CLIN. CHEM. 38/2, 204–210 (1992)

A Tool for the Design and Evaluation of Alternative Quality-Control Procedures

Aristides T. Hatjimihail

I have developed an interactive microcomputer simulation program for the design, comparison, and evaluation of alternative quality-control (QC) procedures. The program estimates the probabilities for rejection under different conditions of random and systematic error when these procedures are used and plots their power function graphs. It also estimates the probabilities for detection of critical errors, the defect rate, and the test yield. To allow a flexible definition of the QC procedures, it includes an interpreter. Various characteristics of the analytical process and the QC procedure can be user-defined. The program extends the concepts of the probability for error detection and of the power function to describe the results of the introduction of error between runs and within a run. The usefulness of this approach is illustrated with some examples.

Additional Keyphrases: *computer simulations* · *data handling*

Analytical quality control (QC) is achieved mainly through the repetitive analysis of stable control materials over long periods of time (*1*) and the application of statistical QC procedures to the data, so that any significant change in the distribution of error in the analytical process can be detected. Upon such detection, the analytical process is then considered to be out of control, and the analytical run is rejected.

Alternative statistics can be calculated and alternative decision rules or multirule procedures can be applied to test statistically the null hypothesis (the analytical process is in control) against the alternative (the analytical process is out of control) (*2*). When a true null hypothesis is rejected, a type I error is committed, a false rejection of an acceptably accurate analytical run. The probability of a type I error is called the "probability for false rejection" (*3*). When a false null hypothesis is accepted, a type II error is committed, the failure to detect a significant change in the distribution of error in the analytical process. The probability for rejection of a false null hypothesis is called "probability for error detection." This latter probability is also called the "power" of the QC procedure and equals 1 minus the probability for a type II error (*3*).

The definition of the probability for error detection implies that the QC procedure is applied to measurements with significant error but does not take into account the transition of an analytical process from an in-control state to an out-of-control state. For a time after that transition, a QC procedure could be applied to some measurements without significant error. In fact, if the QC procedure is applied across k runs and if $j < k$, then during the jth run after the introduction of significant error, the QC procedure will be applied to the control measurements of $k - j$ runs without significant error and to those of j runs with significant error. (Here, the term QC procedure indicates both QC rules and multirule procedures.) Therefore, the probability for error detection of that procedure during the jth run after the introduction of error will differ from the probability for error detection during the nth run after the introduction of error, if $j < k$ and $n \neq j$.

The performance of a QC procedure is described by its power functions (*4*) that relate the probability for rejection (dependent variable) with the analytical error (independent variable) of the measurements to which the QC procedure is applied. Power function graphs are graphs of the probability for rejection vs the analytical error.

Power functions are affected by the number of the controls (*4*), by the presence of a between-run component of variation, by the shape of error distribution (*5*), and by rounding (*6*). The above considerations regarding the probability for error detection obviously apply to the power functions as well.

An acceptable QC procedure should have the lowest possible probability for false rejection and a stated probability for detecting critical errors (*7*). Critical errors are considered the "maximum clinically allowable analytical errors, defined in such a way that an upper bound has been set on the (clinical) type I error" (*2*). Assuming that the distribution function of z is the unit normal distribution, one can calculate the critical errors of an analytical process from the equations:

$$\mathrm{RE_c} = (\mathrm{TE_a} - \mathrm{bias})/[z_1(\mathrm{SD})]$$

$$\mathrm{SE_c} = [(\mathrm{TE_a} - \mathrm{bias})/\mathrm{SD}] - z_2$$

$$P(z > z_1) = P(\mathrm{TE} > \mathrm{TE_a})$$

$$P(z > z_2) = P(\mathrm{TE} > \mathrm{TE_a})/2$$

where $\mathrm{RE_c}$ and $\mathrm{SE_c}$ are respectively the critical random and systematic errors, $\mathrm{TE_a}$ is the total allowable analytical error (*8*), TE is the total analytical error, SD and bias are the standard deviation and systematic error of the process when in control, and z_1 and z_2 are real numbers. To evaluate QC procedures applied to analytical processes having a frequency f of critical errors, one can use the defect rate and test yield (*9*), defined as:

$$\text{Defect rate} = f(1 - P_{\mathrm{ed}})$$

$$\text{Test yield} = (1 - f)(1 - P_{\mathrm{fr}})$$

The Microbiology Laboratory, Health Center of Prosotsane, GR-66200 Prosotsane, Greece. Address for correspondence: P.O. Box 56, GR-66100, Drama, Greece.

Received December 6, 1990; accepted October 24, 1991.

where P_{ed} is the probability for critical error detection and P_{fr} is the probability for false rejection.

QC has been introduced into industry by Shewhart (*10*) and into clinical laboratories by Levey and Jennings (*1*), who proposed rules based on the mean and range of two control measurements. Henry and Segalove (*11*) proposed simpler rules based on a single measurement. Later Westgard et al. (*12*) introduced a multirule procedure based on six decision rules, published as a Proposed Selected Method in *Clinical Chemistry*. All those QC rules and procedures used control charts. Their application was simple, "such that computerized data handling was not necessary" (*12*).

As Westgard and Klee (*7*) have pointed out, "the widespread availability of microprocessors and microcomputers makes it practical to use complicated control procedures on a routine basis in clinical laboratories."

The evaluation of alternative QC procedures can be facilitated by estimating and plotting their power function graphs by computer simulation (*6*), given that the algebraic definition of their power functions is often very complex or practically impossible. Westgard and Groth (*6*) developed an interactive computer program in Fortran IV for a mainframe computer (IBM 370/158) to simulate QC rules and multirule procedures under user-specified conditions. Microcomputer programs for similar simulations have been developed in BASIC for a Commodore 8032 microcomputer by Blum (*13*) and for an IBM or IBM-compatible microcomputer by Smith and Cossitt (*14*). The mainframe and both microcomputer programs estimate probabilities for false rejection and error detection under different conditions of systematic and random error. The mainframe program permits the user to define various analytical conditions and to define a QC procedure by using one or more of 34 QC rules. It can also propose a QC procedure that meets the user-defined specifications. The program by Blum simulates seven QC procedures based on 10 rules with user-defined limits. The program by Smith and Cossitt allows use of as many as 10 rules, chosen in any combination from 20 generic rules.

To assist the design and evaluation of alternative QC procedures, I have developed a flexible microcomputer simulation program that includes an interpreter.[1] The decision rules are derived from 10 generic rules. The interpreter permits the user to define the QC multirule procedures by using Boolean algebra syntax. The number of controls per concentration, the number of concentrations of the controls, the ratio of the between-run standard deviation to the within-run standard deviation, and the number of significant digits in the measurements are also user-defined. The program estimates the probabilities for false rejection and for error detection for various values of random and systematic error. To describe more realistically the performance of a QC procedure, the program assumes that (*a*) there is error in all of the measurements to which the QC procedure is applied, (*b*) the error has been introduced between two consecutive runs, and (*c*) the error has been introduced within a run. The respective power function graphs are plotted. The program also tests whether the QC procedure meets the user-specified requirements for total allowable analytical error detection and estimates the defect rate and test yield of the analytical process.

Materials and Methods

The program is written in Turbo Pascal, version 5.5 (*15,16*), for an IBM or IBM-compatible microcomputer, with an 80×86 central processing unit, 640 kilobytes of RAM, and graphics capability, under MS-DOS, version 3.30. It includes the units described in the following sections.

The Rules Unit

The user defines the decision rules by defining as many as eight parameters of one or more generic rules. The rules (see *Appendix I*) can be divided into six classes: single-value rules (*12, 17*), sum rules (*7, 18*), range rules (*7, 10*), mean rules (*7, 10*), standard deviation rules (*7*), and trend rules (*6, 19*).

The rules can be applied across runs, within a run, or across concentrations or within each concentration of the controls (*12*).

The Interpreter Unit

The QC multirule procedures are defined as Boolean expressions composed of operands (symbols of the user-defined rules or multirule procedures), operators [symbols of the logical operations AND, OR, NOT, XOR (exclusive or)], and parentheses (as required).

The Westgard–Barry–Hunt–Groth procedure (*12*), hereafter referred to as "Westgard," could be defined, for example, as S12 AND [S13 OR G22 OR L22 OR (G12 AND L12) OR G41 OR L41 OR G00 OR L00], where S12, S13, G22, L22, G12, L12, G41, L41, G00, and L00 are user-defined single-value rules (see *Appendix II*).

The interpreter accepts those expressions as input.

The Random Normal Deviates Generator Unit

The program assumes a gaussian (normal) distribution of error. The unit generates Box–Muller series of random normal deviates from pseudorandom numbers (*20*). Two series are generated for each concentration of the controls, to simulate the within-run and between-run standard deviations. The series are standardized to a mean of 0 and a variance of 1. Their coefficients of skewness and kurtosis are calculated. The series are accepted if the coefficients are less than their standard deviations (*21*).

During the simulation of a QC procedure, the same two series per concentration are used to simulate the control measurements under different conditions of random and systematic error. The series are filed and can be used for multiple simulations of alternative QC procedures, as well.

[1] The program is available from the author for $8 (the estimated cost of reproduction, postage, and handling). Requests for the program should be addressed to A. T. Hatjimihail, M.D., P.O. Box 56, GR-66100 Drama, Greece.

The Simulator Unit

The user defines the number of controls per concentration and the number of concentrations (up to three) of the controls. The program simulates 6000 control measurements at each concentration, in consecutive runs. If n is the number of control measurements at each concentration of a run, then the number of the simulated runs is equal to $6000/n$.

The program introduces into the simulated control measurements total random error, from 1 to 6 SD, and systematic error, from 0 to 5 SD, in increments of 0.25 SD. The within-run and the between-run standard deviations are calculated from the equations

$$(s_t)^2 = (s_{wr})^2 + (s_{br})^2 \text{ and } (s_{br})/(s_{wr}) = r,$$

where s_t is the total standard deviation, s_{wr} is the within-run standard deviation, s_{br} is the between-run standard deviation, and r is a user-defined ratio. When the number of the significant digits is specified, the measurements are rounded appropriately.

If a QC procedure is to be applied to c measurements, then during the simulation it is applied to each consecutive simulated control measurement at each concentration of the controls and to the $(c-1)$ previous control measurements, either across or within runs or concentrations of the controls. Depending upon the defined decision rules, the appropriate statistics are calculated and compared with the user-defined limits. If the QC procedure is true, we have a rejection. Then the QC procedure is applied to the first control measurement of the first concentration of controls of the next run and to the previous $(c-1)$ control measurements. When the QC procedure has been applied to all the simulated runs, the number of rejections is counted and the probability for rejection is estimated as the ratio of the number of the runs rejected to the number of the simulated runs.

The program simulates the following situations:

(*a*) There is error in all of the measurements to which the QC procedure is applied.

(*b*) The error is introduced between two consecutive runs. Assume that a rejection occurs when the QC procedure is applied to control measurements of the $(j-k+1)$th to jth run, where $k \geq 1$, then a series of measurements is simulated in which there is no error in runs $(j-k+2)$ to j. The error is introduced between the jth and the $(j+1)$th runs. Therefore, if the QC procedure is applied across runs, it will be applied to at least one measurement without significant error. The error persists until a rejection occurs. If the rejection occurs during the mth run $(m>j)$, then a new series of measurements is simulated where there is no error in the runs $(m-k+2)$ to m. The estimated probability for rejection is the mean probability for error detection after the introduction of the error.

(*c*) The error is introduced within a run after taking the ith control measurement, where $i \geq 1$. The approach is similar to that described in *b*. Assuming that rejection occurs when the QC procedure is applied to control measurements of the $(j-k+1)$th to jth run (where $k \geq 1$), then a series of measurements is simulated where there is no error in the runs $(j-k+2)$ to j and in the first i measurements of the $(j+1)$th run. The error is introduced between the ith and the $(i+1)$th measurement of the $(j+1)$th run. Therefore, if the QC procedure is applied to more than one control measurement per run, the procedure will be applied to at least one measurement without significant error.

The program can estimate the probability for error detection for any user-defined combination of random and systematic error.

Furthermore, the user can define the standard deviation and bias of an analytical process when in control; the allowable analytical error as well as the desired probability that this is greater than the total error; the desired probability for critical error detection; and the frequency of critical errors. Then, the program tests whether the QC procedure meets the requirements for total allowable analytical error detection and estimates the defect rate and the test yield.

The Graphics Unit

The graphics unit plots in high resolution the power function graphs of a QC rule or multirule procedure for random and systematic error. When two QC rules or procedures are compared, the unit plots the graphs of the differences of their probabilities for rejection.

The Statistics Unit

This unit includes the functions of the unit gaussian (normal) and chi-square distributions (*22*).

The Utilities Unit

This part includes utilities for filing the user-defined rules and procedures and the results of the simulations, so that the results can be used for multiple comparisons without repeating the simulation.

Results

As an example, I have simulated two QC procedures: the Westgard QC procedure (see *Appendix II*) and an alternative QC procedure, defined as S12 AND (M01 OR D42).

Let m and SD be the mean and the standard deviation respectively of the control measurements when the analytical process is in control. The definitions of the decision rules of the alternative procedure are

S12: The last one control measurement is greater than m plus 2SD or less than m minus 2SD.

M01: The mean of the last 10 control measurements is greater than m plus 1SD or less than m minus 1SD.

D42: The standard deviation of the last four control measurements is greater than 2SD.

The rules are derived from the generic rules 1, 4, and 6, respectively (see *Appendix I*), and they are applied across runs and concentrations of controls.

The procedures are tested with one control per concentration, two concentrations of controls, no rounding, and between-run error set to 0. The probability for false rejection of both procedures is estimated to be 0.010.

The power function graphs in Figures 1–4 describe the performance of the two QC procedures under different conditions of introduction of random and systematic error. The conditions assumed are as follows: (*a*) there is error in all measurements, (*b*) the error is introduced between two consecutive runs, and (*c*) the error is introduced within a run. When the error is introduced within a run, it is assumed to be introduced after the first control measurement but before the other measurements of the run (bracketed or pre-control mode). In the subsequent figures, the two procedures are compared under the assumption that random and systematic errors exist in all of the measurements (Figures 5 and 6) or that they are introduced between two consecutive runs (Figures 7 and 8).

Discussion

The usefulness of computer simulations for the design and evaluation of alternative QC procedures is well established (*3–6, 12–14*). The programs developed by Westgard et al. (*12*) and by Smith and Cossitt (*14*) use a menu-driven definition of the QC procedures. I use an interpreter, which enhances the flexibility of this process.

The QC procedures are evaluated by estimating the probabilities for false rejection and for error detection and then plotting the power function graphs.

So far, the estimates of the probabilities for error detection and of the power functions have been based on the assumption that error exists in all of the measurements to which the QC procedure is applied. The result of this assumption can be an optimistic estimation of the statistical power of the QC procedures that are applied across runs. Assuming that the error is introduced between two consecutive runs or within a run, one would estimate lower probabilities for error detection and different power functions (Figures 1–4). Consequently, lower probabilities for critical error detection and higher defect rates would also be estimated.

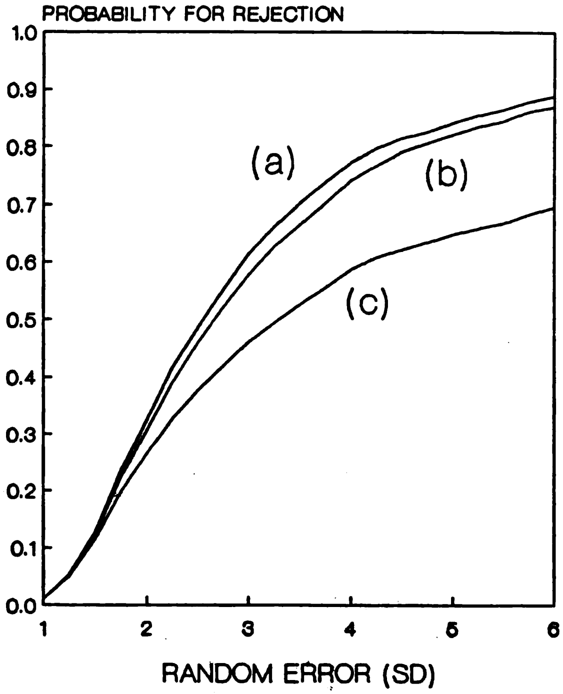

Fig. 1. Power function graphs of the Westgard QC procedure, assuming that (*a*) there is random error in all of the measurements, (*b*) random error is introduced between runs, and (*c*) random error is introduced within a run

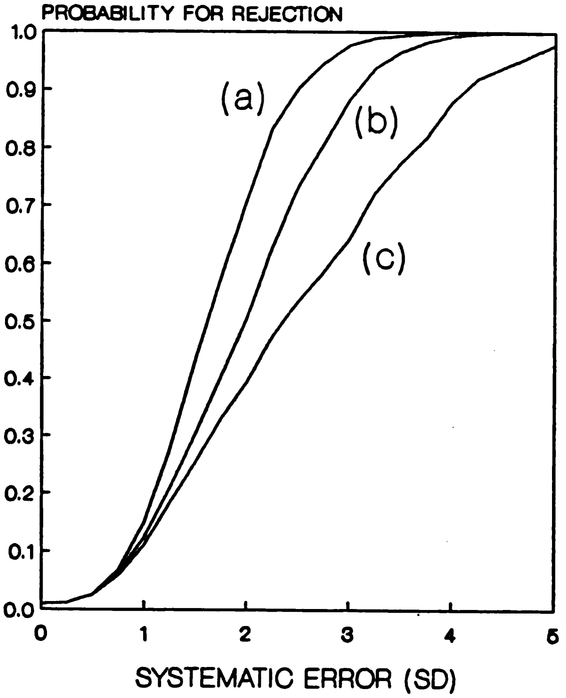

Fig. 2. Power function graphs of the Westgard QC procedure, assuming that (*a*) there is systematic error in all of the measurements, (*b*) systematic error is introduced between runs, and (*c*) systematic error is introduced within a run

Concerning QC procedures applied across runs, the power functions proposed by Westgard and Groth (*4*) describe situations where the error has been introduced into the analytical process before the application of the QC procedure. I propose here power functions describing situations in which error is introduced during the application of the QC procedure, in the period between two consecutive runs or within a run. This approach can be useful, depending on the way the error is introduced into a particular analytical system, as shown in the power function graphs presented in Figures 5–8. The power function graphs proposed by Westgard and Groth (Figures 5 and 6) give the impression that the alternative QC procedure described above has a better performance than does the Westgard procedure. On the other hand, the power function graphs I propose (Figures 7 and 8) show that, when the error is introduced between runs, the performance of the Westgard procedure is better than that of the alternative procedure. Therefore, the extended concepts of the probability for error detection and of the power function offer a better description of the performance of QC procedures applied across runs.

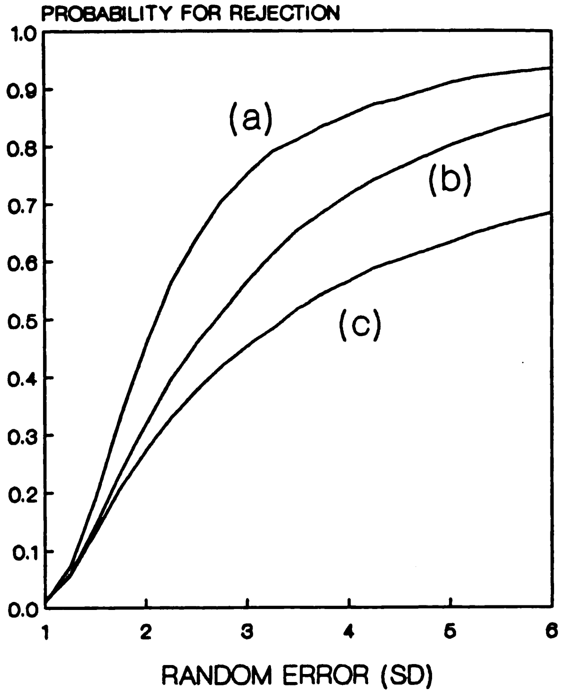

Fig. 3. Power function graphs of the alternative QC procedure
Assumptions as in Fig. 1

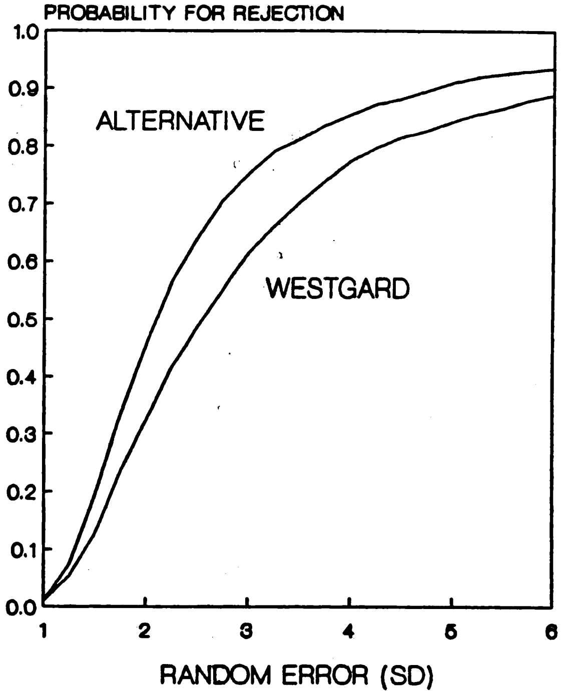

Fig. 5. Power function graphs of the Westgard and alternative QC procedures, assuming random error in all of the measurements

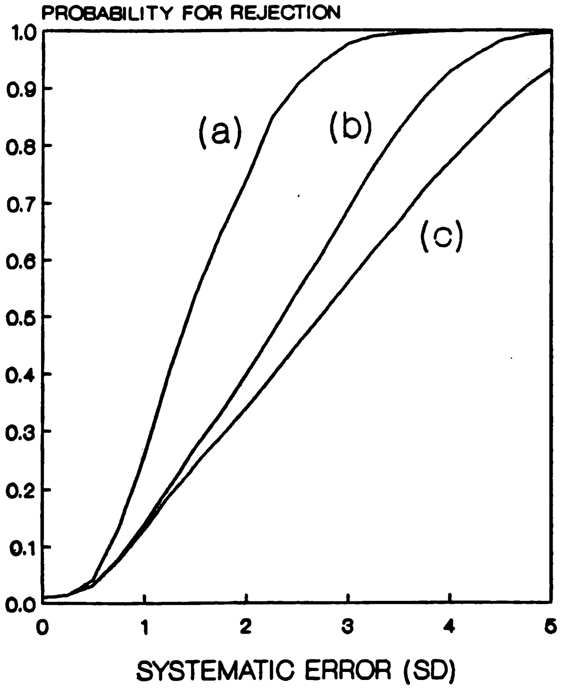

Fig. 4. Power function graphs of the alternative QC procedure
Assumptions as in Fig. 2

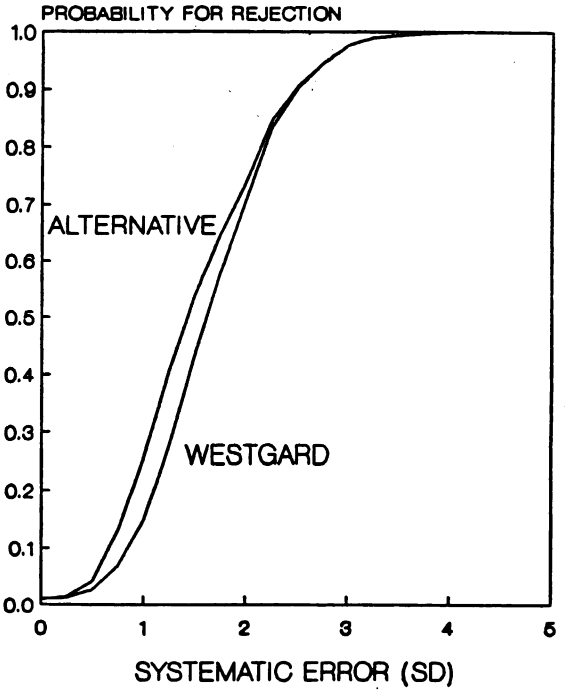

Fig. 6. Power function graphs of the Westgard and alternative QC procedures, assuming systematic error in all of the measurements

In conclusion, the microcomputer simulation program described here allows a flexible design and a more realistic evaluation of alternative QC procedures.

Appendix I

Let SD and m be the standard deviation and mean of the control measurements when the analytical process is in control, and let *c*, *n*, min, max, min_c, max_c, *P*, and *a* be user-defined parameters. The generic rules are

1. The values/absolute values of *c* out of the last *n* control measurements are less/greater than m plus (max)(SD) and/or less/greater than m minus (min)(SD).

2. The value/absolute value of the sum of the differences (control measurement − m) of the last *n* controls is less/greater than (max)(SD) and/or less/greater than (min)(SD).

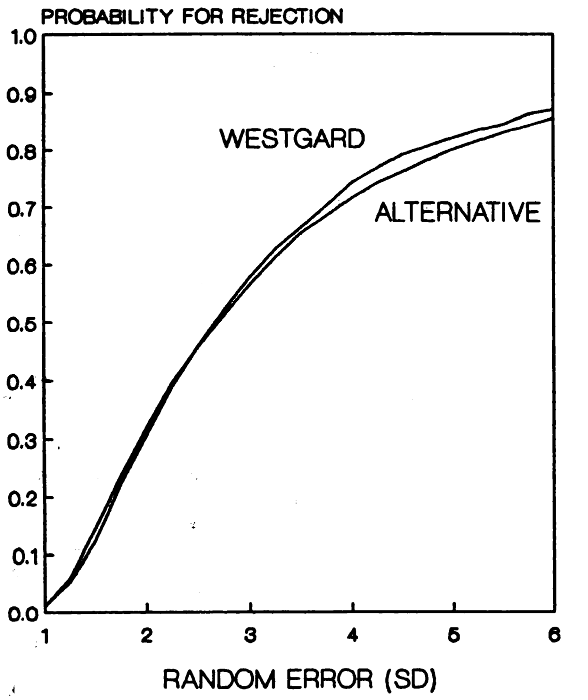

Fig. 7. Power function graphs of the Westgard and alternative QC procedures, assuming random error is introduced between runs

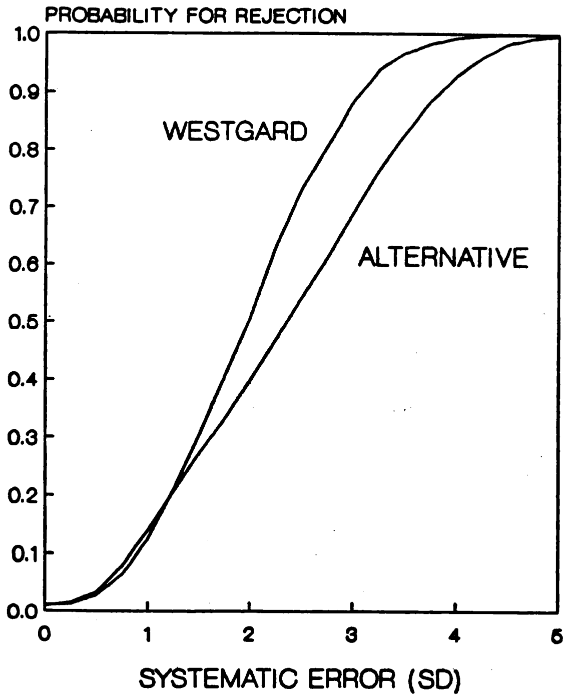

Fig. 8. Power function graphs of the Westgard and alternative QC procedures, assuming systematic error is introduced between runs

The sum can be an algebraic sum, a sum of absolute values, or a cumulative sum. The calculation of the cumulative sum is initiated when a control measurement is less/greater than m plus (max$_c$)(SD) and/or less/greater than m minus (min$_c$)(SD).

3. The range of the last *n* control measurements is less/greater than (max)(SD) and/or less/greater than (min)(SD).

4. The value/absolute value of the difference between the mean of the last *n* control measurements and m is less/greater than (max)(SD) and/or less/greater than (min)(SD).

5. The mean of the last *n* control measurements is different from m at the *P* level of significance.

6. The standard deviation of the last *n* control measurements is less/greater than (max)(SD) and/or less/greater than (min)(SD).

7. The standard deviation of the last *n* control measurements is greater than SD at the *P* level of significance.

8. The smoothed mean of the last *n* control measurements is different from m at the *P* level of significance, when the smoothing constant equals *a*.

9. The smoothed standard deviation of the last *n* control measurements is greater than SD at the *P* level of significance, when the smoothing constant equals *a*.

10. The value/absolute value of the tracking signal of the last *n* control measurements is less/greater than (max)(SD) and/or less/greater than (min)(SD), when the smoothing constant equals *a*.

Appendix II

Let m and SD be the mean and the standard deviation of the control measurements when the analytical process is in control. The definitions of the decision rules of the Westgard procedure are

S12: The last one control measurement is greater than m plus 2SD or less than m minus 2SD.

S13: The last one control measurement is greater than m plus 3SD or less than m minus 3SD.

G22: The last two control measurements are greater than m plus 2SD.

L22: The last two control measurements are less than m minus 2SD.

G12: The one out of the last two control measurements is greater than m plus 2SD.

L12: The one out of the last two control measurements is less than m minus 2SD.

G41: The last four control measurements are greater than m plus 1SD.

L41: The last four control measurements are less than m minus 1SD.

G00: The last 10 control measurements are greater than m plus 0(SD).

L00: The last 10 control measurements are less than m minus 0(SD).

These rules are derived from the generic rule 1 (see *Appendix I*). They are applied across runs and concentrations of controls except for rules G12 and L12, which are applied within-run and across the concentrations of the controls.

The Westgard procedure is defined as S12 AND [S13 OR G22 OR L22 OR (G12 AND L12) OR G41 OR L41 OR G00 OR L00].

I thank Nikos Sotiriadis, systems analyst, for his continuing support and encouragement. I also thank Tima Vlastos for her valuable comments during the preparation of this manuscript.

References

1. Levey S, Jennings ER. The use of control charts in the clinical laboratories. Am J Clin Pathol 1950;20:1059–66.

2. Linnet K. Choosing quality-control systems to detect maximum clinically allowable analytical errors. Clin Chem 1989;35:284–8.

3. Westgard JO, Groth T, Aronsson T, Falk H, de Verdier CH. Performance characteristics of rules for internal quality controls: probabilities for false rejection and error detection. Clin Chem 1977;23:1857–67.

4. Westgard JO, Groth T. Power functions for statistical control rules. Clin Chem 1979;25:863–9.

5. Westgard JO, Falk H, Groth T. Influence of a between-run component of variation, choice of control limits, and shape of error distribution on the performance characteristics of rules for internal quality control. Clin Chem 1979;25:394–400.

6. Westgard JO, Groth T. Design and evaluation of statistical control procedures: applications of a computer "quality control simulator" program. Clin Chem 1981;27:1536–45.

7. Westgard JO, Klee GG. Quality assurance. In: Tietz NW, ed. Textbook of clinical chemistry. Philadelphia: WB Saunders, 1986:424–58.

8. Koch DD, Oryall JJ, Quam EF, et al. Selection of medically useful quality-control procedures for individual tests done in a multitest analytical system. Clin Chem 1990;36:230–3.

9. Westgard JO, Burnett RW. Precision requirements for cost-effective operation of analytical processes. Clin Chem 1990; 36:1629–32.

10. Shewhart WA. Economic control of quality of the manufactured product. New York: Van Nostrand, 1931.

11. Henry RJ, Segalove M. The running of standards in clinical chemistry and the use of the control chart. J Clin Pathol 1952;5: 305–11.

12. Westgard JO, Barry PL, Hunt MR, Groth T. A multi-rule Shewhart chart for quality control in clinical chemistry. Clin Chem 1981;27:493–501.

13. Blum AS. Computer evaluation of statistical procedures, and a new quality-control statistical procedure. Clin Chem 1985;31:206–12.

14. Smith FA, Cossitt NL. Simulated comparison of multi-rule protocols for statistical quality control [Abstract]. Clin Chem 1987;33:909.

15. Turbo Pascal reference guide, version 5.0. Scotts Valley, CA: Borland International, 1989.

16. Turbo Pascal object oriented programming guide, version 5.5. Scotts Valley, CA: Borland International, 1989.

17. Ehrmeyer SS, Laessig RH, Schell K. Use of alternative rules (other than the 1_{2s}) for evaluating interlaboratory performance data. Clin Chem 1988;34:250–6.

18. Westgard JO, Groth T, Aronsson T, de Verdier CH. Combined Shewhart–cusum control chart for improved quality control in clinical chemistry. Clin Chem 1977;23:1881–7.

19. Cembrowski GS, Westgard JO, Eggert AA, Toren EC Jr. Trend detection in control data: optimization and interpretation of Trigg's technique for trend analysis. Clin Chem 1975;21:1396–405.

20. Box GEP, Muller ME. A note on the generation of random normal deviates. Ann Math Stat 1958;29:610–1.

21. Solberg HE. Establishment and use of reference values. *Op. cit.* (ref. 7):356–86.

22. Floegel E. Statistik in Basik. F.R.G.: Ing W Hofacker GmbH, 1984.