# PUMP UP YOUR PT IQ

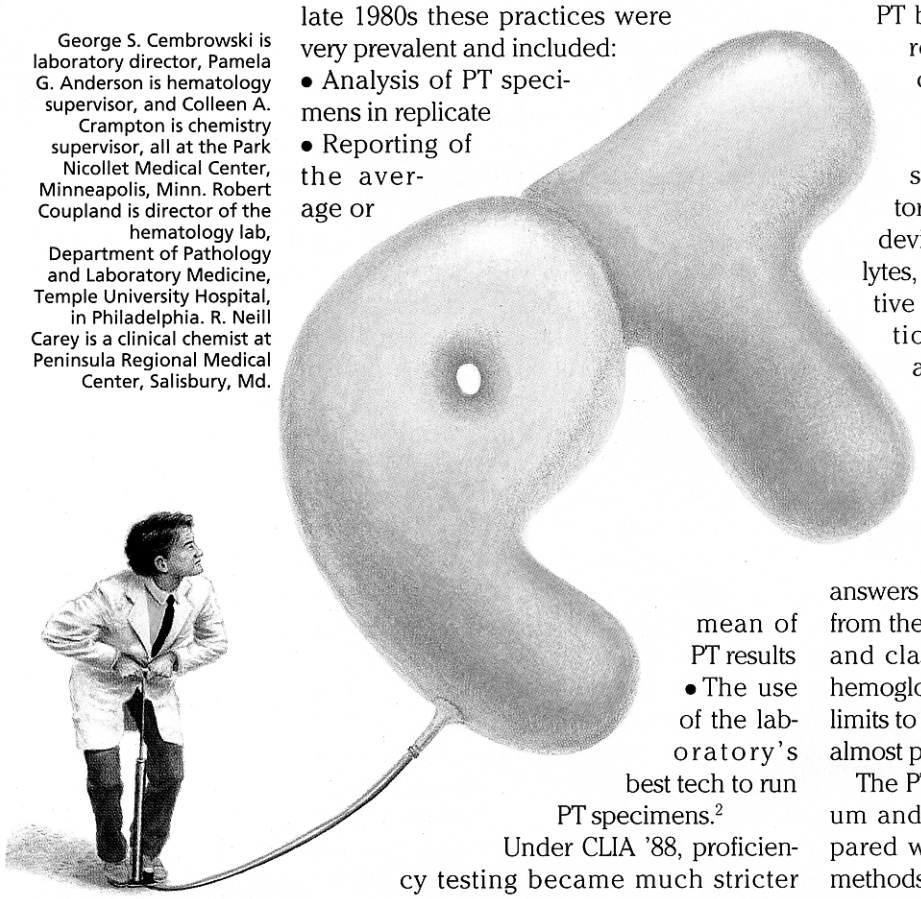
By George S. Cembrowski, M.D., Ph.D., Pamela G. Anderson, MT(ASCP), Colleen A. Crampton, MT(ASCP), Robert Coupland, M.D., and R. Neill Carey, Ph.D.

George S. Cembrowski is laboratory director, Pamela G. Anderson is hematology supervisor, and Colleen A. Crampton is chemistry supervisor, all at the Park Nicollet Medical Center, Minneapolis, Minn. Robert Coupland is director of the hematology lab, Department of Pathology and Laboratory Medicine, Temple University Hospital, in Philadelphia. R. Neill Carey is a clinical chemist at Peninsula Regional Medical Center, Salisbury, Md.

THE FIRST PT PROGRAM was created in 1946 when Belk and Sunderman[1] sent 12 different samples for chemistry and hemoglobin testing to volunteer clinical labs in Pennsylvania, New Jersey, and Delaware. Their survey documented tremendous inter-laboratory differences and a predominance of unsatisfactory results. In response, the College of American Pathologists began the first voluntary PT program.

Participation in PT was mandated by the Social Security Act of 1965 and its associated Medicare regulatory programs, as well as by CLIA '67. During the two decades that followed, most clinical laboratories instituted various preanalytical and analytical practices to improve PT performance. By the late 1980s these practices were very prevalent and included:

- Analysis of PT specimens in replicate
- Reporting of the average or mean of PT results
- The use of the laboratory's best tech to run PT specimens.[2]



Under CLIA '88, proficiency testing became much stricter and many of these special practices were stopped. In addition, the Health Care Financing Administration (HCFA) assembled a long list of analytes that must be tested successfully for a lab to pass PT.

American clinical labs that analyze these HCFA-defined analytes must run a minimum of five PT unknowns three times annually. HCFA can impose various sanctions on laboratories that perform PT unsuccessfully. Included among them are suspension, limitation, or revocation of the CLIA certificate, civil money penalty, civil suit, and even criminal sanction.

## COMPARISON OF HCFA LIMITS

Proficiency testing organizations evaluate PT by comparing the participating lab's results with HCFA limits. Figure 1 compares HCFA proficiency testing limits for some common hematology and chemistry analytes, as measured by typical large clinical laboratory analyzers, with long-term standard deviations (SDs). For many of the analytes, these PT limits are very broad relative to the long-term SDs. Large deviations can be tolerated before an answer is deemed unacceptable.

For example, hemoglobin is analyzed on many hematology analyzers with long-term coefficients of variations (CVs) of 1.0%–1.5%. The PT limits for hemoglobin (mean ±7%) are so broad that only answers deviating by more than 4.7–7.0 SDs from the mean would be outside these limits and classified as errors. For analytes like hemoglobin, use of HCFA proficiency testing limits to detect analytically significant error is almost purposeless.

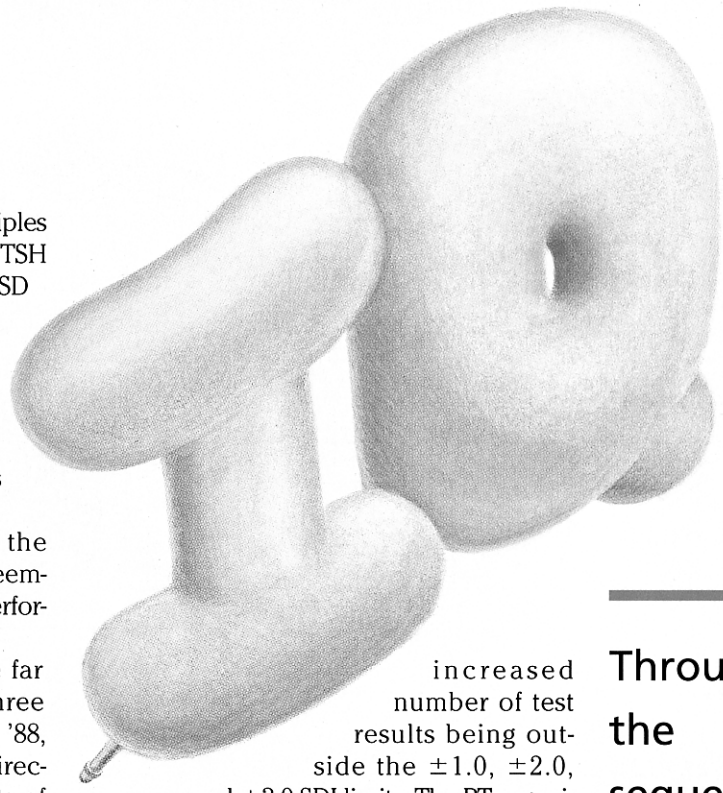
The PT limits for other analytes (e.g., sodium and chloride) are quite narrow compared with the SDs of currently available methods. For a few other analytes, including

Gary Phillips

TSH, HCFA limits are expressed as multiples of the group SD (i.e., mean ±3.0 SD). TSH assays with analytical shifts as little as 1 SD are at risk of failing PT.

While all laboratory supervisors inspect their PT results for failures, many do not evaluate these reports for significant, potentially correctable errors. Three reasons account for this shortcoming:

- CLIA's PT regulations emphasize the importance of passing PT and thus deemphasize efforts to improve laboratory performance continuously.
- While five PT results today provide far more information than the two or three results that were available before CLIA '88, today's supervisors (and laboratory directors) lack the tools to inspect these sets of laboratory data easily and logically.
- The delay between result reporting and receipt of PT results may be so long that investigating PT results is no longer useful.



## GOING ONE BETTER

We have developed a method for inspecting PT results efficiently and identifying important analytical shifts and random errors. We devised the method through a trial-and-error process and then characterized its performance with computer simulations.[3]

The method consists of the sequential application of several QC rules. First, a screening test is applied to the sets of five proficiency results. Then, if the test is positive, the data are tested for significant analytical shifts and random error. This multi-rule procedure is analogous to multi-rule QC procedures proposed by Westgard.[4]

To use this method, it is important the proficiency testing organization report PT results as standard deviation indexes (SDIs). The SDI is calculated as:

SDI = (result - mean)/SD

This index represents the number of standard deviations each result is from the mean. If PT data have a normal or Gaussian distribution, 68% of the results should be within ±1.0 SDI, 95.5% of the results should be within ±2.0 SDI, and 99.7% of the results should be within ±3.0 SDI. The presence of analytical error, either systematic (shifts) or random (increased imprecision), will result in an increased number of test results being outside the ±1.0, ±2.0, and ±3.0 SDI limits. The PT organization can generate two sets of SDI values depending on the mean selected. The mean can be either that of the laboratory's peer group or the all-method mean.

We recommend a lab evaluate the SDI values obtained by comparing its results to those of its peer group. If a lab is using the manufacturer's calibrators and operating an instrument as directed by that firm, the lab can do no better than to obtain proficiency testing results in the middle of its peer group. It is mostly the responsibility of the instrument manufacturer, not the individual lab, to obtain results that are as close as possible to the "true" value.

Figure 2 shows the rules we use to identify the presence of significant systematic error (a shift) or of random error (increased imprecision).

**Screening rule: 2/5 > ±1.0 SDI.** If two or more observations are outside the same +1.0 or –1.0 SDI limit, the screening rule is violated and the data are further tested with rules specific for systematic and random error. This rule usually is violated with shifts exceeding 1.0 SDI or random errors exceeding a doubling of the standard deviation.

**Mean rule: |Mean| > 1.5 SDI.** If the average of the five observations is > 1.5 SDI or <1.5 SDI, significant systematic error is present. The magnitude of the systematic error is equal to the magnitude of the average.

**1–3 SDI.** If one or more observations are outside either the +3.0 SDI or the –3 SDI lim-

**Through the sequential application of several QC rules, you can inspect PT results efficiently and identify important analytical shifts and random errors.**

Figure 1

**Comparison of HCFA PT error limits to long-term SDs**

| Analyte | HCFA error limit | Long-term SD |
|---|---|---|
| Hemoglobin | ±7% | 1 to 1.5% at 10 g/dL |
| White blood cell count | ±15% | 1.5 to 2% (normal range) |
| Platelet count | ±25% | 3% (normal range) |
| Prothrombin time | ±15% | 2% (normal range) |
| Activated partial thromboplastin time | ±15% | 3% (normal range) |
| Sodium | ±4 mmol/L | 1.5 mmol/L |
| Chloride | ±5% | 2.5% |
| Glucose | ±6% or 10 g/dL (greater of the two) | 1.5% |
| Calcium | ±1.0 mg/dL | 0.15 mg/dL |
| Cholesterol | ±10% | 2% |
| HDL cholesterol | ±30% | 6% |
| Thyroid stimulating hormone (TSH) | ±3SD | Not applicable |
| Human chorionic gonadotrophin (hCG) | ±3SD | Not applicable |

its, there is a high probability of random error.

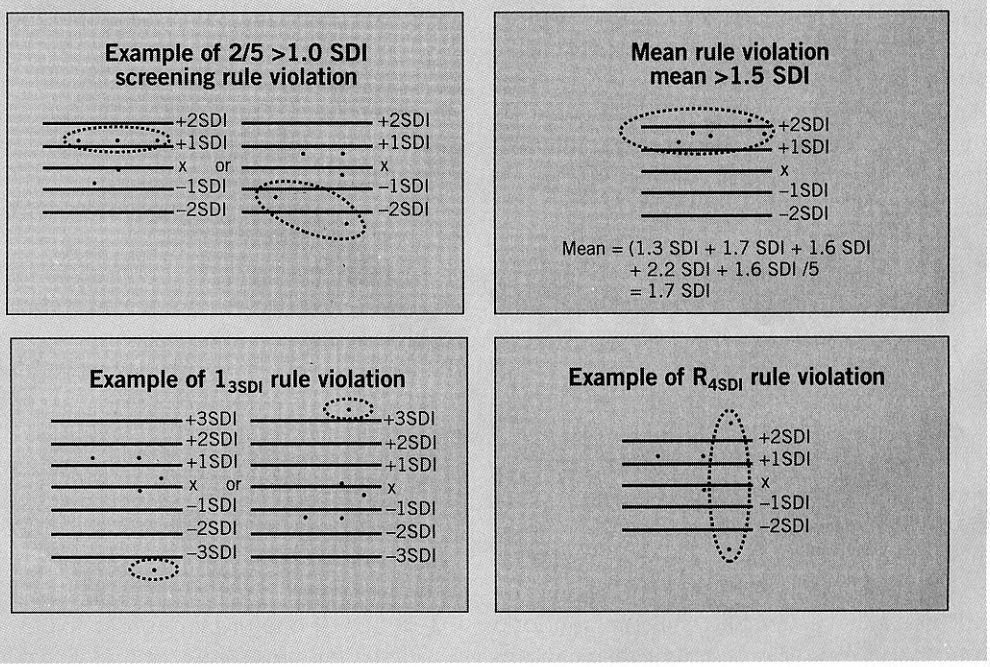
**R-4 SDI.** If the range or difference between the largest and the smallest PT result exceeds 4.0 SDI, there is a high probability of random error.

Figure 3 demonstrates the application of these rules. The five PT SDI data are screened for at least two SDI values, which are both outside either the same +1.0 SDI or –1.0 SDI limit. If this condition is not violated, there is a low probability of analytically significant error. If the screening rule is violated, then the five individual SDI observations are averaged to yield the average shift from the mean (the bias).

Systematic error is significant when the bias exceeds 1.5 SDI. If there is no significant systematic error, the observations are checked for random error. If

Figure 2

**Examples of rule violation**



one or more observations exceed 3.0 SDI or the range (difference between the largest and smallest) exceeds 4.0 SDI, then there is a significant random error. The investigation of random errors is more difficult, especially if the errors are sporadic. If the screening rule is violated and no systematic or random error is discovered, the SDI values of the next analyte are inspected.

## *APPLICATION TO ACTUAL PT EXAMPLES*

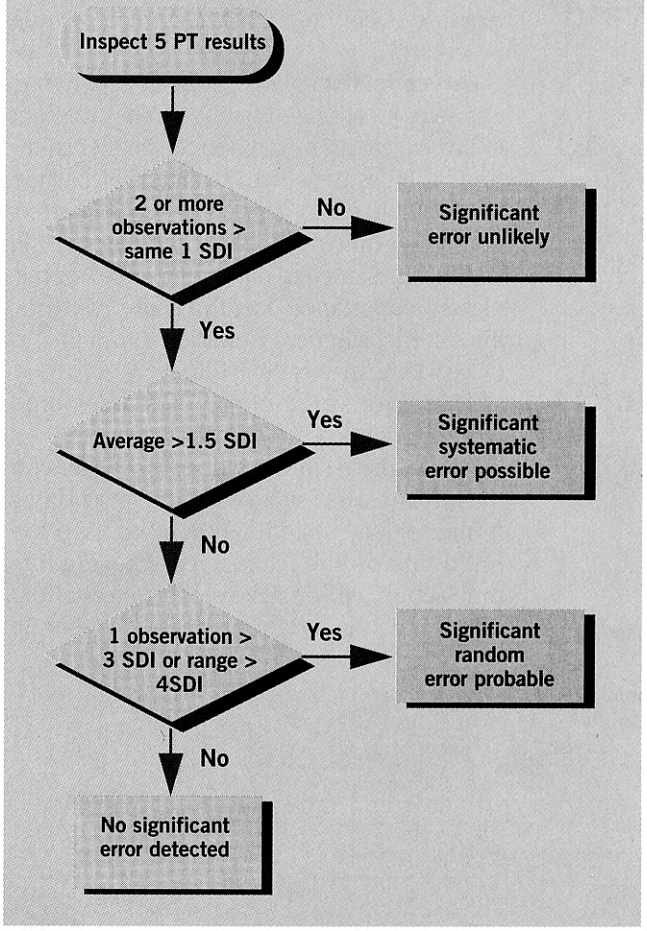
The following examples illustrate the use of the method. Facsimiles of CAP PT reports are shown for each (Figures 4 through 7).

**Example 1—HDL cholesterol.** Figure 4 shows violation of the 2/5 > ±1.0 SDI screening rule with LP-09 and LP-10 exceeding the +1.0 SDI limits. LP-10 that was 4.3 SDIs above the mean was outside the PT limits. We classified the error to be a random error, as it violated the 1–3 SD rule. When we reran LP-10, we obtained 25 mg/dL, which was virtually identical to the original group mean.

**Example 2—HDL cholesterol.** Figure 5 shows violation of the 2/5 > ±1.0 SDI screening rule, with observations LP-12 and LP-13 exceeding the +1.0 SD limits. The average shift was +1.44 SDIs. The occurrence of this shift, coupled with the high outlier from the previous survey (Example 1) and a letter from one of our clinicians complaining about HDL fluctuations, caused us to investigate. Inspection of CAP's unique graphical summary of the last four testing periods indicated sporadic HDL increases above the mean but very few observations less than the mean. Eventually, we attributed these increases to variations in decanting after LDL precipitation and centrifugation. We changed our procedure so our chemistry analyzer aspirates directly from the tube containing the centrifuged precipitate.[5] In this way we eliminated the decanting step, which can disturb the precipitate.

**Example 3—thyroid stimulating hormone.** Figure 6 displays a violation of the screening rule with four of the five PT observations exceeding +1.0 SDI. The average bias is +1.7 SDI, yet none of the TSH observations exceed the +3.0 SDI PT limit. When we examined the QC data at the time of reporting these TSH data, we observed a shift upwards but no violations of our control rules. Several weeks after we reported these PT results, our immunoassay instrument had regularly scheduled maintenance that resulted in the normalization of the QC data.

Figure 3
**Algorithm to evaluate PT results**




This case is a "near miss"; with such a large bias, we were lucky to have gotten all of our PT results within the +3.0 SD limit. The case illustrates the importance of minimizing bias with tests using PT limits of ±3.0 SDIs.

**Example 4—prothrombin time.** As shown in Figure 7, while all the observations exceed the +1.0 SDI limit, none are outside the PT limits. We classified the error to be a systematic one (average shift = 2.2 SDI). When we reran new samples of the same material, we obtained identical results, indicating a long-term systematic error. These results were consistent with long-term shifts in our interlaboratory QC program. These data were

**The multi-rule system for viewing PT data is very easy to teach; using it should result in uniform PT evaluation**

obtained from an older, repeatedly serviced coagulation analyzer. Because we didn't want to risk eventual PT failure, we replaced the instrument.

Following are details of some of the problems we encountered in using the multi-rule procedure.

**Overcorrection.** Some users of highly precise, easily calibrated analyzers may attempt to investigate and even correct smaller shifts, such as those between 1.0 and 1.5 SDI. Often this is unnecessary and leads to overcorrection. Corrections should be attempted if the shift is persistent and either of clinical[6] or regulatory importance, i.e., the shift plus twice the internal standard deviation (from QC) is close to or exceeds the PT limit.

There are several problems associated with attempts to correct small shifts:

- Most of these shifts are due to between-run variations that may occur in today's immunoassay systems. When we used the multi-rule on the PT results of several immunochemical analyzers, we found about 20% of the runs demonstrated systematic errors exceeding 1.0 SDI.[7] Only 5%–12% exhibited larger systematic errors, i.e., those exceeding 1.5 SDI. This is still a high rate of "false rejection"; however, it is necessary to maintain sensitivity to error that could cause future PT failures.
- Calibration is an inexact process. Attempts to lessen the bias through the process of recalibration may, in fact, result in even larger biases.
- Often there is a significant lag between proficiency testing and receipt of PT summaries. In the intervening time, the bias already may have been corrected by recalibration or another action.

**The PT program.** At a minimum, the PT organization should provide peer means and standard deviations. As the calculation of SDI values is tedious, it is almost mandatory PT providers offer SDI values. The more participants there are, the better the estimate of the peer mean and standard deviation. A laboratory should seek PT programs with at least 20 participants in the peer groups.

Inspection of PT data with the objective of error detection and correction is best accomplished if there is little delay between PT reporting and receipt of PT summaries. Even at this time, some PT participants are analyzing new PT samples before they receive summaries from their previous surveys. This is like playing double-blind *Jeopardy*. PT participants should demand rapid PT processing and result reporting.

The format of the proficiency testing report is extremely important. The HDL cholesterol examples demonstrate how the inclusion of information from previous challenges can help demonstrate the presence of long-term trends. While only a few proficiency testing programs currently provide this information in their reports, historical informa-

Figure 4
**HDL cholesterol PT, August 1993**

| Specimen | Your result | Eval. code | Mean | SD | No. labs | SDI |
|---|---|---|---|---|---|---|
| LP-06 | 70 | 13 | 67.9 | 5.8 | 408 | +0.4 |
| LP-07 | 34 | 13 | 33.5 | 3.4 | 412 | +0.1 |
| LP-08 | 24 | 13 | 21.8 | 2.3 | 415 | +1.0 |
| LP-09 | 61 | 13 | 53.8 | 4.7 | 411 | +1.5 |
| LP-10 | 36# | 13 | 24.7 | 2.6 | 412 | +4.3 |

Figure 5
**HDL cholesterol PT, November 1993**

| Specimen | Your result | Eval. code | Mean | SD | No. labs | SDI |
|---|---|---|---|---|---|---|
| LP-11 | 35 | 13 | 33.0 | 3.5 | 392 | +0.6 |
| LP-12 | 27 | 13 | 21.3 | 2.3 | 394 | +2.5 |
| LP-13 | 68 | 13 | 53.2 | 4.7 | 392 | +3.1 |
| LP-14 | 25 | 13 | 24.6 | 2.5 | 393 | +0.2 |
| LP-15 | 71 | 13 | 66.9 | 5.3 | 391 | +0.8 |

tion should be used as a criterion for evaluating PT programs.

**Limited utility.** We use some of the specimens from our main lab's Coulter STKS (Coulter, Miami, Fla.) to calibrate other manufacturers' hematology analyzers that we operate at 12 of our physicians' office labs. With this procedure, we are able to reduce the inter-instrument variation and thus minimize the variation of patients who have hematology testing at more than one site. Last year, we were perplexed when many of the PT results indicated consistent differences between our non-Coulter instruments and the PT means of their peers.

After appropriate investigation and consultation, we decided these differences arose from the calibration process. These differences made it impractical to apply the multi-rule to our non-Coulter hematology proficiency testing data.

Furthermore, as the calibration caused some of our hematology values to be close to the PT error limits, we began to report these instruments under the category "Other." While it would be far more practical to report these instruments by their instrument models and specify they were calibrated to a Coulter instrument, our proficiency testing organization does not allow this particular classification.

Figure 6

**TSH results, December 1994**

| Specimen | Your result | Eval. code | Mean | SD | No. labs | SDI |
|---|---|---|---|---|---|---|
| K-11 | 4.92 | 13 | 4.565 | 0.287 | 267 | +1.2 |
| K-12 | 12.38 | 13 | 11.963 | 0.926 | 268 | +0.5 |
| K-13 | 6.83 | 13 | 6.017 | 0.396 | 268 | +2.1 |
| K-14 | 10.23 | 13 | 8.910 | 0.699 | 269 | +1.9 |
| K-15 | 21.85 | 13 | 18.135 | 1.418 | 270 | +2.6 |

Figure 7

**Prothrombin time PT results, October 1994**

| Specimen | Your result | Eval code | Mean | SD | No labs | SDI |
|---|---|---|---|---|---|---|
| CG1-11 | 15.9 | 13 | 14.89 | 0.40 | 288 | +2.5 |
| CG1-12 | 13.0 | 13 | 12.56 | 0.25 | 287 | +1.8 |
| CG1-13 | 13.2 | 13 | 12.62 | 0.26 | 286 | +2.2 |
| CG1-14 | 23.5 | 13 | 22.02 | 0.86 | 288 | +1.7 |
| CG1-15 | 33.6 | 13 | 29.85 | 1.29 | 288 | +2.9 |

## *UNIFORM PT EVALUATION*

The multi-rule system for inspecting PT data is extremely easy to teach. As such, its use should result in a uniform style of proficiency test evaluation. With this multi-rule system, the need for improvements in technique and even changes in instrument model will be far more obvious. Finally, the laboratorian will be confident the PT data have been evaluated and acted on in the most efficient manner. MLO

**References**


1. Belk WP, Sunderman FW. A survey of the accuracy of chemical analyses in clinical laboratories. *Am J Clin Pathol.* 1947; 17: 853–861.
2. Cembrowski GS, Vanderlinde RE. Survey of special practices associated with CAP proficiency testing in the Commonwealth of Pennsylvania. *Arch Pathol Lab Med.* April 1988; 112: 374–376.
3. Cembrowski GS, Hackney JR, Carey N. The detection of problem analytes in a single proficiency test challenge in the absence of the Health Care Financing Administration rule violations. *Arch Pathol Lab Med.* 1993; 117: 437–443.
4. 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.
5. Overbagh B, Hohnadel D, D'souza JP, Mayer TK. 33% reduction of running time of HDL-cholesterol tests with no pour-off errors optimizing the BMC/Reagent set HDL-cholesterol precipitant method. *Clin Chem.* 1994; 40: 1098. Abstract.
6. Cembrowski GS, Carey RN. Medical usefulness requirements of analytical systems. In Cembrowski GS, Carey RN. *Laboratory Quality Management, QC and QA.* Chicago: ASCP Press; 1989. 80–99.
7. Cembrowski GS, Crampton CA, Byrd J, Carey RN. Detection and classification of proficiency testing errors in HCFA-regulated analytes: Application to ligand assays. *J Clin Immunoassay.* 1994; 17: 210–215.


There can be lags between PT and the receipt of summaries; in that time, bias may have been corrected