Peer-Review Research

Comparing Methods for Determining Out-of-Trend Stability Test Results

Richard Montes

Identifying out-of-trend (OOT) and out-of-specification (OOS) results is a critical part of stability data evaluation. An OOT observation is a measurement of a lot under investigation that deviates from or is atypical of the expected trend of representative historical lots, while an OOS result falls outside of specification. A confirmed OOT may not yet be an OOS, but it is often a warning that underlying issues may be at play. In the past, researchers studying this topic used the terms “analytical alert” for OOT readings that result from an invalid analytical measurement and “process control alert” for an OOT degradation rate (slope) that resulted from some production-related event (1,2). A “compliance alert” occurs when an individual product lot’s degradation rate is steeper and projected to fall short of potency and/or impurity specification limits through the product’s shelf-life.

Instituting corrective and preventive action (CAPA) to minimize the impact of a confirmed OOT is a more proactive control strategy than simply waiting for OOS conditions to occur. For an analytical alert, re-testing can be done while unexpired sample and reagents are still available to replace those that were affected by the OOT result. This practice prevents an invalid test from becoming part of historical lot data, which will affect future trending and setting specification limits of new drugs based on stability data. For a process control alert that is also a compliance alert, the expiry of an investigated lot may be shortened based on statistical projection to avoid OOS if the lot is allowed on the market at its original shelf-life.

To identify the OOT early, trending should ideally be performed by quality control (QC) laboratory personnel who generate the measurements as close to real-time as possible. The identification tools should be simple and practical so that they can be used by QC staffers who lack in-depth statistical training or access to a statistician. Simpler methods are crucial because the workload required for trending can be substantial, with multiple attributes to trend across several products. Moreover, the analyses must be repeatedly performed as new measurements are generated and added to the stability database.

There is no regulatory guidance spelling out which statistical techniques to use for OOT identification. Previously published

Submitted: August 24, 2020
Accepted: September 8, 2020

Comparing Methods for Determining Out-of-Trend Stability Test Results

Richard Montes

Trending stability data to identify out-of-trend (OOT) results is a critical part of avoiding out-of-specification (OOS) events for drug substances and products. Analytical OOT is usually caused by an invalid analytical measurement, while process control OOT is typically due to some production-related event. To date, there is still no regulatory guidance on which statistical techniques are best to use for identifying OOT conditions. Methods such as regression control chart (RegCC), by time tolerance interval (ByTimeTI), slope-by-lot Control Chart (SlopeCC), prediction interval (PI), and Z-score have been described in the literature.

However, there had been no systematic evaluation of the effectiveness of each of these methods. The author has used simulation to compare these methods under varying scenarios of sample size, relative lot-to-lot and within-lot variation, and extent of analytical or process control OOT. Results verified that PI, Z-score, and RegCC were related methods with some slight variation in how each one worked. ByTimeTI was the least effective in detecting either types of OOT. SlopeCC was effective in detecting process control OOT, but, overall, RegCC performed the best at detecting both analytical and process control OOT, although its false alarm rates (Type I error) slightly exceeded 5%. In addition, the method is relatively simple to use, and can be implemented by quality control staffers who have not had in-depth statistical training, or who cannot confer with a statistician.

However, there had been no systematic evaluation of the effectiveness of each of these methods. The author has used simulation to compare these methods under varying scenarios of sample size, relative lot-to-lot and within-lot variation, and extent of analytical or process control OOT. Results verified that PI, Z-score, and RegCC were related methods with some slight variation in how each one worked. ByTimeTI was the least effective in detecting either types of OOT. SlopeCC was effective in detecting process control OOT, but, overall, RegCC performed the best at detecting both analytical and process control OOT, although its false alarm rates (Type I error) slightly exceeded 5%. In addition, the method is relatively simple to use, and can be implemented by quality control staffers who have not had in-depth statistical training, or who cannot confer with a statistician.

However, there had been no systematic evaluation of the effectiveness of each of these methods. The author has used simulation to compare these methods under varying scenarios of sample size, relative lot-to-lot and within-lot variation, and extent of analytical or process control OOT. Results verified that PI, Z-score, and RegCC were related methods with some slight variation in how each one worked. ByTimeTI was the least effective in detecting either types of OOT. SlopeCC was effective in detecting process control OOT, but, overall, RegCC performed the best at detecting both analytical and process control OOT, although its false alarm rates (Type I error) slightly exceeded 5%. In addition, the method is relatively simple to use, and can be implemented by quality control staffers who have not had in-depth statistical training, or who cannot confer with a statistician.
research described three approaches to identify analytical OOT and process control OOT (3). The RegCC method is applicable for both types of OOT, while ByTimeIT and Slope CC methods are applicable for process control OOT. Research that studied the use of all three methods for an actual product advocated their simultaneous use in order to get a visual image of the results (4). In the end, however, it recommended the Z-score method. Other research evaluated those three methods as well as the Z-score method (5).

One study used historical estimates of intermediate precision for the RegCC method (2), and accounted for statistical significance in the slope method and also included prediction interval (2). Yet another study used a mixed effects model in a unified approach that incorporated both inter- and intra-lot variation to improve results (6).

All of this research has provided QC trending specialists with more information on statistical methods, their benefits and limits. The research described in this article was launched to provide a systematic and comparative evaluation of the effectiveness of statistical methods. The methods were compared by simulating various sample sizes, different levels of variation within and between lots, and varying extents of analytical and process control OOT. This article will describe details of the simulation study and results of the methods comparison.

Materials and methods

Statistical model for trended population. The population considered were product lots, the attributes of which were assumed to vary linearly over time, and can be described using a random intercept, common slope model (Equation 1; refer to Appendix for complete definition of symbols):

\[ Y_{ij} = \mu + A_i + \beta \times T_{ij} + E_{ij} \]  

where \( i = 1, \ldots, I \) lots, \( j=1, \ldots, J_i \) timepoints for lot \( i \),  

\( A_i \sim \mathcal{N}(0, \sigma_A^2) \), \( E_{ij} \sim \mathcal{N}(0, \sigma_E^2) \) and \( A_i \) and \( E_{ij} \) are independent so that \( Y_{ij} \) at time \( T_{ij} \sim \mathcal{N}(\mu + \beta T_{ij}, \sigma_A^2 + \sigma_E^2) \).

This assumption is justifiable, because most of the attributes that are evaluated in stability programs can be approximated by zero or first-order kinetics, which lend themselves to linear regression analysis. The release values at time of manufacture (intercepts) were expected to vary from lot to lot around some population mean \( \mu \) within some acceptable range (i.e., release limits). The rate of degradation over time (i.e., slope \( \beta \)) was expected to be common or fixed among the lots for a well-characterized and sufficiently controlled manufacturing process.

Historical lots simulation. To be a valid reference point for the subject lot being investigated, the historical lots should represent the population in terms of both analytical and process variations. Historical lots were simulated by drawing from a population described by Equation 1. Population parameters of \( \mu=100 \) and \( \beta=-0.15/\text{month} \) were chosen to mimic a stability-indicating attribute such as potency. The full stability testing schedule was set at 0, 3, 6, 9, 12, 18, 24, and 36 months, with a shelf-life (expiry) of 36 months. Simulation parameters (i.e., sample size, intra-class correlation coefficient [ICC], and extent of analytical and process control OOT) that were set for the study design are listed in Table I.

The sample size of available historical lots depends on where the product is in its life cycle. A product that is still under development or just recently approved may only have a small sample of stability lots available for trending, while a mature commercial product may already have a large sample. In this simulation study, 6 and 30 historical lots were selected as small and large sample sizes, respectively. The ages of historical lots at the onset of trending were unbalanced, due to staggered enrollment into stability study as lots were manufactured. For simplicity, half of the total number of lots were assumed to have results up to 18 months (i.e., half of a 36-month expiry) and half to have results up to nine months.

The ICC, which represents the correlation among measurements in the same lot is defined in Equation 2, wherein \( \sigma_A^2 \) and \( \sigma_E^2 \) were used without affecting generality.

\[ ICC = \frac{\sigma_A^2}{(\sigma_A^2 + \sigma_E^2)} \]  

[Eq. 2]

The ICC levels used in the study were 0.2 and 0.8. A data set with ICC=0.2 indicated that the between-lot (process) variability was small relative to within-lot (analytical) variability and vice versa for ICC=0.8.

Subject lot simulation. The subject lot was simulated in two ways. If an analytical OOT was being identified, the subject lot was first drawn from the same population as historical lots. The initially drawn, in-trend subject lot results were then induced to be analytically OOT by subtracting \( k_{OOT} \cdot E_{ij} \).

Table I. Levels for simulation parameters. OOT is out of trend.

<table>
<thead>
<tr>
<th>Simulation parameter</th>
<th>Levels studied</th>
</tr>
</thead>
<tbody>
<tr>
<td>Historical lots sample size</td>
<td>Small: [3(18), 3(9)] Large: [15(18), 15(9)]</td>
</tr>
<tr>
<td>Intraclass correlation coefficient ICC</td>
<td>0.2, 0.8</td>
</tr>
<tr>
<td>Extent of analytical OOT for subject lot (k_{OOT} \cdot E_{ij} subtracted from in-trend result)</td>
<td>( k_{OOT}=0.2,0.4 ) for no, moderate, and large analytical OOT, respectively.</td>
</tr>
<tr>
<td>Extent of process control OOT for Subject lot (slope is ( \beta_{OOT} ) steeper than ( \beta=0.15 ))</td>
<td>( \beta_{OOT}=0%,25%,50% ) for no, moderate, and large process control OOT, respectively.</td>
</tr>
</tbody>
</table>

* – Historical sample size notation is # of lots (maximum age of lots in Months). For example, the small sample noted as [3(18), 3(9)] consists of three stability lots at 0,3,6,9,12,18 months and three stability lots at 0,3,6,9 months.
Note that if an attribute wherein $\beta > 0$ (e.g., impurities) was being simulated, $k_{OOT}^*E_{ij}$ was added instead. Only the latest subject lot result was induced to be analytically OOT while all the previous subject lot data were in-trend.

When the next result was generated, the previous analytical OOT would have been re-tested and replaced with an in-trend result. If a process control OOT was being identified, the subject lot was simulated using a slope that is $\beta_{OOT}\%$ steeper than the true population slope $\beta = -0.15$. The three levels of $k_{OOT} (0, 2, 4)$ and $\beta_{OOT} \% (0, 25, 50)$ listed in Table I represent no, moderate, and large extent of analytical and process control OOT, respectively.

An example of simulated data. An example of a simulated data set [small sample, ICC=0.2, subject lot either with $k_{OOT} = 4$ or $\beta_{OOT} = 50\%$] displaying the results for the full stability testing schedule is shown in Figure 1. The available historical lots (A1 to B3) and the subject lot (C) initially drawn to be in-trend at the onset of trending are the filled circles while the ones that have yet to be generated through the remainder of the stability study are the blank circles. The induced analytical OOT at each time $j$ of subject lot are the blue X’s while the induced process control OOT are the red Y’s. At the onset of trending, the subject lot C had a result at 0 month (i.e., release data) while the A and B historical lots had results up to 18 and 9 months, respectively. The stability data set was dynamically added as the measurements per the full testing schedule were generated. For example, when the three months result was generated for the subject lot C, the corresponding $9+3=12$ months results for the B lots would have also been generated. When the six months result for subject lot C was generated, the $18+6=24$ months results for the A lots would have also been generated, and so forth. Each subject lot result was assessed for OOT as it was generated.

Methods to compare. The methods summarized above were systematically compared using simulation. The mixed effects version (6) was excluded from the comparison because of the additional complexity it entails in analyzing random effects and the fact that it requires the use of more advanced statistical software, which may preclude most trending QC personnel from applying this method. The key steps in applying the methods, formulas, and which data are used to construct the trending limits are summarized in Table II. The formulas for trending limits are two-sided, but only the lower limits were applied as appropriate for the attribute degrading over time examined in this paper.

The RegCC method uses common slope estimate $\hat{\beta}$ and root mean square error (RMSE) from fitting a separate-in-
Table II. Summary of the Out of Trend (OOT) identification methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>How it works</th>
<th>Formula for limits</th>
<th>Data used to construct limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression Control Chart (RegCC)</td>
<td>• Fit data with SICS ANCOVA model.</td>
<td>( Y_{c0} + \hat{\beta} \times T_{c_j} \pm Z_{\frac{1-\alpha}{2}} \times \text{RMSE} )</td>
<td>All Lots, all Months.</td>
</tr>
<tr>
<td>By Time Tolerance Interval (ByTimeTI)</td>
<td>• Calculate Mean and SD by Month.</td>
<td>( \hat{Y}<em>j \pm k</em>{1-\alpha} \frac{1-p}{n_j} \times \hat{\delta} )</td>
<td>All lots at each month</td>
</tr>
<tr>
<td>By Time Control Chart (ByTimeCC)</td>
<td>• Calculate Mean and SD by Month.</td>
<td>( \hat{Y}<em>j \pm \frac{Z</em>{1-p}}{2} \times \hat{\delta} )</td>
<td>None</td>
</tr>
<tr>
<td>Slope By Lot Tolerance Interval (SlopeTI)</td>
<td>• Fit simple linear regression by Lot.</td>
<td>( \hat{\beta} \pm k_{1-\alpha} \frac{1-p}{n} \times \hat{\delta} )</td>
<td>All Months by Lot.</td>
</tr>
<tr>
<td>Slope By Lot Control Chart (SlopeCC)</td>
<td>• Fit simple linear regression by Lot.</td>
<td>( \hat{\beta} \pm Z_{1-p} \times \hat{\delta} )</td>
<td>None</td>
</tr>
<tr>
<td>Prediction Interval (PI)</td>
<td>• Fit ANCOVA.</td>
<td>( \hat{Y}<em>{c_j} \pm \frac{\text{StdErr}</em>{c_j}}{n_j} \times Z_{\frac{1-\alpha}{2}} )</td>
<td>All Lots, all Months.</td>
</tr>
<tr>
<td>Z-score (Z)</td>
<td>• Fit ANCOVA.</td>
<td>( Z_j = \frac{Y_{c_j} - \hat{\theta}_{c_j}}{\text{RMSE}} )</td>
<td>All Lots, all Months.</td>
</tr>
</tbody>
</table>

(*) \( \text{StdErr}_{c_j} = \sqrt{\text{RMSE}^2 \times (1 + \frac{c_j}{(X'X)^{-1}})} \) where \( X \) = design matrix.

The significance level \( \alpha \) was fixed at 0.05. The proportion levels \( P \) of 0.90, 0.95, and 0.99 were used in simulation, but only the \( P=0.95 \) results are presented here, due to space limitations.

tercepts, common slope (SICS) analysis of covariance (ANCOVA) model to the historical data. The random effect model (Equation 1) simplifies to a SICS model when lot is treated as a fixed effect. The RegCC trending limit at each time \( j \) is calculated by adding to the subject lot intercept \( Y_{c0} \) the average change up to that timepoint \( (\hat{\beta} \times T_{c_j}) \) and subtracting the product of standard normal quantile and root mean square error, \( \frac{Z_{1-p}}{2} \times \text{RMSE} \).

As previously described (3), the trending limit for ByTime method is the constructed tolerance interval (TI) using the historical lots at each time \( j \). For the SlopeByLot method, the trending limit is the control chart (CC) limit calculated from the individual historical lot slope estimates. For completeness of comparison in this paper, both TI and CC versions were applied to ByTime and SlopeByLot.

Similar to RegCC, prediction interval (PI) and Z-score methods are also based on fitting a SICS ANCOVA model to applicable data. The difference is that, in RegCC, all the subject lot data were excluded from model fitting. In PI and Z-score, only the latest subject lot measurement being evaluated if OOT was excluded from model fitting. The PI trending limit at time \( j \) is the 100P% lower two-sided prediction interval at that timepoint for the subject lot. The Z-score trending limit at time \( j \), \( Z_j \), is the difference between actual and predicted value (i.e., residual) of the subject lot at that timepoint divided by RMSE. RMSE is the square root of the sum of squared residuals divided by the degrees of freedom of the error term from the ANCOVA output. Note that the Z divisor of RMSE used in this paper differs from other research/analyses/work which, instead, used the standard deviation of the residuals of only the subject lot (4,5).

Simulation performance metrics

Probability of correctly identifying OOT. A data set was simulated and the seven methods listed in Table II were applied to assess whether the result under evaluation (i.e., the latest available subject lot result) was OOT. For all methods except Z-score, the result was flagged as OOT if it exceeded (i.e., was lower...
than) the trending limit. For Z-score, the result was flagged as OOT if the absolute value of $Z_j$ was larger than that of the $Z_{0.025}$ standard normal quantile. This OOT flagging was iterated 1000 times. The proportion of the iterations (i.e., probability) flagging the result as analytical or process control OOT was used as the primary metric in the simulation study. A method that has a low false alarm rate (i.e., probability of flagging a result as OOT when it is truly in-trend; also known as Type I error) and high detection rate (i.e., probability of flagging a result as OOT when it is truly OOT; also known as power) is preferred.

Average ratio of the lower specification limit and lower trending limit. Trending limits to identify an OOT serve as an early alert before the underlying issue(s) caused an OOS event. The OOT trending limit should not be so restrictive that it incorrectly flags an in-trend result as OOT. Neither should the limit be so wide that it exceeds the specification and fails its purpose as an early alert before an OOS occurs.

A “hypothetical” lower specification limit (LSL) at each time $j$ was calculated using Equation 3 to bound what would be “acceptable” at that time point. In practice, shelf-life specification limit is a fixed value that drug product has to conform to throughout its shelf-life. The ratio of the hypothetical LSL and the lower OOT trending limit at time $j$, averaged over 1000 iterations, was used as a secondary metric for the simulation study. This secondary metric helps in interpreting the results of the primary metric. For example, a low detection rate may be due to extremely wide trending limits that even exceed the specification (i.e., LSL/OOT limit $>1$). This ratio metric is only applicable to RegCC, ByTimeTI, ByTimeCC, and PI methods, whose limits are measures of individual results from the population, which are compatible to the concept of specification limit.

$$LSL_j = \mu + \beta \times T_j - Z_{1-0.9973} \times \sqrt{\sigma_A^2 + \sigma_E^2} [\text{Eq. 3}]$$

Results

The probability of identifying as OOT the latest subject lot result (Y axis) becoming available from 3 to 36 months (X axis) were graphically summarized. Figures 2 and 3 compare the seven methods when subject lot was induced, at increasing extent from the top to the bottom row, with analytical or process control OOT, respectively. The four combinations of the sample sizes (small, large) and ICC levels (0.2, 0.8) were used to stratify results.

Detection rates for analytical OOT. To identify analytical OOT, the baseline case when the subject lot was in-trend was first
evaluated (Figure 2, $k_{OOT} = 0$). PI had false alarm rates (or Type I error) below 5%. Type I error rates for RegCC were slightly inflated averaging around 12%. Z-score started with about 25% false alarm rates but flattened down to about 14% when more data became available. SlopeTI and SlopeCC started with elevated false alarm rates of about 30% and 40%, but those rates decreased to below 5% and 10%, respectively, with more available data. With moderate analytical OOT ($k_{OOT} = 2$), RegCC and Z-score had comparable detection rates of about 45%, with PI lagging just slightly below. ByTimeTI performed worst, with detection rate as low as <5% for a small sample and ICC=0.8. ByTimeCC performed better than ByTimeTI but not as well as RegCC and Z-score. SlopeTI and SlopeCC started with high detection rates of 65 to 75% but those rates decreased as more data became available. With large analytical OOT ($k_{OOT} = 4$), detection rates increased to at least 65% for most of the methods, with similar relative performances among the methods as $k_{OOT} = 2$. The sample size and/or ICC stratification level had only minimal effect on detection rates for RegCC, SlopeCC, PI, and Z-score, but had a major effect on ByTimeTI and ByTimeCC. Sample size affected SlopeTI results, with worse detection rates found for small sample size.

**Detection rates for process control OOT.** For identifying process control OOT, the baseline case when the subject lot was in-trend, process-wise (Figure 3, $\beta_{OOT} = 0$), was simulated to be equivalent to in-trend, analytical-wise (Figure 2, $k_{OOT} = 0$). The false alarm rates of the methods for these two in-trend cases were therefore identical. With moderate process control OOT (Figure 3, $\beta_{OOT} = 25$%), detection rates of SlopeTI and SlopeCC started at about 30–40% but deteriorated as more data became available, except when the sample size was large and/or ICC=0.8. All the other methods started with low detection rates but improved as more data became available.

RegCC and SlopeCC had the best detection rates up to about 80% for ICC=0.8 when all data were available. The methods performed better when ICC=0.8 than they did when ICC=0.2, with little impact from sample size. ByTimeTI and SlopeTI performed worse on small than on large sample sizes. With large process control OOT ($\beta_{OOT} = 50$%), the detection rates were increased, reaching as high as 100% for the best performing scenarios observed with $\beta_{OOT} = 25$. SlopeTI and SlopeCC now monotonically improved with more available data, except for SlopeTI with small sample size at ICC=0.2, which deteriorated first before improving slightly.

**Ratio of LSL and OOT trending limits.** The width of the OOT limit relative to the hypothetical LSL (Equation 3) was evaluated by taking the ratio of the latter to the former. For a decreasing attribute, the OOT limit should be narrower (i.e., higher)
than the LSL if it is to be useful as an alert before OOS occurs. An LSL-to-OOT limit ratio of less than one is therefore desired. The ratios for RegCC, ByTimeTI, ByTimeCC, and PI with increasing extent of analytical and process control OOT are shown in Figures 4 and 5, respectively. For analytical OOT (Figure 4), ratios remained about the same regardless of $k_{OOT}$ extent and availability of data. This was because hypothetical LSL and OOT limit commensurately adjusted at each time. The only exception was PI, in which case the ratio decreased slightly within the first 12 months and then remained flat.

The ratios were higher (i.e., OOT limits were closer to LSL) when ICC=0.2 than when ICC=0.8 except for ByTimeCC, in which case the ICC level had no impact. Sample size had no impact except for ByTimeTI where ratios at small sample sizes were higher than large sample sizes, even exceeding one (i.e., the OOT limit was wider than specification limit).

For process control OOT (Figure 5), the relative method performances were similar to those in Figure 4, with only a slight change in PI. The ratios for PI increased slightly with the extent of $k_{OOT}$ and also increased as more data became available.

Discussion

Previous research had indicated that RegCC can be applied to detect both analytical and process control OOT’s, while ByTimeTI and SlopeCC are suitable for the latter type of OOT (3). The basis for this recommendation was not given. The current research directly compared prevailing methods to determine which one would perform best under varying scenarios. ByTimeTI performed the worst in detecting either analytical or process control OOT, which may be because the method uses only data at the evaluated time point, thus not utilizing data at other time points. It also ignores intra-lot correlation.

A supportive analysis was performed (results of which are not included in this article) that used TI for longitudinal data with random lot effects accounting for intra-lot correlation as described in prior research (7). This more complex TI method did improve upon ByTimeTI detection rates, but levels were still inferior compared to the results seen using other methods. TI is a direct function of sample size and can be excessively wide with small sample sizes. ByTimeCC eliminates the sample size-dependence of the tolerance multiplier $k_{OOT}$ by using $Z_{1-2}$ and thus improving upon ByTimeTI detection rates. However, ByTimeCC still did not perform well at ICC=0.8.

Simulation results confirmed the unsuitability of the SlopeCC for analytical OOT as previously described in other research (3). Although the method has high detection rates with limited data, it deteriorates as more data become available. This is because an analytical OOT result that occurs at later time points usually do not substantially affect the slope estimate that it exceeds the slope control limits.
However, SlopeCC is very effective in detecting process control OOT, confirming results of prior studies (3). SlopeCC has high detection rates when there is still very limited data available, but these results may just be the baseline performance of the method.

The false alarm rates when the subject lot is in-trend are also high. This is expected since, when only very few time points are available, the subject lot slope estimate and constructed historical lots slope control chart limits are unreliable.

PI, Z-score, and RegCC are related methods in the sense that all use parameter estimates from a SICS ANCOVA model to construct their respective OOT trending limits. The results of the research described in this article indicate that PI, Z-score, and RegCC, ranked in increasing effectiveness although the differences between them were slight. They are versatile enough to detect both analytical and process control OOT.

Comparing RegCC versus PI, the RegCC limit is a modified form of confidence interval of the predicted means. It uses the ‘observed’ subject lot intercept $Y_{c0}$ and $t_{1-\alpha}$ as a multiplier of RMSE to construct its limit. In contrast, PI uses the ‘fitted’ subject lot intercept $\hat{Y}_{c0}$ and $t_{1-\alpha}$ as a multiplier of RMSE. By definition, a prediction interval, which is what the PI method generates, should be wider than a confidence interval. This explains the slightly higher LSL-to-OOT limit ratio (i.e., wider limits) found with PI than with RegCC. The wider OOT limits for PI are less stringent, thus the method’s detection rates are lower than those for RegCC.

Contrasting RegCC versus Z-score, the two are conceptually similar, and the mathematical formulas that they use for trending limits represent different versions of the same expression. However, there are two differences: first, the respective use of observed versus fitted subject lot intercept; and second, the data used to generate RMSE for each method. These two differences are seen when RegCC is compared with PI. The RMSE for Z-score is slightly larger than that for RegCC. This is because both historical and subject lot results are included in the former (i.e., so there are more constraints) whereas only historical lots are used in the latter (i.e., fewer constraints). The larger RMSE results in smaller $Z_2$, and lower likelihood of flagging an OOT. This explains the slightly lower detection rates of Z-score compared to RegCC.

The research described in this article implemented the Z-score in a different way from prior work (4,5), which used the standard deviation of residuals (i.e., observed minus fitted values) as the divisor for $Z_2$. From the methodologies described, it is not clear what type of model was used to generate the fitted values. It is possible that a simple linear regression model was fitted, but only on the subject lot results. In this study, the $Z_2$ divisor used was the RMSE obtained after fitting the available historical and subject lot results (excluding the latest time point) with a SICS ANCOVA model. The RMSE, therefore, represents the deviation from this assumed inter-lot (i.e., common slope with varying intercepts) and intra-lot relationship. Using this RMSE to calculate $Z_2$, therefore, represents a more comprehensive implementation of the Z-score than found in prior research.

Detection rates for the best performing methods are only up to about 45% for a moderate extent of analytical OOT (Figure 2, $k_{OOT}=2$) and increase only up to about 65% at large analytical OOT ($k_{OOT}=4$). These magnitudes of induced analytical OOT are arbitrary and scaled as $k_{OOT}^*E_{0\sigma_u}$ where $E_0\sim N(0,\sigma_z^2)$ and the total variance ($\sigma_z^2+\sigma_u^2$) constrained to equal one per the ICC generalization. The relative performance among the methods is a more meaningful way to evaluate methods rather than the absolute values of the detection rates. To increase detection rates, the OOT trending limits can be rendered more conservative (e.g., by decreasing the proportion level used from $P=0.95$ to $P=0.90$). The challenge, however, is finding the right balance of stringent OOT limits that increase detection rates but do not inflate false alarm rates.

Previous research on this subject noted that the trending limits need only be established once for each product (2). Updating limits is only necessary if changes are made to the product stability profile or to the analytical methods used, or to get more precise estimates when more lots are available to be included as historical data. As supportive simulation analyses in this research, the OOT limits for RegCC were calculated using only the historical data available at the onset of
trending. The OOT limits were then “fixed” and used to trend the remaining subject lot results without updating, even as additional historical data were being generated. This approach contrasts with the dynamic updating of RegCC OOT limits as used in Figures 2–5. The supportive analyses showed very little difference between the performances of fixed vs. dynamic RegCC OOT limits. Thus, the extra effort required to update RegCC OOT limits dynamically does not appear to be worthwhile, although, in some cases, periodic updating may be performed at some pre-determined frequency that is manageable for the QC personnel.

Previous research (4) examined the simultaneous use of RegCC, ByTimeCC, and SlopeCC to better visualize results, but eventually recommended Z-score. Implementing multiple methods simultaneously may be too taxing for trending personnel given the number of attributes and on-going, repeated analyses needed. If a single method can be universally applied to both types of OOTs, it will simplify the implementation of any stability monitoring program. Among the methods compared in the scenarios studied in this research, RegCC performed the best in detecting both analytical and process control OOT with only a slight increase in Type I error rates. RegCC can be easily automated in an Excel spreadsheet to calculate limits at all times after inputting β and RMSE from ANCOVA output. SlopeCC worked just as well, but only for detecting process control OOT. This method requires estimating individual slopes and constructing historical slope-control limits at each time data cut-off. SlopeCC is therefore more complex to implement than RegCC, without resorting to statistically programmed codes for automation. The last supportive analyses performed was specifying β = 0.15. This indicates that the OOT trending methods can be applied for both stability non-indicating and indicating attributes.

Conclusion

Studies used simulation to compare directly statistical methods that are typically used to identify analytical or process control OOT conditions under varying scenarios. Findings indicate that a tolerance interval-based version of either by time or slope by lot methods showed the worst performance. However, the slope by lot control chart is effective in detecting process control OOT.

Prediction interval, Z-score, and regression control chart are three related methods based on fitting a separate intercepts, common-slope ANCOVA model. The three are listed in order of their increasing effectiveness in detecting both analytical and process control OOT. The differences among them are only slight. Overall, the regression control chart performed the best among the seven compared methods, with only a slight increase in false alarm rates (or Type I error). The method is relatively simple to implement in Excel spreadsheets, requiring only ANCOVA model estimates from common statistical software. If a QC laboratory is constrained to use only one method for its stability trending programs, the regression control chart is recommended. Using this method, QC staff can flag potential OOT in real-time, then consult a trained statistician to confirm OOT using more complex models, and institute CAPA as needed.

References