J. clin. Path. (1952), 5, 305.

THE RUNNING OF STANDARDS IN CLINICAL CHEMISTRY AND THE USE OF THE CONTROL CHART

BY

RICHARD J. HENRY AND MILTON SEGALOVE
From the Bio-Science Laboratories, Beverly Hills, California

(RECEIVED FOR PUBLICATION JANUARY 4, 1952)

The progress of clinical medicine has been considerably aided in the past few decades by the increased number of laboratory tests available and the improvement in their specificity and accuracy. Nevertheless, the reliability of analytical results from clinical laboratories has been questioned by many individuals, and several independent surveys have more than justified the suspicion (Belk and Sunderman, 1947 ; Shuey and Cebel, 1949).

In these surveys blame has been placed on poor supervision of personnel, poorly trained and insufficient personnel, poor equipment, poor choice of methods available, and so on. The vast majority of clinical chemical analyses are and probably will continue to be performed by, and supervised by, people not trained as chemists. In such cases laboratory staff follow methods in the so-called "cook-book" fashion with only sparse knowledge of the chemistry involved. Faced with this problem, it becomes the job of the clinical chemist to outline procedures in detail so as to make them as fool-proof as possible.

It is the purpose of this paper to discuss the importance of running routine standards with clinical chemical analyses, the types of standards that can be employed, and the treatment of the data accumulated so as to give reasonable assurance that the results being obtained are within a known permissible error. By such means the laboratory can strengthen the reliability of its results and the faith of the physician in them.

The Use of Standards

Photometric analyses make up the bulk of quantitative clinical chemical analyses. The chemist considers the running of frequent standards with photometric or spectrophotometric analyses routine, but in the clinical laboratory this has seldom been true since the advent of the photocolorimeter. "Precalibrated" photometers are widely advertised to-day with the implication that the tests need not be standardized by the buyer. The acceptance of such precalibrations cannot be too severely condemned. By relying on precalibrations errors can occur as a result of making or buying new reagents, the deterioration or contamination of old reagents, the use of incorrect light filters, changes in the characteristics of the photometer itself, etc. Standardizing each test when a new instrument is obtained is not enough, since the same errors can and do develop subsequent to the standardization. Assuming the necessity of running standards, there is the question of how frequently they should be run. Ideally they should be run with every set of determinations, or once a day (Archibald, 1950). Some have advocated running them twice a week (Levey and Jennings, 1950). Certainly this is far better than not running them at all, but a definite risk is being taken in the latter case, because several days may elapse before serious error is detected. The authors have seen a refrigerated biuret reagent deteriorate within 48 hours to the extent that standards read 30% below what they should have read. Whether this risk is justified is a question which is answered by the individual situation and from experience gained with a particular test over a period of time.

Types of Standards

The Pure Standard.—This is the usually recommended standard and consists of a solution of the substance of known purity with perhaps a preservative added-(e.g., benzoic acid in the case of glucose standards). Any other type of standard must first be standardized against a pure standard. The advantages of using this type of standard are that the purity is usually known with a high degree of accuracy, and, with a few exceptions, such as enzymes standards, such standards are readily available. The disadvantages include (1) the danger of prejudicial handling, since the person running the test knows beforehand what the results should be; (2) substances normally present in blood which may interfere with or augment the colour are not present; and (3) pure standards usually are not run through the entire procedure. With some exceptions, however, there is no reason why pure standards cannot be run through the whole procedure, and, where possible, it is usually advisable. The danger of omitting certain steps is rather obvious. For example, in the determination of blood glucose the usual way to run a standard is to treat the standard glucose solution as if it were a blood protein-free filtrate. The reagents used in preparing the protein-free filtrates of the unknowns are thus by-passed and are not controlled. Any contamination of these reagents leading to alteration of results in the unknowns would not be impressed on the results of the standard.

Pooled Blood or Serum.—This type of standard has been suggested by Archibald (1950) and by Levey and Jennings (1950). The obvious advantage is that it can be introduced as a routine sample, thus eliminating the possibility of prejudicial handling. Furthermore, it must be run through the entire procedure. The primary disadvantage is that it is not at present easily available to all laboratories. Although it is quite probable that most of the chemical values are stable in blood in the frozen state, relatively little actual data are available. Levey and Jennings found that specimens stored in the deep freeze were stable for urea, total protein, albumin, chloride, and CO_2, but not for globulin.

Internal Standard.—This procedure involves the addition of a known " pure standard " to an aliquot of one of the unknown samples. Such a standard would, of course, have to be run through the entire procedure and is readily available. The internal standard is usually utilized in analytical chemistry when there is the possibility of substances present in the unknown which will interfere with the development of colour in the analysis. To check for such interference in every sample would require running an internal standard on every unknown.

Blanks

The use of blanks on the materials used in the tests is nearly as important as the standard. When pooled blood or serum is used as the standard, contamination of a reagent would result in an elevated value for the standard. A reagent blank, therefore, would not be necessary to indicate that trouble of some nature was present; it would, however, indicate the source of the trouble, and if the contamination were not gross the run might be salvaged. But, with the other types of standards, " pure standards " and " internal standards," reagent blanks may be essential. For example, in the analysis of phosphate by the method of Fiske and Subbarow (1925), the standard usually employed is monopotassium phosphate in trichloracetic acid. The trichloracetic acid used for making the protein-free filtrate of the serum sample, however, is not the same trichloracetic acid present in the standard solution. Contamination of this reagent or the water used in the unknown in conjunction with trichloracetic acid would be missed unless a blank were run. In fact, single-distilled water on occasion can contain significant amounts of phosphate and ammonia.

It is suggested, further, that where the absorption of a reagent blank is low, reagent blanks, standards, and unknowns all should be read versus distilled water set at 100% transmission. In this way they are always read against a blank of fixed optical density.

Definitions of Accuracy, Precision, and Reliability

Before discussing the problems of replication of standards and unknowns and the ways of handling data, it is best to define the terms accuracy, precision, and reliability as commonly used to-day in the statistical approach.

Accuracy of a result consists of comparing the observed result with the actual true value. If a known pure standard is used, the accuracy of a procedure frequently can be estimated by recovery experiments with internal standards. In many cases no true value can be determined and a strict determination of accuracy is not possible.

Precision is the degree of variation in results obtained by a method when the same sample is run repeatedly: in other words, the reproducibility of what is observed. The less variation observed, the greater the precision. Obviously, precision can be determined with a high degree of accuracy for every test procedure.

The reliability of a test is the ability to maintain its accuracy and precision into the future. If a test has maintained a steady state of accuracy and precision over a long period of time, then one can predict these characteristics for the future with assurance and the test is said to be reliable. Reliability of the test (which includes the test, analysts, and equipment) can be established only by checks over a period of time.

Control of Accuracy and Precision

The accuracy of a test is a characteristic occasionally beyond the control of the analyst. Various methods for determining the same substance may give different results due to differences in specificity. Blood glucose determinations are a good example of this, since with some methods, for example that of Folin–Wu, reducing substances other than glucose normally present in blood are determined as glucose. Such a test has poor accuracy but may have good precision. This does not necessarily negate the value of the test, especially if the "interfering" substances do not vary markedly, since the clinician becomes accustomed to the definite set of normal values given by the test.

The control of precision requires more serious consideration. For example, if a report on a blood glucose concentration is 140 mg.%, it is important that the clinicians have reasonable assuredness that the true value by the method used is not 110 or 170 mg.%. "Reasonable assuredness" is commonly defined in statistical analyses as "95% confidence," i.e. the value can be reported with a range limit on each side, within which range the true value, by the method used, will be found 95 times out of 100. More rigorous "confidence limits" can, of course, be used. If it is found for the glucose determination used in our example that the confidence limits are ±10% of the observed value, then the clinician can be given the "reasonable assurance" in the form 140 mg.% ± 10%, or 140 mg.% ± 14 mg.%. Ideally each report should be accompanied by some indication of its precision, although routine reporting of clinical chemical determinations with confidence limits is probably impracticable. It is important, however, that laboratories and clinicians become aware of them and think in terms of them.

The precision, i.e. confidence limits, of an analytical procedure can be determined by several methods. The most obvious is to run numerous replicates, say 30, of a sample and analyse mathematically the dispersion of results obtained. Such an analysis will be considered in detail in the subsequent section on control charts. Although this solution of the problem is valid, it is not very practicable in the routine operation of most clinical laboratories. A much simpler approach merely requires running routine unknowns in duplicate at least until about 30 duplicate analyses are accumulated. The duplication must be complete, i.e., through the entire procedure. The difference between each pair, the range R, is calculated and the arithmetic average or mean of the ranges, $\bar{R}$, is calculated. The magnitude of the confidence limit on each side of a *single* observation is then estimated by multiplying $\bar{R}$ by 2.65.*

Table I shows the calculation of the confidence limit in the determination of glucose, using only 10 pairs.

TABLE I

Blood Glucose (mg.%)	R
100 103	3
86 94	8
120 129	9
97 92	5
126 125	1
79 83	4
104 104	0
120 108	12
84 88	4
93 95	2
Total =	48

$$\bar{R}=\frac{48}{10}=4{\cdot}8$$

$$\text{Confidence limit}=\pm\ 2{\cdot}65\ (\bar{R})$$
$$=\pm 12{\cdot}7\ \text{mg.\%}$$

The confidence limit as calculated in Table I is in terms of mg.% which, when converted to per cent error by referring it to the approximate average of the range of analyses used (100 mg.%) becomes ± 12.7%. In neither of the above methods of estimating precision is the error contributed by variation in reagent blanks included in the estimate. This will be dealt with in detail in a future publication, where it will be shown that usually there is only a small increase in the limits as calculated above.

The first decision which must be made before setting up any clinical chemical procedure is the degree of precision required, i.e., how wide dare the confidence limits be. Blood glucose values will probably be satisfactory for clinical use,

* This is an application of the use of range for estimation of standard deviation. A confidence limit of three standard deviations is obtained for the mean of pairs by multiplying $\bar{R}$ by 1.88. Multiplying this by 1.4 gives the limit for a single observation, i.e. 2.65 ($\bar{R}$). Justification for the use of three standard deviation confidence limits is discussed in the section on control charts.

though perhaps not for research, if they are within 10% of the correct value for the method used. Many clinical chemical determinations are satisfactory for clinical use if they have a precision in this range. There are some determinations where an error of this magnitude obviously might be serious. The normal range for serum sodium is approximately 300 to 345 mg. per 100 ml. Thus a value in the middle of this range with confidence limits of ±7% would blanket the entire range. Even for clinical use a result with a maximum error of approximately 3% is thus required. Similarly, a value in the middle of the normal range of serum calcium, if taken as 9 to 11 mg.% with limits of ±10%, would include the whole range.

The next problem to consider is what can be done to decrease the magnitude of the confidence limits (or error) of a determination once it is found that the existing limits are too large. In designing the routine procedures of a laboratory one must stay within the bounds of practicability and still be assured of the accuracy and precision desired. Valuable time in the laboratory is wasted by complicating procedures to give greater precision than is actually required. If a procedure, however, lacks the required precision, one obvious line of attack is to use more precise and careful technique in its performance or improve the method itself, e.g., better control of variables, such as temperature, which may affect the result. Another possible approach to decreasing the error is replication, i.e., running the unknown sample in duplicate, triplicate, etc. Statistics must be called upon to answer just how much of an increase in precision is to be expected (Simon, 1941). If E is the observed confidence limit or error for a single determination and $E_{\bar{x}}$ the maximum confidence limit or error desired, then the number of replicates, N, necessary to achieve this is $\left(\frac{E}{E_{\bar{x}}}\right)^2$.

To demonstrate the magnitude of this increase in precision with replication let us presume an error of ±10% with a single sample. The error for various numbers of replicates can be found by substituting in the transposed version of the formula $E_{\bar{x}}=E/\sqrt{N}$. For one, two, three, four, and five replicates the error would be 10, 7, 5.8, 5, and 4.5% respectively. Thus the error decreases inversely as the square root of the number of replications. If the replication necessary is greater than duplicates, or perhaps rarely triplicates, this method of increasing precision becomes totally impracticable in the clinical laboratory. If neither of these approaches yields the required precision or if they are impracticable, the method must be discarded and a search made for a more precise method.

What has been said for unknown samples also holds to a certain extent for standards. Reagent blanks present the same problem, but in a much attenuated form. In the vast majority of cases the light absorption by the reagent blanks is very small compared with that absorbed by the standard and unknown, so that even a rather large relative error in the reagent blank would not often materially alter the correction factor of the blank. It is an entirely different matter if the reagent blank absorption is relatively high. As previously stated, the frequent running of reagent blanks is usually for the purpose of checking the reagents for contamination.

Control Charts

After a test has been set up and standardized and periodic standards run to check on the procedure, the simplest way to keep track of these results is merely to note them in a notebook reserved for the purpose. As expected, the results will vary from day to day, and if one value is suddenly greatly out of line suspicion will be aroused. There is a simple way, however, to treat these check data as they accumulate which has the advantage of telling the analyst at a glance how much the checks can vary before suspicion is warranted and action indicated. The use of the quality control chart will frequently predict trouble before it actually happens (Mitchell, 1947; Wernimont, 1946).

There are numerous ways in which these charts can be set up, but only a few of the simplest will be discussed, because they will serve the purpose admirably in the clinical laboratory.

Type I.—For this type single standards are run. A sheet of graph paper is employed (it is convenient to have a loose-leaf notebook of graph paper for quality control charts of the various determinations) with an appropriate scale on the ordinates for the observed result of the standard. This scale may be in terms of the final result, e.g. mg. %, but it is usually more convenient to use the density reading of the standard (corrected for reagent blank if read versus water or solvent blank). The abscissae represent successive determinations and the date is noted for each point placed on the graph. The next step cannot be taken until 20 such points have been plotted, so it is advisable to run standards frequently, perhaps once a day, at

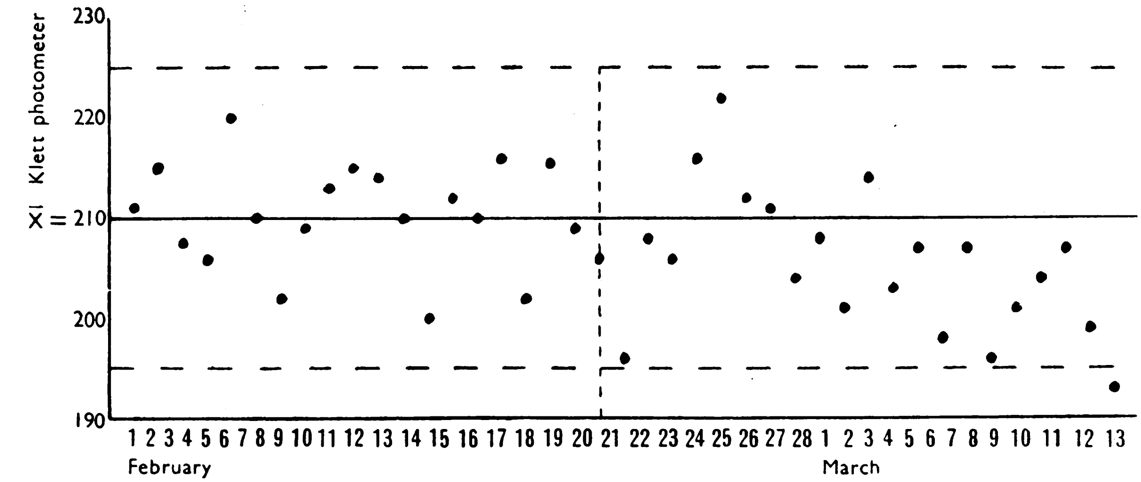

FIG. 1.—Type I control chart. Single standards.

least until this number has been reached. The arithmetic average or mean, $\bar{x}$, is then calculated from these 20 values and represents the best estimate of the true value for the standard. A horizontal line is then drawn on the graph at this point as in Fig. 1.

The next step is to establish confidence limits for the standard values. These are the same confidence limits mentioned before and constitute a range of equal magnitude on both sides of the mean, $\bar{x}$, in which approximately 95 times out of 100 the standard result should fall if *no factors are operating other than those operating during this period in which the 20 values were obtained.* These limits are calculated as follows:

If N = number of determinations

$\bar{x}$ = arithmetic mean of N determinations

Σ = " sum of "

σ = standard deviation

x = a single determination

$(\bar{x}-x)$ = the deviation of a determination from the mean, $\bar{x}$.

$$\bar{x} = \frac{\Sigma(x)}{N}$$

$$\sigma = \sqrt{\frac{\Sigma(\bar{x}-x)^2}{N-1}}$$

Often the mathematical labour is simplified by the use of another form of this equation:

$$\sigma = \sqrt{\frac{\Sigma(x^2) - \frac{(\Sigma x)^2}{N}}{N-1}}$$

As an example the calculations for the confidence limits of Fig. 1 are presented herewith:

TABLE II

(x)	$(\bar{x}-x)$	$(x-x)^2$
211	1	1
215	5	25
207	3	9
206	4	16
220	10	100
210	0	0
202	8	64
209	1	1
213	3	9
215	5	25
214	4	16
210	0	0
200	10	100
212	2	4
210	0	0
216	6	36
202	8	64
213	3	9
209	1	1
206	4	16

$\Sigma(x) = 4200$ $\Sigma(\bar{x}-x)^2 = 496$

N = 20

$$\bar{x} = \frac{\Sigma(x)}{N} = \frac{4200}{20} = 210$$

$$\sigma = \sqrt{\frac{\Sigma(\bar{x}-x)^2}{N-1}} = \sqrt{\frac{496}{19}} = 5{\cdot}1$$

$3 = 15{\cdot}3$

The confidence limits are drawn above and below the line drawn for the x equal to a distance of 3σ. If all goes well for many more determinations, say a total of 100, the $\bar{x}$ and s can be re-evaluated. It will be noted that if the variation of determinations about their mean represented the so-called " normal " or Gaussian distribution (symmetrical bell-shaped distribution), the 95% confidence limits would be represented by $\pm 2\sigma$. It has been emphasized, however, that distributions of this type are frequently not of the " normal " type (Mitchell, 1947; Clancey, 1947). Clancey (1947) analysed the variation of many chemical analyses and found the Gaussian distribution to hold in a minority of cases.

There are always two dangers when using control charts: (1) looking for trouble that does not exist, and (2) not looking for trouble that does exist. Thus, the use of 3σ as compared to 2σ may reduce the number of times that we look for trouble when it does exist, and may increase the number of times that trouble exists and we do not look for it. Taking these points into consideration, and especially when there is no *a priori* knowledge of the nature of the distribution, it is believed that the best compromise, and certainly the safest, is to associate " reasonable assuredness " with confidence limits of $\pm 3\sigma$ (Simon, 1941 ; Mitchell, 1947), especially when 3σ limits are within the desired precision and we do not want to waste time looking for trouble that does not exist. If, however, 3σ limits are too broad compared with the precision required, 2σ limits may be preferred if we are willing to spend the time to investigate out of limit points knowing that frequently in such cases trouble does not actually exist.

What should be done if a value falls outside the confidence limits? First of all it must be remembered that these are only " reasonable assuredness " confidence limits, and therefore approximately one value in every 20 to 100 should fall outside, but not far outside. If a value falls slightly outside the limits, the procedure should be checked carefully for an assignable cause—new reagent used, new standard, possible break in technique, etc. If none of these are obvious and the possible error involved, if it is an error, is not critical, you may wait to see what happens next time. If the same thing happens next time, then there is cause for alarm, since there is a good indication that there is trouble. Many times trouble can be predicted before it actually happens. In Fig. 1 it is noted that after February 26 there is a gradual drift in the distribution in a downward direction, although not until March 13 is there a value falling outside the lower confidence limit. Obviously, trouble began brewing about February 26, but not until March 13 can the test be considered out of control. This type of drift occurs when one reagent gradually deteriorates, e.g., aminonaphtholsulphonic acid reagent in the determination of phosphate.

Type II.—In a Type I control chart there is a certain amount of mathematical labour involved in the calculation of σ. This can be obviated if successive determinations are treated as pairs (Wernimont, 1946). The mean of each pair and the differences between each pair (the range, or R) are plotted as in Fig. 2. The 95% confidence limits for the $\bar{x}$ are $\pm$ 1.88 ($\bar{R}$), where the value of $\bar{R}$ is obtained by averaging the range values, R. The confidence limit for the range is 3.27 ($\bar{R}$). (Only one limit need be considered, since the other limit is obviously 0.)

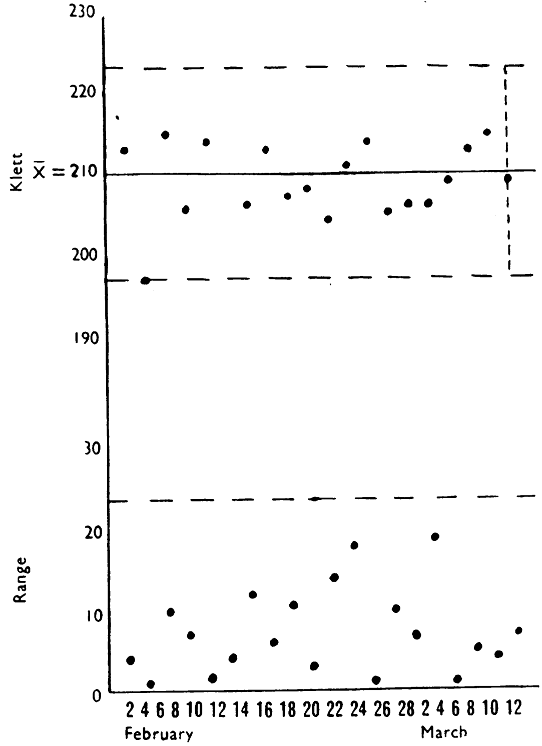

FIG. 2.—Type II control chart. Successive single standards paired.

Type III.—More information can be obtained by using duplicate standards. The difference between the pair of standards or the range, R, is calculated and plotted each time as in Type II and the confidence limits calculated as in Type II. The main advantage of this procedure is that it is possible to tell whether the variation between days is significantly greater than the variation in any one run. With many chemical procedures there is a definite tendency for this to be so, signifying the presence of insufficiently controlled variables, such as temperature. Such a situation would be indicated when the range values are apparently in control but the averages of the duplicates out of control. This type of control chart has been suggested for use in the clinical laboratory by Levey and Jennings (1950), who give some examples of its use and how it detects errors which develop in procedures. Whether the added labour in running duplicate standards is worth while depends on the individual situation.

No matter which type of control chart is used, the value for the standard used in the calculations of the unknowns is the mean $\bar{x}$, not the value for the standard obtained on the same day or even in the same run with the unknowns.

If unknown samples are run singly, the error should be the same as that determined for standards if a Type I control chart is used, *provided the unknowns and standards are treated the same.* Similarly, if unknowns are run in duplicate the error of their average should be the same as that determined for standards run in duplicate (Type III). This assumes that the error is the same for unknowns as for standards and is approximately correct if the per cent transmission for both is in the range of 20 to 60% (optical density 0.2 to 0.7). When pure standards are used and not run through the entire procedure the confidence limits set for the standards obviously cannot be used as an estimate of the precision of unknowns.

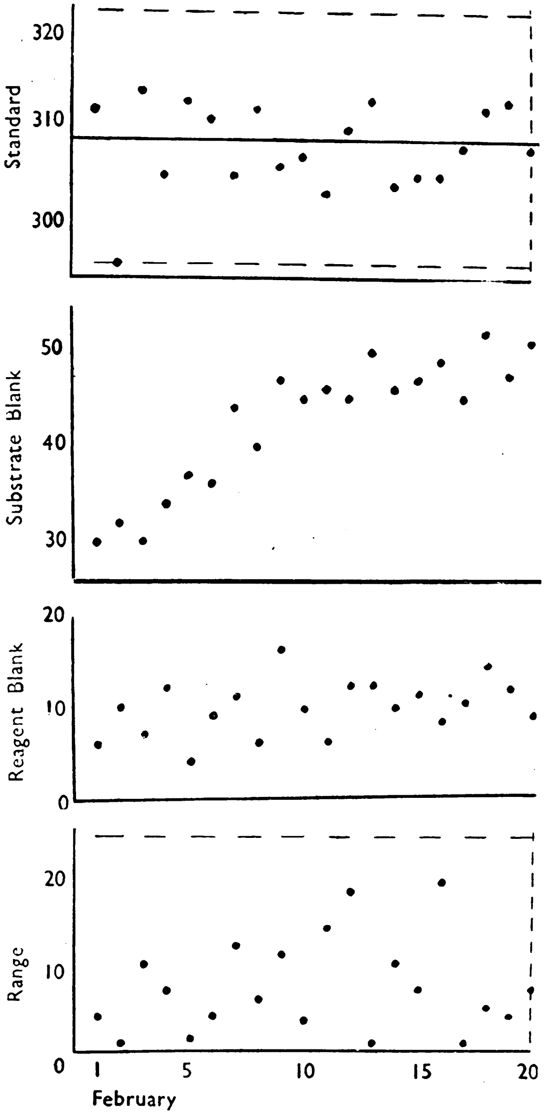

FIG. 3.—Type III control chart with reagent and substrate blanks recorded.

In a procedure such as the determination of serum phosphatase the substrate blank, the reagent blank, and the range for standard, if a Type II or III control chart is being used, can all be plotted together as in Fig. 3.

There is no practical way, and in many cases it is impossible, to be sure that the result of any one determination is correct even with the running of standards and blanks. The procedures outlined here, however, will tell in many cases the accuracy to be expected by a procedure, the precision in all cases, and the reliability as indicated by quality control charts.

There are other factors of importance in setting up determinations which are outside the scope of this paper and have received adequate treatment elsewhere (Archibald, 1950; Ayres, 1949). These include the proper choice of concentration range (yielding final optical densities of 0.2 to 0.7), and the establishment of whether or not the colour produced follows Beer's law.

Summary

It is considered absolutely essential to run frequent standards in all colorimetric clinical laboratory analyses. The various types of standards that can be run are discussed, as well as the estimation and control of accuracy and precision, and the role of duplication in the control of precision. The employment of control charts is advocated, since their use allows the laboratory to be aware at all times of the state of control of the test.

The authors wish to express their appreciation for the advice given by Frank B. Cramer on some of the statistical points.

REFERENCES

Archibald, R. M. (1950). *Analyt. Chem.*, **22**, 639.
Ayres, G. H. (1949). *Ibid.*, **21**, 652.
Belk, W. P., and Sunderman, F. W. (1947). *Amer. J. clin. Path.*, **17**, 853.
Clancey, V. J. (1947). *Nature, Lond.*, **159**, 339.
Fiske, C. H., and Subbarow, Y. (1925). *J. biol. Chem.*, **66**, 375.
Levey, S., and Jennings, E. R. (1950). *Amer. J. clin. Path.*, **20**, 1059.
Mitchell, J. A. (1947). *Analyt. Chem.*, **19**, 961.
Shuey, H. E., and Cebel, J. (1949). *Bull. U.S. Army med. Dep.*, **9**, 799.
Simon, L. E. (1941). *An Engineers' Manual of Statistical Methods.* New York.
Wernimont, G. (1946). *Industr. Engng Chem. Anal. Ed.*, **18**, 587.