Process Capability and Stability of Analytical Systems Assessed from Proficiency Testing Data

Richard W. Jenny

Participation in a proficiency testing (PT) program is a valuable adjunct to laboratory activities dedicated to the maintenance of reliable analytical methods. The PT program may facilitate continuous quality improvement if laboratory performance is presented in the context of expectations espoused by healthcare professionals for optimal patient care. Statistical process control (SPC) and capability analysis are tools used by industry in a Total Quality Management environment to characterize and monitor the performance of its processes relative to performance specifications. I conceptualized the use of an analytical system by many laboratories as a process that periodically produces results from the analysis of PT specimens. I treated a set of five PT results (theophylline) reported by a laboratory as a process sample and subjected the samples collected from many laboratories to SPC and capability analysis. The control charts—\(\bar{x}\) (X-bar) and \(\bar{s}\)-charts—produced by the analysis readily identify significant analytical errors in the context of peer performance and performance specifications provided by the regulatory program and analytical goal setting. The capability index (desirable \(C_p > 1.0\)) determined from clinical specification limits for the three analytical systems evaluated suggests an opportunity for improvement of laboratory performance.

Indexing Terms: statistics/quality control/laboratory performance

Statistical process control (SPC) is used widely in industry to monitor the conformance of the producing process to specifications for reliability and precision (1). SPC is a valuable tool in an environment of Total Quality Management, where the emphasis is to prevent defective work from being produced, as compared with inspection of the final product, where the objective is to remove defects after they have been produced. In SPC, product samples are drawn from the output stream and the variable of interest is measured. The sample is presumed to represent the properties of the process product and the measure of accuracy and precision is used to judge conformance to specifications. If the sample exceeds control limits, the process is judged as unstable and action is taken to correct the problem before production of potentially defective products. Once process stability (statistical control) is achieved, process capability studies can be performed to determine whether the process is capable of meeting the expectations of those who use the product (2).

Much has been published on the value of SPC to industry (1–5). Here, I draw the analogy of an industrial process to laboratory testing, where an analytical system is used by many laboratories to produce test findings that are used by healthcare providers in the management of patients. This on-line process is sampled periodically by challenges with proficiency-testing (PT) specimens. I used a sample of five test results to judge the quality of the respective laboratory's performance relative to that of peers and to judge the capability of the analytical system to meet specifications established by healthcare professionals and regulatory programs. Control charts—\(\bar{x}\) (X-bar) and \(\bar{s}\)-charts—produced by SPC clearly identify process outliers and characterize laboratory error as systematic, random, or a combination of such errors. I applied the concepts of SPC and capability analysis to theophylline data collected by the New York State Department of Health (NYSDOH) Therapeutic Substance Monitoring PT program. The presentation format for laboratory and analytical system performance is a significant improvement over existing formats that provide little insight into the quality of analytical performance.

Materials and Methods

In New York State, 311 laboratories currently determine theophylline concentrations in serum specimens collected by physicians. Before providing this service, the laboratory must perform satisfactorily in the therapeutic substance monitoring PT program administered by NYSDOH. These laboratories are challenged three times each year with five serum specimens that contain theophylline and other therapeutic drugs at concentrations throughout the clinically relevant range. Acceptable performance is the recovery of drug to within 15% of the drug target value. The target value is determined from the weighed amount of drug added to the drug-free normal human serum matrix.

The specimens used in the PT program are prepared as admixtures of two or three master lots. The master lots are prepared by transferring an exact amount of drug to pools of processed normal human serum purchased from Scantibodies Laboratory (Santee, CA). The preparation of the master lots is validated in-house by HPLC and (or) immunoassay (Abbott TDx; Abbott Laboratories, Chicago, IL), and by referee laboratories. The admixture pools are passed through a 0.45-μm (pore size) filter (Gelman Sciences, Ann Arbor, MI), aliquoted into sterile 5-mL Wheaton vials, capped, and stored at −80°C until shipment. The test specimens are shipped.

---

1 New York State Department of Health, Wadsworth Center for Laboratories and Research, P.O. Box 508, Albany, NY 12201-0509. Fax 518-473-2900.

2 Nonstandard abbreviations: PT, proficiency testing; SPC, statistical process control; NYSDOH, New York State Department of Health; CV, process CV; \(C_p\), capability index; USL and LSL, upper and lower specification limits; and UCL and LCL, upper and lower control limits.

Received November 4, 1993; accepted January 14, 1994.
on frozen gel packs (Polar Pack; Mid-Lands Chemical Co., Omaha, NE) via United Parcel Service. Laboratories report their test findings and the method of analysis within 2 weeks of the date of shipment.

In accordance with industry experience that 20 samples of size 5 drawn from an on-line process are adequate to characterize process performance (4), I processed data collected from the three methods that were used by 20 or more laboratories for the analysis of theophylline: the Abbott TDx (n = 175), DuPont acu (DuPont, Wilmington, DE) (n = 28), and the Kodak Ektachem (Eastman Kodak, Rochester, NY) (n = 27). The laboratory test results from each of the five specimens were normalized as $z = (x - \bar{x})/\sigma$, where $z$ is the standard normal deviate, $x$ is the laboratory result, $\bar{x}$ is the method mean, and $\sigma$ is the method standard deviation. The mean and standard deviation of the standard normal deviates were tabulated for the respective laboratories. I then used the Quality Control module of StatSoft StatisticaTM (StatSoft, Tulsa, OK) to generate the $\bar{x}$ and $s$-charts. The process mean and control limits for the $\bar{x}$ chart are determined as $\bar{X} \pm 3\sigma/\sqrt{n}$, where $\bar{X}$ is the overall average of laboratory $x$-values and $\sigma/\sqrt{n}$ is the standard error of the mean. $\sigma$ is an estimate of the process standard deviation and, by the postulates of sampling distributions, $3\sigma/\sqrt{n}$ is estimated by $A_2\bar{s}$, where $\bar{s}$ is the average standard deviation, and the coefficient $A_2$ is a function of sample size obtained from tables of control chart constants (4). The $s$-chart control limits are determined as $B_3\bar{s}$ and $B_4\bar{s}$, where the coefficients are likewise dependent on sample size and are obtained from tables. The Statistica software automatically computes the process mean and control limits given the sampling data from the process. Because control limits are determined as 3 $\sigma$, thereby encompassing 99.73% of the sample means from a stable process, we can assume that outliers occur not by chance, but by a special cause of variation.

Two sets of specification limits are recorded on the theophylline $\bar{x}$-chart: PT limits and analytical goals for optimal patient care. The performance standard for the NYSDOH PT program is the recovery of theophylline from test specimens to within 15% of the target value. The specification limits for PT are determined as $\bar{X} \pm (15%/CV_p)$, where $\bar{X}$ is the process mean $x$-value and $CV_p$ is the process CV expressed as percent. The $CV_p$ is estimated as the average CV of the five specimen data sets. A proposed analytical goal for the total allowable error in the determination of theophylline concentrations in serum is $\pm 9\%$ (6). The $x$-chart specification limits for optimal patient care, therefore, are determined as $\bar{X} \pm (9%/CV_p)$. Given the specification limits, I used the Process Analysis module of Statistica to determine the capability of the respective analytical systems. The nominal value of the process (mean $x$-value) was set to zero and the capability index ($C_p$) was determined as $(USL - LSL)/6\sigma$, where USL and LSL are, respectively, the upper and lower specification limits for proficiency testing or clinical requirements. A $C_p > 1.0$ (desirable) signifies that >99.7% of the process's output will be within specification limits.

### Results

The outcome of the SPC analysis is threefold: characterization of the capability of the analytical system to meet clinical and PT program specifications, characterization of each participant laboratory’s performance relative to specifications and control limits established by peers, and characterization of outlier laboratory analytical error. This information is visualized from charts generated by the SPC analysis.

Figure 1 was produced from PT results reported by 175 laboratories that used the Abbott TDx for analysis of theophylline. This database yielded 175 samples of 5 results each that had been transformed to the standard normal deviate. The first step in the interpretation of SPC analysis is the review of the $\bar{x}$-chart (Fig. 1, bottom), which displays the standard deviation of each sample and flags the outlier samples relative to the process limits. Seven samples (seven laboratories) were flagged as reporting theophylline PT results with ran-

![Fig. 1. $\bar{x}$ (top) and $s$ (bottom) control charts for the Abbott TDx theophylline assay.](image-url)

175 laboratories using the TDx were challenged with a set of 5 PT specimens for theophylline analysis. Test results were normalized as the ratio of bias from method mean to method SD; the mean standard normal deviate and standard deviation were determined, and the 175 data sets were analyzed by SPC. USL and LSL of total allowable analytical error were charted for the PT program and for analytical goals (clinical). Circled points on the $\bar{x}$-chart identify laboratories with impression exceeding control limits on the $s$-chart; numbered points identify laboratories that exceeded performance specification limits. A. UCL = 0.94; B. method mean = -0.06; C. LCL = -1.06; D. UCL = 1.48; E. method mean = 0.70; F. LCL = 0.00.
dom error that exceeded the limits of system imprecision established by the performance of peers. These laboratories, regardless of the PT grade, should investigate the sources of special variation.

The second step is the review of the distribution of sample means around the process mean in the \( \bar{x} \)-chart (Fig. 1, top). The \( \bar{x} \)-chart includes three sets of limits: the upper and lower process control limits (UCL and LCL) determined by system performance as 3 SEM, the total allowable error consistent with optimal patient care (USL and LSL-clinical), and error limits prescribed by the regulatory program (USL and LSL-PT). The average CV of the five TDx data sets is 3.8%; hence, one standard normal deviate on the y-axis represents a systematic error of \( \pm 3.8\% \) [average sample bias = \((1/n) \cdot \Sigma(z_i \cdot CV_i)\); if \( z_i = 1 \), then bias = \((1/n)(\Sigma CV_i)\)]. Laboratories with a sample mean within control limits on both the \( \bar{x} \)- and \( \hat{s} \)-charts demonstrate excellent performance. Laboratories with sample means that exceed control limits but are within clinical specification limits perform at a level acceptable by process specifications; however, these laboratories have a mean bias (systematic error) >3.8% and the chart provides a warning of possible calibration drift.

Laboratories with sample means that exceed clinical specification limits but are within regulatory program limits meet regulatory standards but fail to meet an analytical goal for total allowable error that is consistent with optimal patient care (laboratories 43, 61, and 66). Finally, if the sample mean exceeds the regulatory specification limit, the laboratory has failed the analyte PT challenge (laboratories 7, 71, and 136). All references to the interpretation of the \( \bar{x} \)-chart presuppose that the standard deviation of laboratory results is within acceptable limits. Six of the seven laboratories that demonstrated excessive random error were within the specification limits on the \( \bar{x} \)-chart (Fig. 1, top, circled points), but their performance should be judged unacceptable.

The sample drawn from the stream of results produced by the analytical system is presumed to represent the performance of the TDx within the respective laboratories, and the cumulative samples to represent the quality of laboratory services as assessed by the capability analysis at the time of sampling. The process \( C_p \) is the ratio of the specification limits range (USL-LSL) to the process interval of 6\( \sigma \). The process is deemed capable if the \( C_p \geq 1 \); i.e., the process will generate \( \pm 3 \) or fewer out-of-specification results in 1000. The results of the capability analysis for the TDx are shown in Fig. 2. As shown, the values within 6\( \sigma \) of the nominal value are within the specification limits for the NYSDOH PT program and the \( C_p = 1.15 \). The clinical specification limits are inside the 6\( \sigma \) range and the \( C_p = 0.71 \). Of the test results generated by laboratories using the TDx, the fraction that is outside the clinical specifications can be approximated to be 3.4%: \( z = \text{USL} - x/\sigma \), so \( z = (2.4 - 0)/1.13 \) and, from normal curve probabilities, the fraction outside the specification limits is 2 (1.7) = 3.4%.

The outcome of the SPC and capability analysis for the DuPont aca and Kodak Ektachem theophylline assays are provided in Figs. 3–6. The procedures for interpretation of the charts are as described for the TDx.

**Discussion**

I conceptualized the use of an analytical system by many laboratories as a process by which a stream of test results are produced for use by healthcare personnel in the management of their patients. The device manufacturer provides the materials, training, standard operating procedures, and guidance for quality assurance/quality control with the objective of ensuring reliability in field use of the device. Many factors affect the overall performance of a device in the field. Tools used to monitor and characterize device performance relative to performance specifications designed into the device and those set forth by laboratory professionals (analytical goals) and regulatory agencies must be refined. I have shown that tools used by industry to control production processes are also useful in characterizing the performance of their devices being used in the field.

The precepts of statistical process control are the selection of samples that are representative of process output, statistically sound analysis of the samples to
determine process location (mean) and variability, and interpretation of control charts for identification of special causes of variation and conformance to specifications. The PT model is espoused by regulatory agencies and professional organizations as our current best effort to assess the quality of laboratory tests. The most oft-cited limitation to the model is that the behavior of some PT specimens is uncharacteristic of authentic patients' specimens in some analytical systems. These PT specimen–matrix effects introduce a bias that may result in misjudgment of the performance characteristics of the method. My interest in this investigation was to evaluate the process capability of an analytical system and the sensitivity of SPC for the identification of outliers among peers. To circumvent confounding matrix effects, I used the five theophylline challenges in a PT event to sample the quality of testing and used method mean and standard deviation to determine the standard normal deviate of each result. This represents a necessary first step to assess the capability of a device to meet performance specifications and to identify outlying laboratories that must initiate corrective action. Once the process is in control and found to be capable, the problem of significant method bias from PT program target values can be addressed.

The control limits for the $\bar{x}$- and $s$-charts are determined from the overall variability of the method test results (from the central limit theorem, the SEM of means is $s/\sqrt{n}$). The sample mean that exceeds the process limits (when the sample standard deviation is within limits) indicates a systematic error significantly greater than the inherent process variability. The three systems evaluated are calibration-stable devices for which the laboratory judges the need for recalibration from results for quality-control data. The systems are subject to calibration drift, and the $\bar{x}$-charts flagged systematic errors $>3.8\%$, $5.4\%$, and $6.0\%$ for the TDx, aca, and Ektachem, respectively, as exceeding limits of systematic error established by peer performance. Clearly, laboratories flagged as exceeding the PT program specification limit (PT failure) will investigate and correct error sources. However, the $\bar{x}$-chart provides useful feedback on the level of error relative to peers and specification limits, and laboratories so flagged can judge the need for correc-

![Fig. 3. $\bar{x}$ (top) and $s$ (bottom) control charts for the DuPont aca theophylline assay.](image)

28 laboratories using the aca were challenged with a set of 5 PT specimens for theophylline analysis. Test results were normalized as the ratio of bias from method mean to method SD; the mean standard normal deviate and standard deviation were determined; and the 28 data sets were analyzed by SPC. Chart description as in Fig. 1. A UCL = 0.83; B, method mean = $-0.03$; C, LCL = $-0.89$; D, UCL = 1.26; E, method mean = 0.60; F, LCL = 0.00.

![Fig. 4. Process capability analysis of the DuPont aca theophylline method.](image)

LSL (PT) LSL (clinical) NOMINAL USL (clinical) USL (PT)

$C_{u(PT)} = 0.86$

$C_{u(PT)} = 0.51$

$-3 SD$

$+3 SD$
The determination of the standard normal deviate for each of the five test results produces a data set that is amenable to a statistical analysis of assay performance characteristics. Variability of the standard normal deviates that is significantly greater than that of peers is documented in the $s$-chart and is indicative of either simple random error or concentration-dependent recovery of analyte. Significant systematic error is evident in those situations in which variability is stable but the data set bias exceeds specification limits on the $x$-chart.

I compared findings from the SPC technique with those of our current program's evaluation criteria. Five laboratories failed the June 1993 NYS PT event for theophylline; i.e., two or more of the laboratory's test results were discrepant by >15% from the target value. All five laboratories were also flagged by SPC as exceeding the PT specification limit (PT-USL or -LSL) on the $x$-chart. An additional nine laboratories displayed random error that was atypical of peer performance, exceeding control limits on the $s$-chart: Four data sets documented simple random error where analyte recovery was highly variable across the five challenges, and five data sets documented poor analyte recovery in either the subtherapeutic or toxic range of theophylline concentration. Several other laboratories performed within PT specifications but exceeded clinical specifications for assay performance. This information is clearly lacking in current schemes for the critique of laboratory performance in PT.

Process capability studies provide insight into the ability of a process to produce test results that lie within performance specifications. A $C_p > 1.0$ is desirable for a process and indicates that >99.7% of the process output is within applicable performance specifications. Typically, process capability is determined when the process is stable and special causes of variability are removed. As shown in Figs. 1, 3, and 5, outlying laboratories exist

**Fig. 5.** $x$ (top) and $s$ (bottom) control charts for the Kodak Ektachem theophylline assay. 27 laboratories using the Ektachem were challenged with a set of 5 PT specimens for theophylline analysis. Test results were normalized as the ratio of bias from method mean to method SD; the mean standard normal deviate and standard deviation were determined; and the 27 data sets were analyzed by SPC. Chart description as in Fig. 1. A, UCL = 0.86; B, method mean = 0.02; C, LCL = -0.82; D, UCL = 1.22; E, method mean = 0.59; F, LCL = 0.00.

An outcome of SPC that is lacking in conventional performance evaluation schemes is a clear presentation of the type and magnitude of analytical errors in the context of clinical and regulatory performance specifications. The determination of the standard normal deviate for each of the five test results produces a data set that is amenable to a statistical analysis of assay performance characteristics. Variability of the standard normal deviates that is significantly greater than that of peers is documented in the $s$-chart and is indicative of either simple random error or concentration-dependent recovery of analyte. Significant systematic error is evident in those situations in which variability is stable but the data set bias exceeds specification limits on the $x$-chart.

I compared findings from the SPC technique with those of our current program's evaluation criteria. Five laboratories failed the June 1993 NYS PT event for theophylline; i.e., two or more of the laboratory's test results were discrepant by >15% from the target value. All five laboratories were also flagged by SPC as exceeding the PT specification limit (PT-USL or -LSL) on the $x$-chart. An additional nine laboratories displayed random error that was atypical of peer performance, exceeding control limits on the $s$-chart: Four data sets documented simple random error where analyte recovery was highly variable across the five challenges, and five data sets documented poor analyte recovery in either the subtherapeutic or toxic range of theophylline concentration. Several other laboratories performed within PT specifications but exceeded clinical specifications for assay performance. This information is clearly lacking in current schemes for the critique of laboratory performance in PT.

Process capability studies provide insight into the ability of a process to produce test results that lie within performance specifications. A $C_p > 1.0$ is desirable for a process and indicates that >99.7% of the process output is within applicable performance specifications. Typically, process capability is determined when the process is stable and special causes of variability are removed. As shown in Figs. 1, 3, and 5, outlying laboratories exist

![Fig. 6. Process capability analysis of the Kodak Ektachem theophylline method.](image)
that are classified as special causes of variability. However, my interest was to determine the $C_p$ for the analytical system, to reflect the performance of the method as it is used in the field. All laboratories included in this study are reporting test results to physicians, and a sampling of method output by analysis of PT specimens is presumed to reflect the quality of their analytical services. Only the TDx had a $C_p > 1.0$ when calculated with the PT program specification limits of $\pm 15\%$ around the method mean. None of the analytical systems produced a $C_p > 1.0$ for clinical specification limits. The $C_p$ should prove to be a valuable statistic to characterize and monitor process performance. Clearly, the limiting factor in process capability is the performance specification designed into the analytical system. I anticipate that the $C_p$ for each analytical system will approach or exceed 1.0 as laboratories classified as special causes of variation identify the sources of error and improve their performance.

The x-axis of control charts typically records the time sequence as samples are drawn from the process output. The charts can then be interpreted for process trends, shifts, and special causes of random error. SPC as I have applied it to PT data does not produce charts that are amenable to such interpretation. However, the charts may be useful to visualize differences in performance among categories of laboratories or among laboratories with identifiable characteristics. For example, physician’s office laboratories, rural clinics, and tertiary-care centers can be grouped along the x-axis of the charts and the outcome of the SPC reviewed for relative analytical reliability. To illustrate the concept, I grouped laboratories participating in the NYSDOH PT program by the method of analysis for theophylline, selecting a random subset from the 175 TDx reports. The data were transformed as (reported value - target value)/0.075(target value), where the target value is determined from the weighed amount of drug, and 2 [(0.075)(target value)] is the allowable error. The charts produced from this analysis are shown in Fig. 7. Because the target value rather than the method mean is used in data transformation, the method bias from the nominal value may be an inherent method bias or be introduced by PT specimen-matrix effects. The analysis shows the positive bias of Stratus (Baxter, McGaw Park, IL) and Ektachem assays (Ektachem performance is adversely affected by PT specimen matrix) and the relatively low imprecision of the TDx and Roche (Nutley, NJ) fluorescence polarization immunoassays. For theophylline recovery, Emit (Syva, Palo Alto, CA) and aca were comparable with TDx and Roche, but interlaboratory variability was higher.

In conclusion, the control charts produced from analysis of PT data by the techniques of SPC and process capability analysis convey laboratory performance in the context of performance expectations. Interpretation of the charts is multidimensional: The PT provider can readily identify PT failures, the laboratory can assess the level and type of error relative to peers, the device manufacturer can assess the capability of its system in the field, and the healthcare provider can easily discern how well the system meets needs for patients’ care. The charts support activities of continuous improvement where, analogous to the Deming PDCA cycle (Plan, Do, Check, Act), the device manufacturer, the laboratory, healthcare providers, and regulatory agencies each has a role in defining performance specifications (planning), using the device according to accepted protocols (doing), monitoring device performance (checking), and responding to improve performance where it falls short of expectations (acting).

**References**