exact probability uncertain, it reduces the discrimination which the test can achieve. Consequently,
whenever any overlap is present, the smallest possible standard deviation is needed.

Even if the "true" distribution curves do not overlap, but the deviations are greater than the concentration differences between the adjacent "ends" of the normal and abnormal curves, an artificial area of overlap may be generated. If the deviations are less than the concentration difference between the adjacent ends of the normal and abnormal curves, then the lack of precision does not decrease the discriminatory power of the determination.

These considerations can be illustrated by serum creatine phosphokinase determinations, performed at our medical center on normal individuals and on patients with acute myocardial infarctions as seen in Fig. 7 (24). The standard deviation calculated after determining a sample 10 times, with a mean value of 15.0 spectrophotometric units, was 0.8. Figure 7 shows individual frequency distribution curves which represent the experimental data. It also shows calculated frequency distribution curves which would have resulted if the standard deviation had been 50% greater. The increase in standard deviation is just sufficient to result in a significant overlap in the normal and abnormal curves and in a decrease in the ability to distinguish between the two classifications on the basis of this test. If the standard deviation increased further, the overlap and the loss of discrimination would become greater. If, on the other hand, the calculated curve represented

---

**Fig. 6.** Dependence of range of possible probability values on magnitude of 2 standard deviations (arrowed line). Probability of answers corresponding to an abnormal individual.
the original data, then the discriminatory ability of the determinations would be improved by an increase in precision. These considerations indicate that in this example a standard deviation of less than 1.2 is needed in order to avoid a loss in discrimination.

![Graph showing the effect of precision on discrimination.](image)

**Fig. 7.** Effect of precision on discrimination. Separate frequency distributions for normals and patients who had had myocardial infarctions 6–36 hr. previous to time blood was taken. Upper curves compiled from laboratory data and lower curves calculated: shows alteration produced by increase of 50% in standard deviation about means. Relative areas under curves kept constant.

Tonks (25) has proposed the following empirical formula for acceptable deviation limits:

$$\left[ \frac{\frac{1}{4} \text{ (normal range)}}{\text{mean of normal range}}} \right] \times 100$$

It is based on the premise that deviations should not exceed one quarter of the normal range. Sparapani and Berry have also suggested an
empirically based limit for allowable deviations, which is a coefficient of variation of up to 8% (26).

The availability of both normal and abnormal frequency distributions is necessary in order to determine the precision that is required in any given case. The normal range for two different determinations may be the same, but if the abnormal distribution curves and the overlaps differ, the accuracy requirements are not the same (Fig. 1). The necessity of using empirical or arbitrary limits can be obviated by a consideration of the reproducibility needed in terms of the overlap of individual frequency distributions (Fig. 7).

Even though the smallest possible standard deviation may be needed, the precision which can be achieved with any given method is limited by the manipulations, and is generally indicated in relevant publications or textbooks (27). Comparison with these values may therefore provide the most useful guide for the determination of the precision which can reasonably be expected, or considered acceptable. Procedures in which the standard deviations result in the greatest losses in discrimination (or greatest percentage of uncertain answers) are those in which more precise methods are most needed. The precision can be improved by running samples in duplicate or triplicate when indicated.

An additional use of individual frequency distribution curves which can be mentioned is the detection of inconsistent results obtained with a given patient. If, in Fig. 4A, for example, a patient has a hippuric acid excretion value of 0.55 gm., this would indicate the presence of liver disease. If, on the basis of other tests done on this patient, there is little probability that he has liver disease, the data-processing system could signal the need for a repetition of the hippuric acid excretion test. It should be noted that it may not be possible to determine accurately if a given value is normal from a composite curve, and a value of 0.55 gm. would be within the estimated normal range in Fig. 4B.

The use of probabilities to represent correlations actually obtained between each test and different diagnostic classifications makes it possible to consider the significance of combinations of tests by the combination of individual probabilities into a single resultant value. This procedure is discussed in a subsequent paper.

References
