Design and Evaluation of Statistical Control Procedures: Applications of a Computer "Quality Control Simulator" Program

James O. Westgard¹ and Torgny Groth²

A computer simulation program has been developed to aid in designing and evaluating statistical control procedures. This "QC simulator" (quality control) program permits the user to study the effects of different factors on the performance of quality-control procedures. These factors may be properties of the analytical procedure, characteristics of the instrument system, or conditions for the quality-control procedure. The performance of a control procedure is characterized by its probability for rejection, as estimated at several different magnitudes of random and systematic error. These performance characteristics are presented graphically by power functions—plots of the probability for rejection vs the size of the analytical errors. The utility of this simulation tool is illustrated by application to multi-rule single-value procedures, mean and range procedures, and a trend analysis procedure. Careful choice of control rules is necessary to minimize false rejections and to optimize error detection with multi-rule procedures. Control limits must be carefully calculated for optimum performance of mean and range procedures. The level of significance for testing control must be carefully selected for the trend analysis procedure.

Additional Keyphrases: variation, source of - statistics

The current practices in quality control in clinical laboratories are not much different from those introduced in the 1950s. Levey and Jennings (1) described procedures based on the mean and range of duplicate measurements, following Shewhart’s classical procedures (2). Henry and Segalove (3) simplified these by using single measurements instead of duplicates and by plotting those single values directly on a control chart. Today this "single-value" chart is probably the predominant control procedure in clinical laboratories in the United States, though there is some use of mean-and-range control procedures. Other control procedures such as cusum (4) have been tried but have not found widespread acceptance, most likely because of the additional effort required for data calculations and the difficulty in judging when the analytical process is out of control.

In contrast to this rather static state in the development of quality-control techniques, measurement techniques have continued to evolve. Today’s instrument systems are much more sophisticated than those from the 1950s. Recent advances incorporating microprocessors into the hardware of many instrument systems have provided many improvements in controlling instrument functions, and one would also expect improvements in controlling the quality of the instrument measurements, such as the introduction of more sophisticated control procedures, with data analysis and interpretation provided by the instrument system itself. One example is the application of the algorithm of Bull et al. (5) in some of the new cell-counting instruments.

Improved statistical quality-control procedures have yet to be implemented on many instrument systems, but there is an increasing awareness of the potential for making such improvements with microprocessor-controlled instruments. Microcomputers will offer an even greater opportunity for introducing improved statistical-control procedures in many clinical laboratories.

One present difficulty concerns the design and evaluation of new control procedures. This problem has traditionally been analyzed in mathematical and statistical terms that are not generally familiar to clinical chemists, and this work has usually been done by theorists who may not be familiar with analytical methodology. As a result, the theoretical models may be oversimplified, either because important characteristics of the analytical method are overlooked, or because the mathematical complexity becomes too great when the detailed model is to be treated. Consequently, the performance characteristics of the statistical-control procedures have seldom been adequately characterized.

The performance of statistical-control procedures is described in the quality-control literature in terms of the average number of analytical runs before the control procedure gives a rejection signal. When an analytical method is in a stable state of operation, this average run length (ARL) should ideally be long; when the operation is unstable, average run length should ideally be short. ARLs have not often been presented in the clinical chemistry literature, though there are some examples, such as their use by Rowlands et al. (6) to compare the performance of cusum and Shewhart-type control procedures.

In describing performance of control procedures in the clinical chemistry literature (7), we have preferred to describe the probability for rejecting an analytical run. This probability for rejection is related to average run length, being its reciprocal (p = 1/ARL). Performance of a control procedure is described by its probability for false rejection and its probability for error detection. Ideally, the probability for false rejection should be low (near 0.00) and the probability for error detection should be high (near 1.00).

Information on the performance of control procedures can be summarized graphically by plotting the probability for rejection vs the size of the analytical errors. We have called such plots "power functions" because they describe the statistical power of the control procedure (8). Saracci (9) also used the term "power" in this manner to describe the sensitivity of different quality-control plans.

We have presented power functions for many commonly used control procedures (8). These provide some general information about the performance of different control procedures, but do not describe the exact performance when a particular control procedure is applied to a particular analytical method. Unfortunately, the statistical power of a
control procedure depends on several factors, which vary from one application to another. Some of these factors are analytical characteristics, such as the relative sizes of within- and between run standard deviations; some are instrument characteristics, such as the number of significant figures in the measurement; and some are the conditions chosen for the control procedure itself, such as the number of control observations, the choice of control limits (calculated from the within-run, between-run, or total standard deviations), and the decision criteria (control rules).

To take these many factors into account, we have developed a computer simulation tool to aid clinical chemists in designing and evaluating statistical-control procedures (10). This "QC simulator" is an interactive program that allows the user to describe the conditions for the measurement and control procedures, after which the program simulates several hundred analytical runs conforming to these conditions. Different amounts of analytical error are introduced in these runs and the probability for rejection is estimated, and power functions are plotted.

For improving the design of statistical control procedures, we recommend the following approach (11). First, study the probability for false rejection and eliminate those control procedures having high values for this. Then consider those remaining procedures that have the best error detection. Select at least one control procedure or control rule that is responsive to random error and at least one that is responsive to systematic error. Assess the probability for error detection for this combination of rules, describing this by power functions. From these power functions select the number of control observations that will provide the desired probability for error detection. This iterative design process permits clinical chemists to explore the performance of different statistical-control procedures, an experimental type of approach with which they are both comfortable and skilled.

We have also developed a more formal design approach (12) in which a quality goal or quality specification is the starting point. From that quality specification, the sizes of errors that should be detected are calculated. Then, using power functions, one can determine which control procedures and what number of control observations are necessary to assure that the quality specification is achieved. Although this design procedure is in principle more objective, the lack of quality specifications is currently a limitation. Consequently, here we emphasize applications in which the informal, iterative design procedure is used; however, the simulation program provides for both approaches.

Materials and Methods

Simulation Program

The program has been described in detail by Groth et al. (10). It is written in FORTRAN IV for an IBM 370/158 computer run under OS/MVT and the Gothenburg University Terminal System (13). It requires 105 kbytes of memory, makes use of the IMSL library (14), and has several subroutines specific for the computer installation at the Uppsala Data Center. Another version of the program is written in ANSI FORTRAN and has been installed at the Madison (WI) Academic Computing Center, where it runs on a Sperry Univac 1100/82.

The "QC simulator" is an interactive program that queries the user for a series of input parameters needed to describe both the analytical procedure and the control procedure. The analytical parameters are used to set up a random-number generator, which simulates test data, whereas the control parameters are used to select the data analysis or quality-control testing of that data. The program provides two modes of operation. Mode A permits the user to determine the power functions for a particular set of analytical conditions and a particular design of a control procedure. Mode B permits the user to start with a quality specification and determine which designs from a series of recommended control procedures will achieve that level of quality.

When in use, the simulation program generates representative control data, which are then inspected to determine when a given control procedure gives a rejection signal. Four hundred analytical runs are simulated and the proportion of runs rejected is taken as the estimate of the probability for rejection. To study the response of the control procedure to systematic errors, systematic shifts are introduced that correspond to errors of 0.0, 0.5s, 1.0s, 1.5s, 2.0s, 2.5s, and 3.0s (where s is the total standard deviation of the analytical method). To study the response of a control procedure to random error, the within-run standard deviation is increased by multiplying s by 1.0, 1.25, 1.5, 1.75, 2.0, 2.5, and 3.0. For each of these error conditions, 400 runs are simulated to estimate the probability of rejection. Two power functions for the control procedure are plotted, one describing the response of the control procedure to systematic error and the second describing the response to increases in random error.

In mode A, the user enters values describing the stable analytical performance of the method, such as the total standard deviation (s), the relative sizes of the within-run (sw) and between-run (sb) standard deviations, the mean concentration of the control material, possible uncertainty in the estimate of the total standard deviation, and the number of decimal places in the results reported. A "menu" of control rules is presented, giving 14 possible choices for detecting random error and 20 possible choices for detecting systematic error. The user can select one or more rules from each category. For each rule chosen, the user must specify whether the control limits are to be calculated from sw or sb, or an appropriate combination of sw and sb. When rules for trend analysis are selected, additional information must be given to define the alpha value used in exponential smoothing and the level of significance for setting control limits. Finally, the number of control observations (N) is entered, the simulations are performed, and the probabilities for rejection for different error conditions are printed.

In mode B, the user starts by entering a quality specification (a 95% limit of allowable error), using the same concept of total error as used by Westgard et al. in method-verification testing (15). The user also enters the concentration at which medical interpretation is most critical, describes the analytical procedure in a manner similar to mode A, but provides additional information on method bias. The program calculates the sizes of the random and systematic errors that must be detected to maintain errors within the quality specification (12, 13). Next a menu of recommended combinations of control rules is presented. The default state is that all the recommendations will be tested; however, the user can eliminate any that would not be of interest, owing to practical limitations in implementing such procedures. Finally, the user must specify the desired probability for error detection, and the computer simulations start. The printout lists those control procedures tested, the number of control observations necessary to detect the critical errors at the desired probability, and the probability for false rejection under these conditions.

3 The imprecision of the measurement procedure is generally estimated from a replication experiment. By proper design and with application of analysis-of-variance techniques, the within-run and between-run components of variance may be estimated; sw and sb denote the corresponding standard deviations for the components, and s stands for the total standard deviation of the measurement procedure.
Control Procedures

The use of the QC simulator is illustrated by application to the design and evaluation of the following control procedures.

Multi-rule single-value control procedures. We use the name "multi-rule single-value" to describe a control procedure whereby the control observations are plotted directly on a control chart having more than one set of control limits, or more than one decision criterion or control rule. One example of a multi-rule control chart has been described in detail (16); that chart was designed for two or four control observations per run and for manual implementation. Other designs may be better when the number of control observations is different and when the procedure is implemented with computerized data handling.

Two extensions of the multi-rule single-value procedures have been investigated here and the results are shown in the Figures.

(a) Application when N = 3. This is of particular interest when there are three different concentrations of control materials and each is analyzed once. This is often the case for blood gas measurements where control materials are available in sets corresponding to alkalosis, normal, and acidosis specimens. Several possible control rules could be used in adapting the multi-rule procedure for three observations: 4

1s Reject when one control observation exceeds control limits set as the mean ± 3s.

(2 of 3)s Reject when two of three consecutive control observations exceed the same limit, which is either the mean + 2s or mean − 2s control limit.

3s Reject when three consecutive control observations exceed the same limit, which is either the mean + 1s or mean − 1s control limit.

R3 Reject when one control observation (in the group of three) exceeds the mean + 2s control limit and another exceeds the mean − 2s control limit.

(b) Application with exact range rule. The approximate range rule (R3s) is convenient for manual applications because it eliminates any data calculations. This rule is violated when one observation exceeds the +2s limit and another the −2s limit, but it would not signal a range error when one observation is ±2.5s and the other is −1.6s. The control limit of 4s is approximate and is used with N from 2 to 4, even though the exact limit depends on N and the chosen level of significance (or selected probability for false rejection). These limitations would be reduced by calculation of the exact range and use of exact control limits, such as provided by the following control rules:

R3 Reject when the difference between the highest and lowest values in a group of N control measurements exceeds a control limit set such that the probability for false rejection is 0.05.

R3 Reject when the difference between the highest and lowest values in a group of N control measurements exceeds a control limit set such that the probability for false rejection is 0.01.

Use of an exact range rule is possible with computerized data handling and may provide improved detection of random error. This application has been investigated as an off-line quality-control procedure for an automated chemistry analyzer.

Shewhart mean and range control procedures. This is the classical Shewhart procedure (2), in which the mean and range are calculated from a group of control measurements. The range rules are the same as the "exact" range rules defined above. The mean rules similarly have limits that are carefully set to provide a specified probability for false rejection:

R0.05 Reject when the mean of a group of N control measurements exceeds control limits set such that the probability for false rejection is 0.05.

R0.01 Reject when the mean exceeds control limits set such that the probability for false rejection is 0.01.

Mean and range procedures have been implemented for on-line monitoring of a multi-channel chemistry analyzer. Control measurements are made on three different control materials, the results are normalized by comparing each with the mean and standard deviation for that material, and the normalized results are then processed to determine the mean and range. These procedures are applied to the two to six control measurements obtained during one calibration cycle of the instrument. This procedure would be expected to provide good statistical power; however, the design may need to consider how performance depends on the between-run and between-material components of variance. Optimum performance may require that control limits for the range be calculated from the within-run standard deviation (sN) (17) and that control limits for the mean be set from an appropriate calculation involving the between-run (sN) and within-run standard deviations (18):

\[ s = (s^2_N + s^2_b/N)^{1/2} \]

Trend analysis. Cembrowski et al. (19) described a trend analysis control procedure that recently has been implemented on a microcomputer.5 In this application (Micro-Sure®, Multi-Sure®; Laboratory Consulting, Inc., Madison, WI 53701) color graphics display the control data and indicate control status, first signaling a yellow warning and then a red alert. This adaptation is somewhat different from the original application by Cembrowski et al. in that the two estimates that were combined into a ratio to give the "Trigg's tracking signal" are treated separately, one as an "accuracy trend" indicator and the other as a "precision trend" indicator. This procedure is analogous to the classical Shewhart control procedure in which the mean is used to monitor systematic error and the range or standard deviation to monitor random error. However, the estimates of the mean (accuracy trend) and standard deviation (precision trend) are obtained by an exponential smoothing technique.

This exponential smoothing calculation weights recent control observations more heavily than past observations, thus increasing the response to trends that are just beginning. The equations for the calculations are more complicated, but the technique does not require storage of the original control observations. Instead, new estimates of the mean and standard deviation are made each time a new control observation is obtained, and only the updated estimates need be saved. In addition, these estimates can be used for immediate and continuing assessment of control, in contrast to the end-of-run information provided by the classical mean-and-range procedures.

The estimate of the mean obtained by exponential smoothing, hereafter called the smoothed mean (sm-mean), is calculated as follows: the new control observation is multiplied by a smoothing constant (alpha), and to this product is added a second term, which is the previous estimate of the smoothed mean multiplied by 1 − alpha.

---


sm-mean = alpha (new obs)
+ (1 - alpha)(previous sm-mean)

When alpha is set to 0.10, for example, the following equation is obtained:

sm-mean = 0.1 (new obs) + 0.90 (previous sm-mean)

The new observation would be weighted by 0.1 and the previous smoothed-mean by 0.9.

A similar calculation is performed for the standard deviation, though this is more complicated because the difference between the new control observation and the estimate of the mean must be calculated first. This difference is called the forecast error:

forecast error = new obs - previous sm-mean

The forecast error is then processed by an exponential smoothing calculation to give an estimate of precision called the mean absolute deviation (MAD).

MAD = alpha (new forecast error)
+ (1 - alpha)(previous MAD)

Because MAD is proportional to the standard deviation, the smoothed standard deviation (sm-s) can be obtained by multiplying MAD by a factor that depends on alpha. This factor is obtained as follows:

factor = [0.25\pi (2 - alpha)]^{1/2}

When alpha equals 0.10, this factor becomes 1.22, and sm-s = 1.22 MAD.

Initially, sm-mean is set equal to the “true” mean (T-mean) and sm-s to the “true” standard deviation (T-s). The “true” values are those values obtained when the quality-control materials are analyzed under conditions of stable analytical performance.

To assess control status, the estimate of the smoothed-mean is tested against the T-mean or established mean for the analytical method by using a Z-test.

Z = (N)^{1/2} (sm-mean - T-mean)/T-s

This statistical test is analogous to setting limits for the mean in the same way that limits have been set in the classical Shewhart procedure. This is more obvious when the equation is rearranged as follows:

sm-mean - T-mean = Z \cdot T-s/N^{1/2}

where Z corresponds to the number of standard deviations and is related to the “level of significance” for the statistical test.

When a chosen Z-value is substituted into this equation, control limits for the difference between sm-mean and T-mean are obtained. The limits for the yellow warning signal correspond to a Z-value of 2.58, or the 1% level of significance. For an alpha of 0.10, sm-mean corresponds to a classical mean calculated from 20 observations, thus N is taken as 20.

sm-mean - T-mean = 0.58 T-s

Limits for the red alert signal correspond to a Z-value of 3.09, or the 0.2% level of significance.

sm-mean - T-mean = 0.69 T-s

Because the observations on control materials having different concentrations are expressed as the number of standard deviations from the mean of that material, T-mean has a value of 0.00 and T-s has a value of 1.00. Therefore, the yellow warning signal occurs when sm-mean exceeds 0.58 and the red alert signal when sm-mean exceeds 0.69.

The estimate of the smoothed standard deviation (sm-s) is tested for significant change by using a chi-square statistical test, which compares the observed variance (sm-s squared) with the true variance (T-s squared):

\[ \text{chi-square} = (\text{sm-s}^2/T-s^2)(N - 1) \]

This equation can be rearranged to provide control limits for sm-s.

\[ \text{sm-s} = \left( \frac{\text{chi-square}}{(N - 1)} \right)^{1/2} \text{T-s} \]

With alpha set to 0.10, corresponding to N equal to 20, the critical chi-square values are 36.2 (at a 1% level of significance, and N - 1 or 19 degrees of freedom) and 43.8 (at a 0.2% level of significance). The yellow warning signal occurs when the estimate of the smoothed standard deviation (sm-s) is 1.38-fold greater than the “true” standard deviation.

\[ \text{sm-s} = (36.2/19)^{1/2}\text{T-s} = 1.38\text{T-s} \]

The red alert signal occurs when sm-s exceeds 1.52 times T-s.

\[ \text{sm-s} = (43.8/19)^{1/2}\text{T-s} = 1.52\text{T-s} \]

Because the control observations are expressed as the number of standard deviations from the mean, T-s is 1.00. The yellow warning signal occurs when sm-s exceeds 1.38 and the red alert signal when sm-s exceeds 1.52.

Results

Some simulations of the multi-rule single-value procedures are shown in Figure 1. When N equals 3, the designs to be considered make use of the 1\text{st}, (2 of 3)\text{st}, 3\text{rd}, and R\text{th} rules. The performance of such procedures is shown in Figure 1, where the different lines represent different designs. The detection of systematic error is shown in Figure 1a-c. The probability for false rejection increases when the 3\text{rd} rule is included, thus limiting any control procedure involving that rule to applications where the between-run standard deviation is small.

Procedures involving the R\text{th} control rule show an increase in false rejections as the between-run component gets very large, limiting applications of such procedures, though not as severely as for procedures where the 3\text{rd} rule is included. Because the R\text{th} rule does improve the detection of random error, as shown in Figure 1d, it is desirable to include this rule whenever possible.

All of these simulations indicate that a combination of control rules can increase error detection over that obtained with the 1\text{st} rule alone. The detection of systematic error is improved by addition of the (2 of 3)\text{rd} rule. The detection of random error is improved by the addition of the R\text{th} rule.

The results of more detailed studies on the performance of range rules are shown in Figure 2. When the between-run standard deviation is zero, performance of the R\text{th} approximate range rule is similar to that of the R_{0.05} and R_{0.01} exact range rules, as shown in Figure 2a. When the between-run standard deviation equals the within-run standard deviation, more notable differences in performance are seen (Figure 2b). This figure shows how performance changes when control limits are calculated from the within-run standard deviation or from the total standard deviation. The probability for error detection is highest when the R_{0.05} rule is used with control limits calculated from the within-run standard deviation. Use of the R_{0.01} rule with control limits calculated from the within-run standard deviation reduces false rejections but also causes some reduction in error detection. Error detection is still better than with the R_{0.05} rule and limits calculated from the total standard deviation. The R\text{th} procedure is clearly less responsive for error detection, though it is better than the R_{0.01}
When the between-run component is twice as large as the within-run component, the differences in performance become more pronounced (Figure 2c). Use of exact range rules with control limits calculated from the within-run standard deviation clearly provide better performance.

In the simulations in Figures 2a–c the example analytical procedure was a creatinine method having a total standard deviation of 0.04 mg/dL on a control material with a mean of 1.5 mg/dL. These simulations further specified that the analytical method provided results significant to two decimal places (DP = 2). This number of significant figures is seldom obtained because data are usually rounded to one decimal place for clinical relevance. The consequence of this rounding is illustrated in Figure 2d, which shows the performance of the R_{0.05} and R_{4s} rules when results are obtained to only one decimal place. Note that the R_{0.05} rule gives about a 30% level...
Fig. 2. Power functions for some range control procedures, where the number of control measurements (N) is 4 and the between-run standard deviation is (a) zero, (b) equal to the within-run standard deviation, or (c and d) equal to two times the within-run standard deviation

(a–c) two decimal places in the reported results, (d) one decimal place. Control limits (CL) indicated by $s_i$ have been calculated from the total standard deviation, and those indicated by $s_w$ from the within-run standard deviation.

of false rejections, compared with 5% in the Figure 2c. This is due to data rounding.$^6$ The $R_{4s}$ procedure provides a suitably low level of false rejections, and in this situation performs better than the exact range rules.

The results in Figures 3 and 4 show how the performance of Shewhart's mean rule is affected by a between-run component and the calculation of control limits. The number of control observations is set to 6 for these simulations, corresponding to the N expected during an analytical run of the analyzer for which the control procedure was being designed. Figure 3 shows the performance of a mean rule having control limits calculated to provide a 5% level of false rejections. When the between-run standard deviation is zero, the observed false rejection rate is about 5%, as expected. When the between-run standard deviation equals the within-run standard deviation, performance depends on how the control limits are calculated.

When calculated from the total standard deviation, the level of false rejections is high, nearly 30%. When calculated from the between-run standard deviation and the within-run standard deviation (see equation under Materials and Methods), the expected 5% false rejection rate is obtained, but there is some loss in error detection.

This loss in error detection is shown in Figure 4. The figure compares the performance when the between-run standard deviation is zero with the performance when the between-run standard deviation is equal to the within-run standard deviation. The family of lines at the left all show greater error detection than the family of lines to the right. Observe that the N = 2 line for a zero between-run component shows better performance than the N = 6 line when the between-run component is equal to the within-run component.

Figures 5–8 show the simulation results for the precision and accuracy trend analysis procedures. With alpha set as 0.10 and statistical significance tested at a 1% level, the power functions are as shown in Figures 5 and 6. When tests of significance are carried out at a 0.2% level, the power functions are as shown in Figures 7 and 8. These Figures show that the trend analysis procedures offer high error-detection capabilities, but attention should also be paid to the probability for false rejection. When these trends are tested at the 1% significance level (the "yellow warning" signal), the probability for false rejection is approximately 15%. When tested at the 0.2% significance level (the "red alert" signal), the false rejection level is 6–7%, which is much better. Error detection is greater than attained with the classical Shewhart mean and standard deviation procedures, demonstrated here by comparing the rules tested at the 0.2% level of significance (the dotted lines in Figures 7 and 8).

Discussion

The applications we have presented were chosen to illustrate the use of computer simulation to determine the performance characteristics of different quality-control procedures. The simulation results were obtained from an interactive "QC simulator" program that can be used by clinical
chemists who have no special background in computers, simulation studies, or statistics and probability theory.

Multi-rule single-value procedures can be designed to provide better performance than existing single-rule procedures, provided the control rules have been properly selected. Certain control rules may cause the probability for false rejection to be too large, as illustrated in the example where the \( R_{14} \) control rule resulted in a high proportion of false rejections when the between-run standard deviation increased. The probability for error detection may be improved by addition of certain control rules, as shown by the effects when the \( R_{14} \) and the (2 of 3)\(_{2a} \) rules are added to the 1\(_a\) rule. The 1\(_a\) rule improves the detection of random error and the (2 of 3)\(_{2a} \) rule improves the detection of systematic error.

Mean and range procedures must also be carefully designed to take into account the effects of a between-run component of variation. The simulation studies demonstrate that the performance of a range procedure will be optimum when control limits are calculated from the within-run standard deviation. Performance of a mean procedure will be optimum when the control limits are calculated from a standard deviation that appropriately weights the within-run and between-run components of variance.

The trend-analysis example illustrates the use of computer simulation to evaluate the performance characteristics of a more sophisticated control procedure involving an exponential smoothing technique for estimating the mean and standard deviation. The simulations show that the technique provides error detection comparable with that of mean and range (or standard deviation) procedures involving large numbers of measurements. There are differences in the calculation and testing procedures, most notably that the trend-analysis procedures provide repeated testing of the control data each time a new measurement is added, whereas the classical mean and range procedures test the data only once, after all the required measurements have been obtained. This repeated testing provides control information that is more immediately useful, but also contributes to a higher proportion of false rejections. Consequently, significance levels for the statistical

---

Fig. 6. Power functions for the "precision trend analysis" procedure, where the exponentially smoothed standard deviation is tested by a chi-square rule having limits set for the 0.01 (1%) level of significance
Alpha is 0.10 (corresponding to an N of 20)

Fig. 7. Power functions for the "accuracy trend analysis" procedure when the exponentially smoothed mean is tested at the 0.002 (0.2%) level of significance
Alpha is 0.10 (corresponding to an N of 20). The dotted line is for the classical Shewhart mean rule with \( N = 20 \) and 0.002 significance

Fig. 8. Power functions for the "precision trend analysis" procedure when the exponentially smoothed standard deviation is tested at the 0.002 (0.2%) level of significance
Alpha is 0.10 (corresponding to an N of 20). The dotted line is for the classical Shewhart standard deviation rule (chi-square rule) with \( N = 20 \) and 0.002 significance
tests must be carefully chosen. If the tests of significance were applied at the 5% significance level, the rate of false rejections would be very high. The 1% significance level seems appropriate for a warning signal, and the 0.2% significance level is more suitable for an alert signal.

These examples illustrate that computer simulation can be used to evaluate a wide variety of control procedures. Application can be extended to other computational algorithms, including algorithms with patients' data. For example, the usefulness of the anion-gap calculation as a quality-control procedure has recently been investigated. Routine patients' electrolyte data were analyzed to characterize the distribution of the anion gap and of individual electrolytes, and to assess any correlations between the individual measurements. This information was used to construct a simulation model that would generate numerical data with these same characteristics. To study the response of the anion-gap calculation, analytical error was introduced into the individual electrolyte measurements of the model. The probabilities of exceeding certain anion gap limits were determined and plotted vs the size of the analytical errors to give power functions. These plots demonstrated that anion gap is not very responsive for monitoring the analytical errors occurring with the individual patient measurements. They also permitted comparison of the anion gap procedure with conventional control procedures, which show considerably more power for error detection.

Other algorithms for patients' data could be studied similarly. It would be particularly interesting to determine the performance characteristics for Bull's algorithm for erythrocyte indices (5). There have been attempts to evaluate the performance of this and similar algorithms by experimentally comparing them with other control procedures (20, 21). Such evaluations are difficult to perform because it is nearly impossible to determine which rejections are true rejections and which are false rejections. It would be better first to determine the performance characteristics of the control procedure (from computer simulation studies or probability theory), compare these with the performance characteristics of other control procedures, and then decide whether the new procedure would provide improved error detection or fewer false rejections. Laboratory studies can provide useful experience for assessing practical limitations and developing guidelines for problem solving, such as illustrated by Koepke and Protexor's report on their experience with Bull's algorithm (22). However, such studies do not establish the statistical performance characteristics of a new control procedure.

Besides demonstrating the usefulness of computer simulation for the design and evaluation of statistical control procedures, we think the applications provide insight into some of the limitations of present analytical methods. Some limitations to the improvement of quality control are due to the design of instrument systems, not to the design of control procedures alone. Instrument systems have not been designed with proper understanding of the relationship between instrumental parameters and the performance of statistical-control procedures. The effect of data rounding is an obvious example; however, the effects of a between-run component are more widespread and more serious. This between-run component is probably related to systematic errors that vary from run to run, rather than true random errors (23). Some sources of systematic errors can be identified and the between-run errors can be reduced. One of those sources of error undoubtedly is the calibration procedure being used. Improving calibration will reduce the between-run standard deviation. Addition of a calibration module to the simulation program would perhaps be one way to study effectively the interactions between different calibration designs and the performance of quality-control procedures.

We have presented this discussion to emphasize that those persons making improvements in analytical methodology must consider both the measurement procedure and the control procedure, and the interactions between the two. With the potential applications of micro-processors, improvements in quality control should be achieved. Computerized simulation provides one approach that should be useful for designing new control procedures and for evaluating their performance.

We thank Professor Carl Henrik de Verdier for his continuing support and encouragement. The computer program was developed at Uppsala University Data Center, financed by funds from the University to the Medical Faculty. Hans FALK provided the programming expertise. Alice Erickson adapted the program to run at the Madison Academic Computing Center. We thank Marlene Brown for preparing the manuscript and George Cembrowski and Patricia Barry for their helpful comments.

References


