Estimating the Performance Characteristics of Quality-Control Procedures When Error Persists until Detection  
Curtis A. Parvin

The concepts of the power function for a quality-control rule, the error detection rate, and the false rejection rate were major advances in evaluating the performance characteristics of quality-control procedures. Most early articles published in this area evaluated the performance characteristics of quality-control rules with the assumption that an intermittent error condition occurred only within the current run, as opposed to a persistent error that continued until detection. Difficulties occur when current simulation methods are applied to the persistent error case. Here, I examine these difficulties and propose an alternative method that handles persistent error conditions effectively when evaluating and quantifying the performance characteristics of a quality-control rule.

Additional Keyphrases: data handling, variation, source of statistics

The customary approach to evaluating the performance of a quality-control rule is to determine the probability of the rule's giving a rejection signal when no analytical error is present (the probability of false rejection, \(p_r\)), or giving a rejection signal in the presence of error (the probability of error detection, \(p_{ed}\)).\(^1\) A convenient way to convey this information is in a power function graph of the probability of rejection vs the magnitude of the error condition (1, 2). When an out-of-control state is assumed to occur only in an individual run (an intermittent error), the concept of \(p_{ed}\) is unambiguous and easily understood (3). When an out-of-control state persists from run to run until detected and corrected, however, the meaning of \(p_{ed}\) has to be expanded to incorporate the number of runs for which the error has persisted and gone undetected. Even then, at least three different probabilities can be defined that relate to error-detection capability. First, there is the probability that an error has gone undetected up to the current run, but is detected during this run. Second, there is the cumulative probability that an error is detected during or before the current run. Third, there is the conditional probability that the error is detected during the current run, given that it has not been detected before this run. These three probabilities are defined in more detail in Methods.

For simple control rules, these three probabilities can be calculated exactly. More complicated rules, e.g., Westgard's multi-rule procedure (4), usually involve simulation techniques. The simulation approach currently used for evaluating the error detection capabilities of a quality-control rule to detect a systematic shift (1, 5–7) proceeds as follows:

1. Set the number of control samples per run (\(N\)), the true concentration of the control (\(\mu\)), the inherent imprecision of the analytical method (\(\sigma\)), and the magnitude of the systematic error (SE).

2. Simulate a control observation (C) by adding the systematic shift and an inherent random-error component to the true value, \(\mu\). That is, \(C = \mu + SE + \epsilon\), with \(\epsilon\) having a Gaussian distribution with zero mean and standard deviation = \(\sigma\).

3. Repeat step 2 for the required number of controls per run.

4. Apply the Quality-control rule and determine whether the run is rejected.

5. Repeat steps 2–4 numerous times (400–1000 is common).

6. Calculate the fraction of runs that are rejected. When there is no error, this provides an estimate of the \(p_{tr}\), whereas, for a systematic error, \(p_{ed}\) is estimated.

Although acceptable for evaluating intermittent error detection, this simulation approach is less than ideal for evaluating detection of persistent errors. In this paper, I describe an alternative simulation approach and use it to demonstrate several important concepts related to the characterization of error detection when error persists until detection.

Methods

The unconditional probability that an error has persisted and gone undetected for \(i - 1\) runs (\(i - 1\) runs accepted) and is detected on the \(i\)th run (\(i\)th run rejected) is denoted by \(P(R_i)\). The conditional probability that the error is detected on the \(i\)th run, given that the \(i - 1\) previous runs with error have been accepted, is denoted by \(P(R_i|A_{i-1})\). The cumulative probability that the error has been detected on or before the \(i\)th run is denoted by \(P(\leq R_i)\). The appendix includes a recursive formula that permits exact calculation of the three probabilities, \(P(R_i), P(R_i|A_{i-1}),\) and \(P(\leq R_i)\), for individual counting rules as a function of run number. For the more complicated multi-rule procedures, I use a simulation that allows direct estimation of the three probabilities. This simulation approach proceeds as follows, with steps 1–4 being as listed previously:

5. Repeat steps 2–4 until a rejection is encountered, and store the run number at which rejection occurs.
6. Repeat steps 2–5 numerous times.
7. Calculate \( P(R_i) \) as the fraction of the total trials that were rejected during the \( i \)th run; \( P(R_i | A_{i-1}) \) as the number of trials rejected during the \( i \)th run divided by the number of trials that had still not been rejected through run \( i - 1 \); and \( P(\leq R_i) \) as the sum of \( P(R_i) \) from 1 to \( i \).

Table 1 illustrates example calculations by the above simulation approach for these three probabilities from a simulation of 1000 trials.

The current simulation approach is based on the assumption that SE exists from the first run. An interesting alternative to this “startup” case is the possibility that SE occurs during ongoing operations after a period of being in control. To accommodate the “ongoing” case, the following modifications can be made to the simulation strategy:

1. Set \( N, \mu, \sigma \), and SE as before.
2. Simulate an in-control observation as \( C = \mu + \epsilon \).
3. Repeat step 2 for the required number of controls per run.
4. Repeat steps 2 and 3 for the number of runs necessary to apply the quality-control rule.
5. Apply the quality-control rule. If the rule rejects the run, start over at step 2; otherwise continue with step 6.
6. Simulate an out-of-control observation by adding the systematic shift to the control observation, \( C = \mu + \text{SE} + \epsilon \).
7. Repeat step 6 for the required number of controls per run.
8. Apply the quality-control rule and determine whether the run is rejected.
9. Repeat steps 6–8 until a rejection is encountered and store the run number, counting from the first run with SE.
10. Repeat steps 2–9 a large number of times.
11. Calculate \( P(R_i) \) as the fraction of the trials that were not rejected at step 5 but were rejected on the \( i \)th run with error. Calculate \( P(R_i | A_{i-1}) \) and \( P(\leq R_i) \) as before.

Note that because the estimate of the conditional probability of rejection, \( P(R_i | A_{i-1}) \), has as its denominator the number of trials that “survived” the first \( i - 1 \) runs, the denominator progressively decreases as the run number increases. Therefore, for estimating the conditional probability of rejection with reasonable precision, a much larger number of trials may be necessary than when using the current simulation approach. Only persistent SE is evaluated here. Simulation of a persistent increase in random error would proceed similarly.

The Figures and discussion that follow are based on two quality-control samples per run and a persistent systematic shift of 1.5\( \sigma \). I use Westgard’s notation to denote the various control rules (1). For Figures 1 and 2, I used a computer to calculate the recursive formulas given in the Appendix.

**Results**

Figure 1 shows a graph of the unconditional, conditional, and cumulative probabilities of rejection as a function of run number for the \( 1a \) and \( 2a \) rules. With two control observations per run, these rules involve only control data within a run; therefore, the probability of rejecting the \( i \)th run, given that the previous runs were accepted, \( P(R_i | A_{i-1}) \), is constant from run to run. Figure 2 (top panels) displays the error-detection characteristics for the \( 4a \) and \( 10c \) rule in the startup case. With two control observations per run, these rules must incorporate control samples from previous runs. The conditional probability of rejection is no longer constant, but initially drops and then oscillates until eventually converging to a constant value. Note that the curves for the \( 4a \) rule begin at the second run, and the curves for the \( 10c \) rule begin at the fifth run, which are the first runs with enough control observations accumulated to be testable. Figure 2 (bottom panels) shows the error-detection characteristics for the \( 4a \) and \( 10c \) rule in the ongoing case. These curves begin at the first run (with error). The error-detection capabilities increase as the fraction of the runs with error used in the rule increases. From the point at which all runs used in the rule contain the error, the pattern of behavior is similar to that of the startup case. Because there is some probability of detecting the error in the early runs (before all runs used in the rule contain error), the cumulative probability of detecting the error in the ongoing case is larger than in the startup case; in these examples, however, the difference is small.

Using the alternative simulation approach described...
in Methods, I simulated use of the Westgard multi-rule (4) for the ongoing case. To ensure that the estimate for the probability of rejecting run 15 (conditional on accepting the first 14 runs) had an adequate sample size (≥500), I simulated 600 000 trials. Figure 3 displays the results. Of 600 000 trials, 638 had not been rejected by run 15. The same oscillating pattern is seen in the estimate of $P(R_1 | A_{15})$ (the conditional probability of rejection), but because all of the rules are being applied simultaneously, the pattern reflects their combination. Figure 4 displays the cumulative probability of rejection for runs 1, 2, and 5 as a function of the magnitude of the SE. Only cumulative probabilities were being estimated, so 1000 trials were adequate to simulate each error condition.

Discussion

When evaluating quality-control rules that involve only control observations within a single run, the distinction between intermittent and persistent errors is inconsequential. The conditional probability of rejecting a run, given that previous runs have been accepted, is constant and equal to the $P_{w0}$ as originally defined and estimated by current simulation techniques. However,
when evaluating quality-control rules involving control observations from more than one run, there is no single number that can serve the role of \( p_{\text{ed}} \). Although the conditional probability of rejection is the closest conceptually to \( p_{\text{ed}} \), it is not constant from run to run. Despite this, several recent articles have evaluated the Westgard multi-rule procedure applied across runs and have reported its "probability of error detection" (8, 9).

The oscillatory behavior exhibited by the conditional probability of rejection for the rules shown in Figures 2 and 3 is real and provides a good example of one of the advantages of mathematical evaluation of \( p_{\text{ed}} \) over simulation. The curves shown in Figure 2 were constructed from exact mathematical expressions and are not subject to the random fluctuation that would be present had the curves been constructed by simulation. The purpose in the simulation of 500 000 trials to obtain the curves in Figure 3 was to assure that random fluctuations in the estimates of the conditional probabilities of rejection would be so small that they would not obscure any true oscillatory pattern.

With existing simulation approaches, \( p_{\text{ed}} \) is calculated at the \( i \)th run by simulating control observations for \( i \) runs, applying the control rule to the \( i \)th run only, repeating this process many times, and determining the fraction of rejections. The estimates obtained reflect the unconditional probability of rejecting the \( i \)th run independently of the acceptance or rejection of any prior run. Thus, a simulation trial that violates one of the control rules during the first \( i - 1 \) runs has the possibility of being counted as an "accept" at the \( i \)th run. Consequently, this estimate will produce values less than the cumulative probability of rejection. Conversely, a simulation trial that violates one of the control rules during the first \( i - 1 \) runs could be counted as a rejection at the \( i \)th run. If more trials are counted as rejections at the \( i \)th run (when earlier runs are also rejected) than when none of the earlier runs are rejected (a characteristic that should be true of any quality-control rule that is based on control observations from more than one run), then this estimate will produce values greater than the conditional probability of rejection.

Westgard and Barry (3) indicated that, for the persistent error case, a single \( p_{\text{ed}} \) is inadequate for evaluating rules that are based on control observations from more than one run; they described two alternatives for characterizing quality-control performance. First, their power function graphs were generalized to include separate curves for different run numbers. Second, they discussed the average run length (ARL, the average number of runs that will occur before a run is rejected) as a measure of error detection capability in the persistent error case, and they described a method for calculating ARL from the conditional probabilities of rejection.

If power function graphs are to be used to characterize the performance of quality-control rules that are based on control observations from more than one run, then a decision must be made as to which error-detection probability should be displayed. \( P(\leq R_0) \), the cumulative probability of rejection, seems to be most meaningful, and is shown in Figure 4. However, current simulation techniques, which estimate the unconditional probability of rejecting the \( i \)th run independently of the acceptance or rejection of the previous \( i - 1 \) runs, will produce power function graphs that give curves with lower values than those obtained by the alternative simulation technique described here and illustrated in Figure 4. Table 2 illustrates this with a comparison of values for \( p_{\text{ed}} \) obtained by using the current simulation approach and values for \( P(\leq R_0) \) obtained by using the alternative simulation approach for the \( 1_{\text{ARL}}/2_{\text{ARL}}/4_{\text{ARL}}/10_0 \) rule. Likewise, if values for the unconditional probability of rejecting the \( i \)th run independent of the acceptance or rejection of the previous \( i - 1 \) runs are substituted for the conditional probability of rejection in the formula for ARL, then estimates of the error-detection probability will tend to be too high, resulting in ARL estimates that are too low. The ARL obtained from the 500 000 trials that were simulated to produce Figure 3 is 3.274, compared with Westgard's previously reported value of 2.982 (3). Note that when the alternative simulation approach is used, the ARL, as well as the distribution of run lengths or any percentile of the distribution, can be obtained directly from the simulation results rather than requiring an indirect calculation based on the \( p_{\text{ed}} \) as a function of run number. For example, from the data in Table 1, the ARL can be calculated as \( [407(1) + 360(2) + 134(3) + 50(4) + 42(5) + 5(6)]/1000 = 1.984 \).

Published evaluations of the error-detection capabilities of quality-control rules that are based on control observations from more than one run have not distinguished between error detection at startup and that in an ongoing operation, but apparently the startup case is generally simulated. Evaluating the startup case makes the comparison of alternative multi-rule procedures difficult. For example, if existing power function graphs for the \( 1_{\text{ARL}}/2_{\text{ARL}}/4_{\text{ARL}}/10_0 \) rule are compared with the \( 1_{\text{ARL}}/2_{\text{ARL}}/4_{\text{ARL}}/10_0 \) rule with two control observations per run, the curves for runs 1 and 2 appear identical (3). This implies that the \( 10_0 \) rule contributes no error-detection

<table>
<thead>
<tr>
<th>Table 2. Error Detection Probabilities Obtained by the Current Method and an Alternative Simulation Approacha</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Run no.</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>( 0.0 )</td>
</tr>
<tr>
<td>( 0.5 )</td>
</tr>
<tr>
<td>( 1.0 )</td>
</tr>
<tr>
<td>( 1.5 )</td>
</tr>
<tr>
<td>( 2.0 )</td>
</tr>
<tr>
<td>( 3.0 )</td>
</tr>
</tbody>
</table>

* Simulations of 1000 trials evaluating the \( 1_{\text{ARL}}/2_{\text{ARL}}/4_{\text{ARL}}/10_0 \) rule with two controls per run.

a Systematic error in analytical standard deviation units.

CLINICAL CHEMISTRY, Vol. 37, No. 10, 1991 1723
capability in the first two runs with error. However, this is true only for the first two runs after startup, where the rule cannot be applied for lack of sufficient control observations. Otherwise, there is a nonzero probability that the 10k rule will be the only rule that rejects in the first or second runs with error during ongoing operation. This additional error-detection capability of the \( \frac{P(10_n)}{P(A_n)} \) rule will be reflected if the ongoing case, rather than the startup case, is simulated.

The simulation strategy described here offers numerous advantages over simulation approaches described previously for evaluating the performance characteristics of quality-control procedures to detect persistent error. It allows for the simulation of startup as well as ongoing quality-control operation, providing a more meaningful comparison of different quality-control rule combinations and direct estimation of the ARL to error detection or any other percentile of interest from the run-length distribution. This strategy provides the estimates necessary to produce power function graphs of the cumulative probability of rejection for different numbers of runs. Application of this simulation approach should provide new insight into the evaluation of performance characteristics of quality-control rules.

Appendix

Assume a quality-control rule that requires \( M \) consecutive control observations (Cs) to be \( > k \) analytic standard deviations from the quality-control mean in the same direction, when there are \( N \) control observations per run. Let \( n = M/N \), the number of runs necessary to apply the rule, and assume \( n \) is an integer.

Let \( P(A_0) = 1 \), \( P(H_0, A_0) = P(L_0, A_0) = P(O_0, A_0) = 0 \), \( P(H_i, A_j) = P(H_i) \), \( P(L_i, A_j) = P(L_i) \), \( P(O_i, A_j) = P(O_i) \)

where \( i = 1, \ldots, n-1 \); \( H_i \) = all \( N \) Cs are abnormally high in run \( i \); \( L_i \) = all \( N \) Cs are abnormally low in run \( i \); \( O_i \) = \( H_i \), \( H_i \), \( L_i \), \( L_i \); \( A_i \) = accept up through the \( i \)th run; \( R_i \) = reject the \( i \)th run; and \( P(0) \) = probability. The probability that the \( N \) Cs in the \( i \)th run are all abnormally high and the \( i \)th run is accepted is, for \( i = n, n+1, \ldots, n+1 \),

\[
P(H_i, A_j) = P(H_i, \sim H_i, A_i, \sim A_i) + \ldots + P(H_i, \ldots, H_i, \sim A_i, A_i, \sim A_i)
\]

which can be expressed as the recursive formula:

\[
P(H_i, A_j) = \sum_{j=i-n+1}^{i} [P(A_j) - P(H_j, A_j)] \prod_{k=j+1}^{i} P(H_k)
\]

Likewise,

\[
P(L_i, A_j) = \sum_{j=i-n+1}^{i} [P(A_j) - P(L_j, A_j)] \prod_{k=j+1}^{i} P(L_k)
\]

Also, \( P(O_i, A_j) = P(O_i)P(A_i, \sim 1) \) and \( P(A_i) = P(H_i, A_i) + P(L_i, A_i) + P(O_i, A_i) \).

Finally, the probability of accepting the first \( i - 1 \) runs and rejecting the \( i \)th run is

\[
P(R_i) = P(H_i, \ldots, H_{i-n+1}, \sim H_{i-n}, A_{i-n}) + P(L_i, \ldots, L_{i-n+1}, \sim L_{i-n}, A_{i-n})
\]

or

\[
P(R_i) = \left[ P(A_{i-n}) - P(H_{i-n}, A_{i-n}) \right] \prod_{j=i-n+1}^{i} P(H_j)
\]

\[
+ \left[ P(A_{i-n}) - P(L_{i-n}, A_{i-n}) \right] \prod_{j=i-n+1}^{i} P(L_j)
\]

For the "startup" case, \( P(H_0) \), \( P(L_0) \) are constant for \( i = 1, \ldots, n \), whereas for the "ongoing" case, \( P(H_i) \) and \( P(L_i) \) equal their in-control values for \( i = 1, \ldots, n \) and then shift to their values for the out-of-control condition for \( i = n+1, \ldots \). In the ongoing case, probabilities are computed to be conditional on accepting the in-control runs. The conditional probability of rejecting the \( i \)th run given that the first \( i - 1 \) runs have been accepted is, for the startup case,

\[
P(R_i | A_{i-1}) = \frac{P(R_i)}{P(A_{i-1})}, \quad i = n, n+1, \ldots
\]

and for the ongoing case,

\[
P(R_i | A_{i-1}) = \frac{P(R_i | A_{n} \cap P(A_{i-1}) | A_{\omega})}{i = n+1, n+2, \ldots
\]

References