[Reprinted from CLINICAL CHEMISTRY, 36, 230 (1990).]


# Selection of Medically Useful Quality-Control Procedures for Individual Tests Done in a Multitest Analytical System

**David D. Koch,[1-3] Jeffrey J. Oryall,[2,3] Elsa F. Quam,[3] Donald H. Feldbruegge,[3] Dennis E. Dowd,[3] Patricia L. Barry,[3] and James O. Westgard[1-3]**

Quality-control (QC) procedures (i.e., decision rules used, numbers of control measurements collected per run) have been selected for individual tests of a multitest analyzer, to see that clinical or "medical usefulness" requirements for quality are met. The approach for designing appropriate QC procedures includes the following steps: (*a*) defining requirements for quality in the form of the "total allowable analytical error" for each test, (*b*) determining the imprecision of each measurement procedure, (*c*) calculating the medically important systematic and random errors for each test, and (*d*) assessing the probabilities for error detection and false rejection for candidate control procedures. In applying this approach to the Hitachi 737 analyzer, a design objective of 90% (or greater) detection of systematic errors was met for most tests (sodium, potassium, glucose, urea nitrogen, creatinine, phosphorus, uric acid, cholesterol, total protein, total bilirubin, γ-glutamyltransferase, alkaline phosphatase, aspartate aminotransferase, lactate dehydrogenase) by use of 3.5s control limits with two control measurements per run (N). For the remaining tests (albumin, chloride, total $CO_2$, calcium), requirements for QC procedures were more stringent, and 2.5s limits (with N = 2) were selected.

The selection or design of quality-control (QC) procedures to detect "maximum clinically allowable analytical errors" was discussed recently by Linnet (*1*).[4] Using published values of "medically important changes" and analytical imprecision, Linnet illustrated that careful selection of QC procedures is necessary to satisfy clinical requirements for quality. Linnet also recommended critical evaluation of recent literature suggestions for the implementation of "clinically useful control limits."

We have applied an approach similar to Linnet's for designing cost-effective QC procedures (*2*) to assess the correctness of the QC procedures we used with a multitest analyzer. After this analyzer was introduced into our laboratory, we became concerned whether previous QC practices were still appropriate. As is the habit in most laboratories, we carried over the QC procedures used on the previous instrument system to the new analytical system, even though the two systems were quite different in design and performance. Analysts observed that the new analytical system was both more precise and more stable and questioned whether the previous statistical control criteria were still appropriate for maintaining medically useful performance while minimizing costs.

The QC design approach has been applied to each individual test in the multitest system. The objectives were to select control rules (decision criteria) and numbers of control measurements (N) that would simultaneously minimize false rejection of runs and maximize detection of runs having medically important errors. "Cost" would be optimized by reducing the number of runs that are falsely rejected; "effectiveness" would be optimized by increasing the detection of runs having medically important errors.

For optimal performance, a QC procedure should have a low probability for false rejection of analytical runs ($P_{fr}$) and a high probability for error detection ($P_{ed}$). $P_{fr}$ and $P_{ed}$ will depend on the control rules and number of control measurements (N) in the analytical run. Both $P_{fr}$ and $P_{ed}$ generally increase as control limits are narrowed and the number of measurements is increased. $P_{ed}$ will also depend on the size of medically important errors (*2*), which is related to the clinical requirement for quality and the analytical characteristics of the method being controlled. These key characteristics are *precision*, measured by "s," the stable standard deviation of each method calculated for this study from the determination of control products over a 15-month period, and *accuracy* (bias) of the measurement procedure. Errors that are large multiples of the stable standard deviation will be easier to detect than those that are small multiples. Measurement procedures having high precision (low s) will be easier to monitor than those having low precision, as pointed out previously by Arkin (*3*), who stated that "improved precision may lessen the requirements for the numbers of control specimens per run," and more recently by Campbell (*4*), who commented that "the limiting factor [in improving QC systems] is inadequate analytical precision."

## Materials and Methods

*Analytical system:* The Hitachi 737 analyzer (Boehringer Mannheim Diagnostics, Indianapolis, IN 46250) was in routine operation at the University of Wisconsin Hospital and Clinics, 18 of the 23 available tests being used. All reagents were also from Boehringer Mannheim Diagnostics, with the AST procedure modified as described elsewhere (*5*). The QC procedure initially in use for all tests was a $1_{3s}/2_{2s}/R_{3.64s}$ multirule procedure having two control samples bracketing 18 patients' samples. The control measurements per run were from three different control materials used in an alternating sequence throughout daily operation.

*Project team:* This study originated as a quality-improvement project in the laboratory's quality-assurance program. A project team was established to bring together the expertise and support necessary to evaluate the QC prob-

[1] Department of Pathology and Laboratory Medicine, Medical School; [2] Medical Technology Program, School of Allied Health Professions; and [3] Clinical Laboratories, University of Wisconsin, 600 Highland Avenue, Madison, WI 53792.

[4] Nonstandard abbreviations: QC, quality control; $P_{fr}$, probability for false rejection; $P_{ed}$, probability for error detection; N, number of control measurements per analytical run; s, analytical standard deviation; $TE_a$, allowable total error specification; $SE_c$, critical systematic error; $RE_c$, critical random error; GGT, γ-glutamyltransferase (EC 2.3.2.2); ALP, alkaline phosphatase (EC 3.1.3.1); AST, aspartate aminotransferase (EC 2.6.1.1); and LD, lactate dehydrogenase (EC 1.1.1.27).

Received September 15, 1989; accepted November 16, 1989.

lems and design new QC procedures. The team was the mechanism for defining medical usefulness goals for analytical quality, collecting information and performance data about the analytical system, reviewing performance characteristics of control procedures, and developing recommendations for new QC designs.

*QC design approach:*

- Clinical requirements for quality were defined for each test in terms of "total error specifications" ($TE_a$), which represented the performance required for medical usefulness.
- Estimates of the stable standard deviation (s) of each test done with the analytical system were calculated from measurement of control products over a 15-month period.
- The sizes of critical systematic ($\Delta SE_c$) and critical random ($\Delta RE_c$) analytical errors that need to be detected to maintain the clinical requirements for quality were calculated from the following equations (*2*):

$$\Delta SE_c = [(TE_a - |bias|)/s] - 1.65$$

$$\Delta RE_c = (TE_a - |bias|)/1.96s$$

where $TE_a$ is the total error requirement for the method in question, s is the stable method standard deviation, and bias is the observed inaccuracy vs reference or comparative methods. The bias term was set to zero in all calculations in this application, because prior method-evaluation studies (*5*) had indicated that the systematic differences between comparative systems in our laboratory were small, and we designed ongoing accuracy efforts to maintain near-zero biases. By calculating the critical errors $\Delta SE_c$ and $\Delta RE_c$ as multiples of s, the performance capabilities of potential QC rules can be tested by using power-function graphs.

- The performance characteristics of candidate QC procedures were determined by a computer simulation program (*6*). Power-function graphs were prepared (*7*) and the probabilities for false rejection and detection of the critical-size errors were estimated by graphical interpolation.
- QC procedures were selected with the objectives of having 90% detection of *systematic error,* while maintaining false rejections as low as possible. A high detection rate for random error was considered a secondary objective because of the instrument's high precision observed over 20 months of use.

*QC computer simulation program:* The program by Groth et al. (*6*) has been implemented on an in-house laboratory computer (PDP 11/84; Digital Equipment Corp., Marlboro, MA). Power-function graphs (*7*) were generated for a variety of candidate control rules, including $1_{2.5s}$, $1_{3.0s}$, $1_{3.5s}$, $1_{4.0s}$, $1_{5.0s}$, and $1_{6.0s}$ rules having one to four control measurements per run. Estimates of probabilities for rejection were obtained from 1000 simulated runs.

## Results

*Medically important errors:* Table 1 summarizes the information procured in the first three steps of the design approach. Clinical requirements for quality were defined in the form of total-error goals ($TE_a$) through a consensus process. Project team members reviewed literature recommendations for medical usefulness, the fixed limits suggested by the College of American Pathologists proficiency programs, and the internal limits for agreement of measurements between analytical systems in routine operation. Estimates of s for each analytical test on the Hitachi 737 were obtained from measurements of QC materials during 15 months. A representative control material was selected from the three different materials in use, based on the observed stability of that material and the agreement of its mean with an appropriate medical decision value. The results of calculations of critical systematic and random errors are presented as multiples of the standard deviation of the measurement procedure; i.e., the value of 4.32 for $\Delta SE_c$ for sodium corresponds to a systematic error equivalent to 4.32 times the standard deviation; the value of 3.05 for $\Delta RE_c$ corresponds to a tripling of the standard deviation.

**Table 1. Clinical Requirements for Quality, Analytical Imprecision, and Medically Important Systematic and Random Errors for Individual Tests on the Hitachi 737 Analyzer**

| Test | Units | Total error, $TE_a$ | Observed precision, s | Critical systematic error, $\Delta SE_c$ | Critical random error, $\Delta RE_c$ |
|---|---|---|---|---|---|
| Sodium | mmol/L | 4.0 | 0.67 | 4.32[a] | 3.05[a] |
| Potassium | mmol/L | 0.3 | 0.035 | 6.92 | 4.37 |
| Chloride | mmol/L | 4.0 | 1.04 | 2.20 | 1.96 |
| Total $CO_2$ | mmol/L | 3.0 | 0.75 | 2.35 | 2.04 |
| Glucose | mg/L | 80 | 12.0 | 5.02 | 3.40 |
| Urea nitrogen | mg/L | 30 | 4.0 | 5.85 | 3.38 |
| Creatinine | mg/L | 3.0 | 0.30 | 8.35 | 5.10 |
| Calcium (1) | mg/L | 5.0 | 1.58 | 1.52 | 1.62 |
| Calcium (2) | mg/L | 5.0 | 1.68 | 1.33 | 1.52 |
| Phosphorus | mg/L | 4.0 | 0.51 | 6.19 | 4.00 |
| Uric acid | mg/L | 8.0 | 0.88 | 7.44 | 4.64 |
| Cholesterol | mg/L | 200 | 27.0 | 5.76 | 3.78 |
| Total protein | g/L | 6.0 | 0.92 | 4.87 | 3.33 |
| Albumin | g/L | 3.0 | 0.64 | 3.04 | 2.39 |
| Total bilirubin | mg/L | 4.0 | 0.44 | 7.44 | 4.64 |
| GGT | U/L | 12.0 | 1.40 | 6.92 | 4.37 |
| ALP | U/L | 12.0 | 1.40 | 6.92 | 4.37 |
| AST | U/L | 10.0 | 1.50 | 5.02 | 3.40 |
| LD | U/L | 40.0 | 6.00 | 5.02 | 3.40 |

[a] $\Delta SE_c$ and $\Delta RE_c$ are expressed in multiples of s; $\Delta SE_c$ of 2.00 corresponds to a 2s shift; $\Delta RE_c$ of 2.00 corresponds to a doubling of the standard deviation.

*Performance capabilities of candidate control procedures:* Figure 1 shows the performance characteristics for several control rules, all with N = 2, for detecting systematic errors. We began with N = 2 because we wanted to continue using a "bracketing" mode of operation, with a control measurement preceding a group of patients' samples and another control measurement following the group. A decision to report the patients' results would be based on both control values; hence, N = 2. Probabilities for rejection are plotted vs the size of systematic error. The medically important systematic errors for several tests ($\Delta SE_c$ in Table 1) are indicated on the *x*-axis of Figure 1 to illustrate how easily the control rule requirements for each test are educed. The objective of 90% error detection can be readily achieved with the $1_{3.5s}$ rule for tests such as sodium, AST, and others for which critical error values exceed 4.0s. Using the $1_{2.5s}$ rule, one can achieve nearly 90% error detection for albumin, but only about 60% error detection for chloride and total $CO_2$, and only 25% for calcium.

Figure 2 shows the performance of these same control rules (with N = 2) for detection of random error. Clearly,

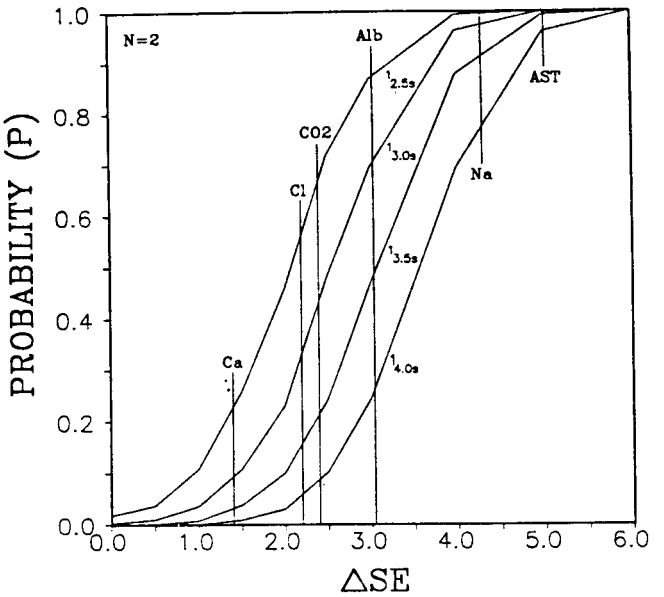


Fig. 1. Power functions illustrating the probabilities for rejecting runs with increasing *systematic* errors when different control rules with two control measurements per run (N = 2) are used

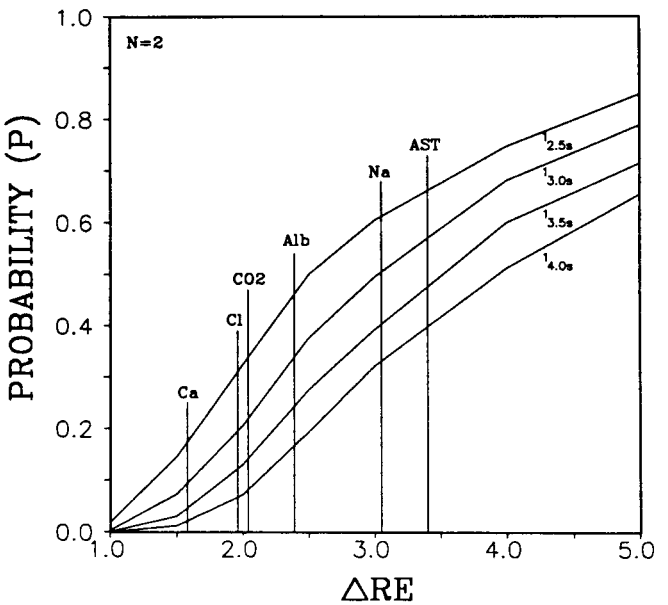


Fig. 2. Power functions illustrating the probabilities of rejecting runs with increases in *random* error when different control rules having two control measurements per run (N = 2) are used

medically important random errors are more difficult to detect than are medically important systematic errors.

*Recommendations and implementation of medically useful QC procedures for Hitachi 737:* QC procedures using a bracketing mode, and therefore N = 2, were chosen. The QC rules are applied to both control measurements independent of each other. If either control value exceeds the control rule limits for that method, the run is rejected. A QC rule having 3.5s control limits achieves the design objectives for sodium, potassium, urea nitrogen, creatinine, phosphorus, uric acid, cholesterol, total protein, total bilirubin, GGT, ALP, AST, and LD. A QC rule having 2.5s limits achieves the design objectives for albumin and provides improved performance of chloride, total $CO_2$, and calcium. For chloride and $CO_2$, use of a multirule procedure could further improve performance. The choice of rules for calcium is problematic because precision for this analyte during stable operation is barely adequate.

## Discussion

The cost-effectiveness of an automated multitest analytical system, such as the Hitachi 737, can be optimized by selecting proper QC procedures for each individual test. Clearly, different QC procedures are appropriate for different tests done in this analyzer—and probably in most analyzers. Careful design is necessary and requires the use of a systematic approach that begins with a definition of the clinical requirements for quality (total error requirements), takes the performance (imprecision) of the measurement procedure into account, and considers the performance capabilities of the control procedure itself (probabilities for error detection and false rejection).

In using this design approach, once the sizes of the critical systematic and random errors are known the task of selecting a QC procedure becomes clear. Large critical errors suggest that a QC procedure with low N and wide limits will be appropriate. Small critical errors indicate the need for higher N, narrower control limits, and possibly multirule procedures.

It is important to recognize that the sizes of the critical errors depend on the analytical performance of the measurement system. Both $\Delta SE_c$ and $\Delta RE_c$ are a function of the ratio of the clinical requirement and the standard deviation of the measurement procedure. A given clinical requirement will result in small critical errors when the measurement procedure is imprecise and large critical errors when the measurement procedure is very precise. Different QC procedures will be appropriate for instrument systems with different precision; improved precision will generally permit simpler and more cost-effective QC designs. We agree with Campbell's recommendation (*4*) that instrument standard deviations should be at least better by a factor of two than the medically necessary standard deviations. When that level of precision is achieved, the values for $\Delta SE_c$ will be 4.0 or larger. Such large errors will be relatively easy to detect, as shown in Figure 1.

The QC procedures selected for the Hitachi 737 illustrate how improved precision can influence QC practices. For 14 tests, the critical systematic errors exceed 4.0 and can be effectively controlled by using 3.5s control limits with two control measurements per run. Four other tests have critical systematic errors between 1.3 and 3.0 and require different control procedures; for these, we chose to use control limits of 2.5s with two control measurements per run. For albumin, this limit permits nearly ideal error detection (0.90, or 90% error detection). For chloride and total $CO_2$, error detection is about 60%, which could be increased by use of a multirule procedure. Addition of a $4_{1s}$ rule would increase error detection to the 70–80% range; addition of $4_{1s}$ and $8_{\bar{x}}$ rules would increase error detection to the 90% objective. We have chosen to handle calcium in a special way, making duplicate measurements of all samples and controls, and have installed two calcium test-channels in the instrument. The duplicate measurements are averaged by a special program in the laboratory computer that includes additional QC checks.

Application of these specific QC procedures to other laboratories depends on whether similar clinical requirements exist and whether similar analytical performance is

achieved. Differences in either may support use of different QC procedures as being more appropriate in other laboratories. The general design approach, however, is applicable in all laboratories, to all instrument systems, and to all analytes, and will lead to selection of control rules and numbers of measurements that are appropriate for each test in that laboratory.

The approach we outline here depends on the availability and proper application of power-function graphs. An even more practical development of cost-effective QC is available through simplification of the design approach by use of "quality-control selection grids" (*8*). These QC selection grids are 3 × 3 tables that identify QC procedures having error-detection and false-rejection characteristics appropriate for measurement procedures with different values of $\Delta SE_c$ and with different frequencies of occurrence of those errors. The grids should provide a practical planning tool that can be used in any service laboratory.

## References


1. Linnet K. Choosing quality-control systems to detect maximum clinically allowable analytical errors. Clin Chem 1989;35:284–8.
2. Westgard JO, Barry PL. Cost-effective quality control: managing the quality and productivity of analytical processes. Washington, DC: AACC Press, 1986.
3. Arkin CF. Quality control: What are our goals? How much is necessary? Pathologist 1985;August:19–25.
4. Campbell BG. Evaluation of two types of "medically significant error limits" and two quality control procedures on a multichannel analyzer. Arch Pathol Lab Med 1989;113:834–7.
5. Feldbruegge D, Liddicoat L, Dowd D, Koch DD. Complete evaluation of the Hitachi 737, with modification of the AST method to allow preincubation with pyridoxal 5'-phosphate (P5P) [Abstract]. Clin Chem 1986;32:1103.
6. Groth T, Falk H, Westgard JO. An interactive computer simulation program for the design of statistical control procedures in clinical chemistry. Comput Programs Biomed 1981;27:1536–45.
7. Westgard JO, Groth T. Power functions for statistical control rules. Clin Chem 1979;25:863–9.
8. Westgard JO, Quam EF, Barry PL. QC selection grids for planning QC procedures. J Clin Lab Sci, in press.