PharmaSUG 2020 - Paper SA-284

Implementing Quality Tolerance Limits at a Large Pharmaceutical Company

Steven A. Gilbert, Pfizer

ABSTRACT

Predefined quality tolerance limits (QTLs) were introduced in the revised ICH E6 (R2) Section 5 update to help identify systematic issues that can impact subject safety or reliability of trial results. This paper will focus on Pfizer's implementation of this requirement. The key focus will concern the approach with respect to loss of evaluable subjects, patient discontinuation and inclusion/exclusion errors, that is easily measurable attribute data. We discuss a team approach in setting tolerance limit as well as lessons learned in monitoring the progress and the important role of simulations in defining best practices for monitoring trials. Examples of simulation methods, signal detection through control charts suitable for short-run attribute data such as variable life adjusted displays and other graphical methods will be demonstrated along with example code. We reflect on preferred methodology and challenges inherit in various clinical trial designs ending with a look at future work needed to maximize the use of QTLs in mitigating trial risk and ensuring the integrity of published results.

INTRODUCTION

In this paper, we discuss the use of quality tolerance limits (QTLs), a statistical process control methodology, to facilitate the execution of high-quality clinical trials. With the release of the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH) E6(R2), QTLs are a requirement. There are challenges in the interpretation and implementation of these guidelines. This paper demonstrates Pfizer approaches to these challenges, shares SAS code for simulating and monitoring clinical trials.

The ICH guidelines provide a sparse definition of QTLs and how to use them. "Predefined quality tolerance limits should be established, taking into consideration the medical and statistical characteristics of the variables as well as the statistical design of the trial…" TransCelerate has published guidance in "Risk-Based Quality Management: Quality Tolerance Limits and Risk Reporting," and interpretation of QTLs as "the principles of QTLs presented in this paper are based on control limit methodology and extrapolations of other industries 'practices'."

Pfizer defines a QTL as a limit that defines when a trial may be in danger of having reduced validity (e.g., too many subjects have been lost to follow-up, and where reviewers question the reliability of the primary efficacy and safety results). The study teams are often asked, "what keeps you up at night thinking about this protocol design?". This approach is different from control limits, sometimes called alarm or action limits, which indicate that the trial is not progressing as expected (e.g., losing subjects to follow-up at a greater rate than historical trials) but not yet crossing the QTL. Note that literature on the QTLs can be a little confusing because the term, QTL, is used to denote both the actual limit and the entire process of monitoring with a QTL.

Also, there is confusion between the difference between key risk indicators (KRIs) and QTLs. Study teams use KRIs, to monitor similar parameters to QTLs, such as dropout rate. The difference is that KRIs are at the site-level; intended to find sites that are performing poorly by comparing them to either other sites or pre-defined fixed limits. QTLs are at the trial-level, intended to assess the quality of an entire trial. It is beneficial to leverage the QTL parameter, (e.g., dropout) as a KRI as well, because typically the first step in a root-cause analysis of a study level issue is to see if the issue is driven by a single or small number of sites.

OUR EXPERIENCE WITH EARLY ADOPTERS

Currently, Pfizer has had over 60 studies participating in the QTL process. Therapeutic area champions volunteered their time to help make the process a success across the company. A small cross-functional group took the lead as subject matter experts and worked with early adopters to provide just-in-time training and advice. The goal was to have teams describe and document the QTLs using the following template fields taken from the TransCelerate RBQM and Risk Reporting Appendix 1 (2017).

1. Parameter: The subject of the QTL
2. Definition: How will the QTL be measured and calculated
3. Justification for the Parameter
4. Unit of Measure: e.g., number, proportion
5. Expected Values: What is the expected value based on historical data and expert opinion
6. Justification for Expected Values
7. Quality Tolerance Limit: What is the QTL limit.
8. Justification for the Quality Tolerance Limit.
9. Planned Mitigation Actions: What will be done if there is a risk of crossing the QTL or if it is crossed. Secondary/Action limits can be described here as well.

Three general strategies were used to implement QTLs with early adopter study teams:

1. Exempt the trial from the QTL process based on trial type (e.g., non-interventional) and size.
2. Use of a “Bright line” rule where a single QTL and control limit is chosen. This was used in many early adopter trials that have initiated the QTL process after the trial had been ongoing. With new clinical trials, this approach will be used less often.
3. The use of statistical process control methods.

The remainder of this paper focuses on the use of statistical process control methods, particularly the use of control charts.

STATISTICAL CONSIDERATIONS

Underlying the Pfizer approach is a control chart methodology modified from statistical process control (SPC). Control chart methods in manufacturing go back nearly 100 years to Walter Shewhart while working at Western Electric. A reference states, “In the 1920’s Western Electric’s Dr. Walter Shewhart took manufacturing quality to the next level – employing statistical techniques to control processes to minimize defective output.” (Western Electric History). Control charts also have a long history in medicine going back over 60 years, first in laboratory settings and later introduced to clinical medicine in 1977 (Blackstone, 2004).

Control charts are graphs monitoring a process sequentially over time, for example, the number of subjects who did not meet an inclusion or exclusion criteria versus the calendar date. Control charts differ from a simple time series plot by augmenting the time series plot of the parameter over time with action or secondary limit lines to indicate when the process needs to be corrected. The term process in SPC indicates a production method (e.g., assembly line) in traditional use and a clinical trial in the current context. Control charts also differ from another standard quality control method, acceptance sampling. In a sampling framework, the focus is on the product. A random sample is taken from a manufactured batch, and the entire batch is either accepted or rejected based on the results of the random sample. For example, a soda bottling plant will fill bottles with soda. The bottles can then be inspected to make sure they are filled with the correct amount of soda. In a sampling framework, the batch of bottles can be sampled, and if the bottles are underfilled, the batch will be discarded. Process control, in contrast, focuses on the assembly line; bottles are examined on a regular basis, and if it appears that the bottles are beginning to be underfilled, action is taken to fix the bottling equipment before any bottles need to be discarded. Obviously, the process approach is more appropriate for a clinical trial where patients and their data are never discarded.

In SPC, the monitoring or charting of a process is the second, and more straightforward, part of a two-step process. In the first step, a process is optimized to minimize defective output and establish the average and variability of the output. Following up on the soda example, this would be the average volume of soda in a bottle and the standard deviation of the volume. If the average is too high or too low, it needs to be corrected, and as importantly, if the variability is too high, it needs to be lowered, or there will always be too many bottles that are over or underfilled.

In the clinical trial QTL process, the first step is skipped since there is no opportunity to have a run-in period just to see how the operational aspects of a trail are behaving. Instead, the best historical and subject matter expert knowledge is used to make an educated guess on how the process should behave.

THREE TYPES OF CONTROL CHARTS

Three useful and related control charts for attribute data (yes/no, fail/pass) methods that were explored are an observed minus expected difference chart (O-E difference), an O/E difference chart (Grunkemeier, 2009), and lastly, a cumulative probability plot. Underlying all three methods is a cumulative count of events of the parameter of interest. These methods were found to be more appropriate for clinical trials than more common SPC charts used in manufacturing. For example, *p* charts calculate the proportion of events in independent samples; the first 50 patients, the second 50 patients, etc. Samples of 50 are considered a minimum size for these methods but many clinical trials are run where the trials are too small to monitor the trial sufficiently often; furthermore, they are monitored at irregular time periods adding another layer of complications. Using the cumulative sum of events naturally includes all data, mitigating sample size concerns and can be used at any point in a trial. Another benefit is most trials have a fixed sample size and it is clear what the total count of events is expected and what is unacceptable at the end of the trial.

O-E DIFFERENCE CHARTS

The O-E difference chart plots the difference between the observed and expected counts on the y-axis and the cumulative number of subjects on the x-axis (see Figure 1). This chart is useful when each subject contributes a single Event/No-event to the cumulative sum, and the total number of subjects in the clinical trial is known at the beginning of the trial. The y-axis will be above zero when there are more events than expected and below zero when there are fewer than expected events.

The O-E difference chart is a modification of a method used to monitor surgeons introduced by Lovegrove (Lovegrove, 1997), known as a variable life adjusted displays (VLADs) or expected-observed cumulative sum charts (O'Neill, 2015). Lovegrove's original charts compared the cumulative sum of expected mortality versus the cumulative sum of observed mortality. Pfizer modified Lovegrove's VLAD methodology for use with QTLs in two ways.

First, the VLAD methodology uses a risk model to produce patient-specific predictions for the expected outcome, the probability of death. A simple version of this is used where every patient in the trial has the same probability of an event of interest, not necessarily death, but usually something far less dire such as being lost to follow-up. This probability is a single number used for all patients based on historical data and subject matter expertise, as described earlier.

Second, reverse the order of the observed and expected cumulative sums, and subtract the expected cumulative sum from the observed cumulative sum. This reverse action results in a chart that trends up when the process is out of control, and the number of events increases. Study teams found this approach to be intuitively reasonable as study teams think in terms of increasing numbers of adverse events as being bad; VLADs on the other hand trend up when things are going well, and the surgeon has lower than expected mortality rates. For the remainder of this paper this version of the chart will be called an O-E difference chart, though it is an example of a VLAD chart

The O-E Difference Method.

Every subject in the trial is observed to have an event $O_i=1$ or not have an event $O_i=0$ for i in $1…n$ subjects. Each subject will also have an expected number of events, E_i, where E_i equals the probability of an event, a number between 0 and 1, assumed to be the same for every subject in the trial. In other words, the O_i are Bernoulli random variables (Ross, 1988) with the probability of success E_i. The probability is chosen as the observed historical proportion seen in previous trials or from expert opinion if sufficiently similar trials are not available. For example, if events occur 5% of the time in previous trials, the expected proportion is $p_e=0.05$ and $E_i=p_e=0.05$ for each patient, i in $1…n$. Calculate at each time point, the difference O_i-E_i. The cumulative sum of the differences, V_n is then plotted over time, where;

$$V_n = \sum_{i=1}^{n}(O_i - E_i) = \sum_{i=1}^{n} O_i - \sum_{i=1}^{n} E_i = \sum_{i=1}^{n} O_i - np_e = (Number\ Observed) - (Number\ Expected)$$

If the events are independent, the cumulative sum of Bernoulli random variables has a binomial distribution with parameters n and p_e, Bin(n,p_e). Therefore, at any time point, n,

$$E[V_n] = np - np_e.$$

Where p is the unknown true probability of an event, and p_e is the expected probability of an event. Note, this assumes subjects enter the trial one at a time, and time is measured by patient number (i.e., number of patients in the trial at the time of monitoring), not calendar time. With the assumption, $p= p_e$, or, that the trial is behaving as expected and is 'in-control,' the expected value is 0. Furthermore,

$$Var[V_n] = np(1-p), \qquad SD[V_n] = \sqrt{np(1-p)}$$

Therefore, pointwise limits can be constructed. A normal approximation can be used with a mean of 0 and a standard deviation of √np(1-p). However, a more easily calculated approach in a data step is to use the quantiles of a Bin(n,p) distribution subtracting off np_e, to center the limits on 0, as illustrated below.

Simulating the O-E Difference Chart

Virtual (i.e., simulated) data for O-E difference charts can be simulated in a data step and used to determine operating characteristics of the method. An outside loop keeps track of trials while an inside loop keeps track of subjects in a trial. The following code illustrates the code for 20 simulated trials of 400 subjects each with a 0.1 expected proportion of events and a true event rate of 0.1:

```
data ds;
    pt= .1;             *true event probability;
    pe= .1;             *expected event probability;
retain cumsum 0 alarmn 1001;
call streaminit(1965);  *set random number seed for reproducibility;
do trial = 1 to 20;     *number of simulated trials;
do i = 1 to 400;        *number of patients per trial;
    subject = i;
    if subject = 1 then cumsum =0;  *reset count to zero for each trial;
    x= rand(“BERN”, pt); *generate random 0-1 event variables;
    cumusum = cumsum+x;
    expected = pe*i;
    vlad  = cumsum – expected ;
    ucl1  = quantile(“BINOMIAL”, .99, pe, subject);
    ucl   = ucl1-expected;
    alarm = (vlad > ucl1);
    if alarm = 1 then alarmn = subject;
    else if alarm = 0 then alarmn = 1001;
    output;
end;end;run;
```

The results of the data step simulation can be plotted in PROC SGPLOT see Figure 1. The amber line is the 99th quantile of a binomial distribution with *n* cumulative number of observations minus the expected count, np_e. Only the upper limit is plotted because excessive numbers of events are of concern. Also note the jagged nature of the line, which is due to using exact quantiles from a binomial distribution. The limit would be smooth if an approximation using a Normal distribution was displayed instead. The binomial approach is favored because it is more accurate, especially early on the trial when only a small number of subjects have been observed. Since the standard deviation is $\sqrt{np(1-p)}$ the amber limit line grows with the square root of *n*, resulting in the non-linear 'bullet' or 'rocket-tail' shape to the upper alarm limit.

The red line is the total excess number of events allowable for the trial, which is how the QTL was defined. The chart and limits were calculated assuming a historical event rate of 10%; therefore, at the end of the trial, it was expected to see 0.1x400=40 events on average. The QTL is based on a 15% event rate or 0.15x400=60 events. Please do not get confused by this. The O-E chart is centered at 0 by subtracting the expected count from the observed count. At the end of the trial, after 400 subjects, if the expected 40 events were observed, the chart will plot 0-E=40-40=0,and fall on the zero line. The QTL at the end of a trial with 60 events is charted as QTL-E=60-40=20. That is, if there are 20 or more events above the expected count at the end of the trial, the quality tolerance limit has been exceeded. During the clinical trial, the amber action limit is used to determine when mitigation actions should be taken to avoid crossing the quality tolerance limit at the end of the trial.

The calculations to plot the QTL on the O-E difference chart required knowing the exact size of the trial and the exact number of unacceptable events. The next two charting methods will not require knowledge of those two exact numbers, although they can be used when the sample size is known. Also, note that it is possible in this chart to cross the QTL early on and still have an acceptable result at the end of the trial. For example, many events might occur early on, however, the problem is mitigated, and the trial runs perfectly afterwards. The chart plots observed – expected, so if the observed count remains constant, the charted line will decrease over time if the observed count remains fixed while the expected count continues to grow.

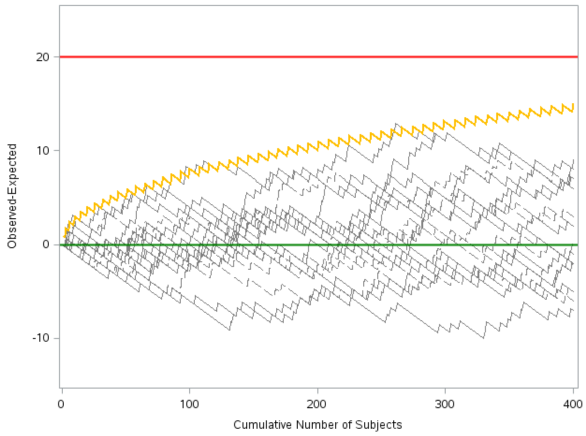

Figure 1 O-E Chart: 20 Simulated Trials with Upper Alarm Limit

THE O/E RATIO CHART

Some study teams focused on the number of positive events, for example, (e.g., the number of patients enrolled), instead of negative events (e.g. the number of patients lost to follow-up). This can be important for trials that need to meet a minimum database size for a regulatory commitment. The key difference here is not that the study teams keep track of a positive event, but rather they are targeting a count with no clear denominator to use to convert it to a proportion. For example, if the study team are looking to accrue subjects, the closest they can find to a proportion is the screen failure rate (number of subjects enrolled/number screened). The screen failure rate could be very low, if the trial accepts all subjects, but if enough subjects are not screened, the study will still fall short of its goals.

In this charting method, replace the total observed number of subjects with the total expected number of events. For example, compare the actual number of enrolled patients s at June 1st with the projected number of patients to be enrolled by June 1st. In one clinical study looking at an enrollment rate, a clinician, felt that if the observed enrollment at any time during the trials was less than 90% of the expected number, he needed to start mitigating a low enrollment problem before it got out of hand. Can the clinician's intuition be backed up using the methods introduced in the previous section? To some extent it can; if the approach uses a little more theory and modifies the data step program in a couple of crucial spots.

To address the problem, a Poisson process can be introduced, while using the previous code with minor modifications. To avoid confusion, note the term 'process' is used in two different ways in this paper. Earlier, a process was described as a production method, assembly line or clinical trial; now, process is defined in the sense of a stochastic or random process. The term process merely means random variables that are observed or simulated in code are related over time, or more colloquially a time-series.

A Poisson process is a process that counts events over time, starting with a zero count at time zero and continues to increase or remain the same over time, but never decreases. Assume the number of events between times t_1 and t_2, where t_2 is later than t_1, has a Poisson distribution with mean parameter $\mu=\lambda(t_2-t_1)$, where λ is the rate of events per unit time. Poisson distributions have the convenient property that the mean and variance are both equal to μ. For completeness the probability of k events is:

$$P(k) = e^{-\mu}\frac{\mu^k}{k!}.$$

A Poisson process is homogeneous if λ is constant over time, λ varies over time in a non-homogenous process. A scenario where the non-homogenous process could be useful is for events that are more likely early in a clinical trial when sites are still being trained. In addition, the random number of events on non-overlapping time intervals, say between day 2 and day 3, and day 5 and day 6 are independent. The astute reader may have already concluded this is very similar to the Bernoulli random variable simulation for the O-E difference charts. There is, however, a subtle difference. In those simulations, independent zero-one Bernoulli variables were generated independently for each subject. If there were *n* subjects, then the cumulative count could range from 0 to *n.* The Poisson process extends that methodology to allow counts greater than *n*.

Suppose a trial is expected to enroll 1000 subjects. Arbitrarily, break up the time axis into 10,000 non-overlapping intervals, t= {0, 0.0001, 0.0002, …., 1} using data step code, simulate 10,000 Bernoulli random variables with mean values of 1,000/10,000=0.1 (note the mean value of Bernoulli random variable is equal to the probability of an event). This simulation method should make sense; essentially, it is like flipping a biased coin with a probability of heads equal to 1/10, 10,000 times. Expectations are to see approximately 1,000 heads on average, with some simulations having more than 1,000 and others less. SAS makes it easy to change the code to generate Poisson random variables directly, replace x=rand("BERN",pt) with x=rand("POISSON, pt). Generating random variables from the Poisson distribution allows the cumulative count to increase by more than one per time interval, a Poisson random variable with a mean of 0.1 will be greater than 1 with a probability a little less than one-half a percent, or about 50 times in the 10,000 simulations. The probability of more than one event will decrease as the time intervals shorten, and the expected value used in the simulation decreases; however, using many short time intervals is inefficient, and the Poisson approach is recommended.

It is also necessary to change the output from the difference between the observed and expected cumulative counts to the ratio of the observed and expected counts. Change `observed-expected` to `observed/expected`. Finally, calculate a lower action limit as `quantile("Poisson," pt, expected)/expected`. Figure 2 displays the outcome of 100 simulated trials; the amber action limit captures the variability of the data well, which is simulated under the controlled condition of 1,000 expected rate of lambda=1,000/10,000=0.1. The figure cuts off wide swings at the very beginning of the trial. This is due to the instability of the O/E ratio with small samples. The O/E ratio does compensate for its early instability by a slow change later in the trial. Note that using 0.9 as an action limit is reasonable for most of the trial. The trial mentioned at the beginning of this section was an early adopter and did not start monitoring until it was well underway, and 0.9 action limit well approximated the statistically based amber action limit.

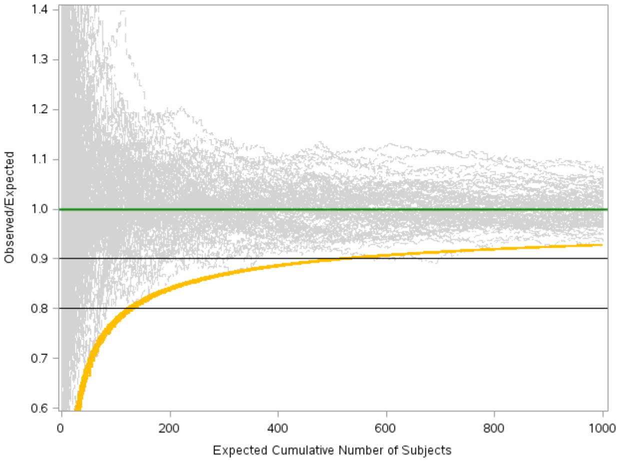

Figure 2 O/E Chart - 100 Simulated In-Control Trials

Figure 3 displays the results from simulating 100 trials, where the true average enrollment is only 800. Note how the amber action limit and 0.9 rule catch most of the trials early on.

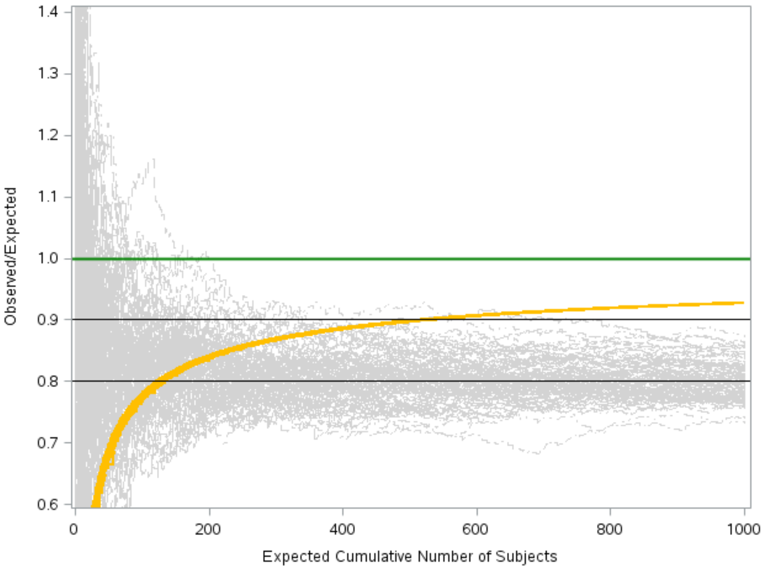

Figure 3 O/E Chart - 100 Simulated Out-of-Control Trials

These simulations create large datasets that can strain the ability of PROC SGPLOT. The data sets can be reduced in size by noting that the only records needed for plotting are the first and last for each trial, along with the records where the random count increases by one or more. The SGPLOT code can then be modified to plot each series with a step function instead of interpolating a straight line between points.

THE CUMULATIVE PROPORTION CHART

As a quick recap, the O-E difference chart was a useful approach when the study team knows during the planning of the clinical trial, the exact sample size and number of events that are undesirable; the x-axis, or 'time' axis, was the cumulative patient number. The O/E chart was useful when the exact number of subjects to use in calculating a percentage was not known, but the expected number of events at any point in the trial was known; the x-axis was now the expected number of events or time.

The cumulative proportion chart is another approach to define a cumulative or rolling proportion of events that are expected to be constant throughout a trial, whether the final size is known or not.

As an example, consider a vaccine trial where the parameter of interest was the collection of an important blood sample for immunogenicity testing. Assume we expect 95% of subjects to provide a sample at month 3. The cumulative or rolling proportion is calculated as the number of subjects with a lab collection at month three divided by the number of subjects in the trial long enough to have had a month three collection. An action limit is calculated using the 1st percentile of a binomial distribution. An action limit is calculated in a data step as `quantile("BINOMIAL", .01, .95, n)/n;` where *n* is the number in the trial long enough to have had a blood collection. Figure 4 displays 100 virtual in-control trials. The QTL was chosen as 80% based on subject matter criteria, and most all simulated trials are within the amber action limit as expected for an in-control process. One shortcoming of this method is that the charting and control limits are very wide early in the trial. However, this is not usually an issue since, in practice, it takes some time to get reliable data for monitoring, which will come as a bolus of patients. In this case, the results appear to be sufficiently precise after the first 50 subjects.

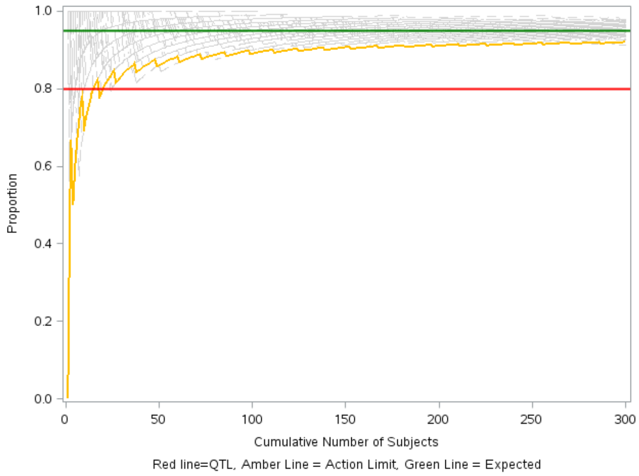

Figure 4: Cumulative Proportion Chart: 100 Simulated In Control Trials

PROCESS CAPABILITY

An important concept in statistical process control is process capability, which loosely speaking is the ability of a process to stay within its tolerance limits. In an O-E chart, like Figure 1, it means that the action limit (amber line) is less than the control limit (red line). For example, consider a 200-patient oncology trial where the parameter of interest is the number of violations of inclusion/exclusion criteria. The study team finds a historical rate of 10%; that is, they expect 20 violations and set a QTL of 25 violations. Does this make sense? This can be checked several different ways, starting with a simulation. Figure 5 plots 200 simulated trials. Since the team set the QTL at 25 patients and expects 20 patients, on this O-E figure, the red QTL line is set at 25-20=5, which is less than the amber action limit and many of the grey lines representing virtual trials. What has happened is the variability of the trial ‘process’ leading to events is too large. Even if the trial were to behave similarly to historical trials, there is a high probability of crossing the QTL of 25.

A simulation is not needed to know the trial results will be too variable to meet the QTL requirement. The same binomial quantile function used in our simulation code could have been used to determine likely outcomes at the end of the trial. For example, using quantile(“BINOMIAL”, .025, .1, 200) and quantile(“BINOMIAL”, .025, .1, 200) to calculate a 95% confidence interval yields limits of 12 and 29. A back of the envelope (or least calculator) computation can be done with the Poisson approximation. The expected value is 20, and the standard deviation is √20=4.47≈4.5, the mean plus/minus two standard deviations is now approximately 11 to 29, very close to the more precise binomial calculation. Lastly, a Normal approximation to the Binomial distribution with standard deviations √np(1-p) = √200*0.1*0.8=4.24 will quickly lead to the same conclusion. The process is too variable to have high confidence that the QTL will not be crossed. Actions will need to be taken from the beginning of the trial to lower the number of expected violations in this trial compared to historical trials.

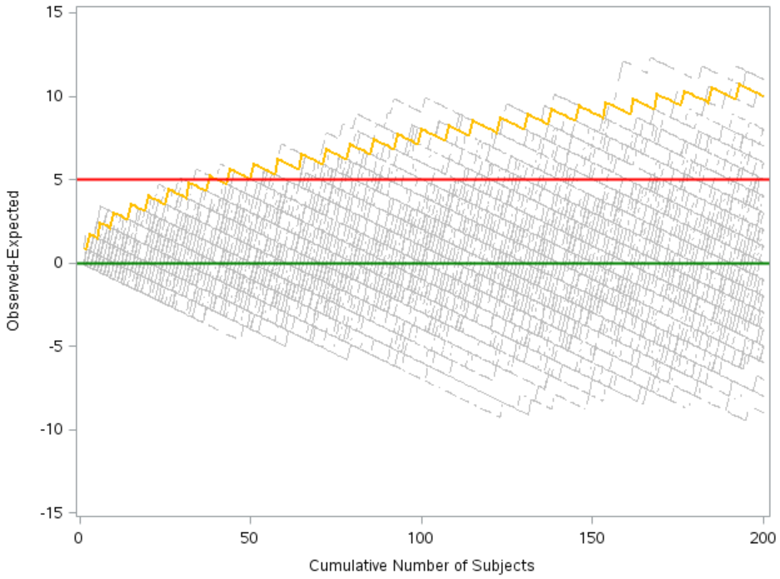

Figure 5 Inadequate Capability in and O-E Chart

SIMULATIONS FOR SMALL SUBJECT RULE

Simulations are also useful for deciding when a trial is too small to benefit from the QTL approach. Recall that the general idea behind SPC is to monitor a process that is in control, this means the trial behaves like historical trials and look for signs that it is drifting out of control. Control charts provide an early warning when either an action limit is crossed or some other criteria such as Westinghouse rules (Montgomery, 2009) indicate a signal. Mitigation actions are employed to fix the trial before a QTL is crossed. However, if a trial is too small there will be either a low probability of crossing an action limit, or a low probability of crossing the action limit early enough in the trial to mitigate any issues. The problem would go away if the trial were larger with more subjects. It is not recommended to increase the clinical trial size to better monitor a QTL, trial size should be determined by statistical and ethical considerations. A small trial can be difficult to monitor because proportions and other statistics can be imprecise due to a small sample size. This is different from process capability where at the planning phase of a clinical trial we are unsure the clinical trial can reliably avoid crossing a QTL if it behaves like previous trials regardless of the size of the trial.

An important property of SPC methods is how long it takes until the action limit is crossed, or the average run length or ARL (Montgomery, 2009). In the theoretical case of an infinitely long process (infinitely large trial) the control chart will eventually indicate an alarm at some point. If the chart works well, out-of-control processes will be detected quickly, and in-control processes will take a long time until a false alarm is raised. This concept is used to explore control charts for a variety of trial sizes and scenarios with a slight modification of the original simulation code to make the simulations more efficient.

O-E difference Chart Demonstration.

Consider categorical data again, with true error rates of 1%, 5%, 10% and 15% and determine when the QTL methodology can reasonably detect a doubling of the true rates to, 2%, 10%, 20% and 30%. To make the code more efficient, simulate a trial until the first alarm is reached, crossing the action limit, up to a maximum of 200 subjects to explore operating characteristics in small trials. The code uses a “do until loop” to stop after the first alarm, and only outputs one record per simulated trial, either at the virtual subject, where the first alarm occurs, or at the 201st iteration were the subject number is set to 250, indicating an alarm was never set. Flags are created in the code to make it easy to find what proportion of the time the trial is flagged after 37, 70, 100, 150 and 200 subjects by calculating the average of the flag using PROC MEANS. The code is illustrated below:

```
data s2;
      pt= .30;        *true event probability;
      pe= .15;        *expected event probability;
      retain cumsum 0 alarmn 250;
      call streaminit(1965);  *set seed to make simulation reproducible;
      do trial = 1 to 10000;  *number of simulated trials;
          alarm = 0;
          i= 0;
          done=0;
     do until(done);        *number of patients per trial;
          i= i + 1;
          subject=i;
          if subject = 1 then cumsum=0; *
          x=rand("BERN", pt); *generate random 0-1 variable;
          cumsum=cumsum+x;    *cumulative sume of variables;
          expected= pe*i;     *expected value;
          vlad = cumsum - expected;
          ucl1=quantile("BINOMIAL", .99, pe,subject) ;
          ucl=ucl1-expected;
          alarm=(vlad > ucl);
          if alarm = 1 then alarmn=subject;
          else if alarm = 0 then alarmn=250;
          done= (alarm = 1 or i = 200);
             *Flags for OCs;
             flag37 = (alarmn <= 37);
             flag75 = (alarmn <=75 );
             flag100 = (alarmn <= 100);
             flag150 = (alarmn <= 150);
             flag200 = (alarmn <= 200);
             flagNone = (alarmn = 250);
          if done = 1 then output;
          end;
    end;
  run;
```

In this simulation, the underlying variable needed to calculate the ARL is alarmn, the subject number when the first alarm is found. A complete summary of alarmn can be obtained using PROC UNIVARIATE, with the CDFPLOT option. Figure 6 displays the cumulative distribution function (CDF) for alarmn when the historical event rate is 5% (p_e=0.05) and the true rate is 10% (p_t=0.10). Due to the simulation code, which iterated up to 200 subjects, this plot is only useful for ARLs less than 200. ARLs were set arbitrarily to 250 for cases where the action limit was never crossed. This is seen on the far right of the plot where the CDF jumps up from approximately 80% to 100% indicating that about 20% of the time a doubling of a 5% rate will not be detected in a 200-patient trial. Recall that the focus is not on detecting a problem at the end of a trial when it is too late to mitigate but rather want to detect a problem early on. Checking the probability of an alarm half-way through the trial is a reasonable guide. Therefore, using this code to look for alarm rates after 37, 75, 100, 150, and 200 subjects is a way to assess whether it is effective in trials with total sample size of 75, 150, 200, 300, and 400 subjects.

Table 1 displays the alarm rates for the four scenarios, doubling of a 1%, 5%, 10%, and 15% historical rate. The column for 5% presents the probability of detecting a 10% rate after 37, 75, 100, 150, and 200 subjects respectively. These numbers match the values that can read from Figure 6, where the vertical grid lines cross the blue CDF curve. Note that even with a doubling of a 15% rate (last column in the table), there is only a 56% chance of an alarm after 37 subjects and doubling of a very low rate, 1% is difficult to detect even after 200 subjects. These results can be used to make general guidance on how large a trial should be to benefit from QTL monitoring.

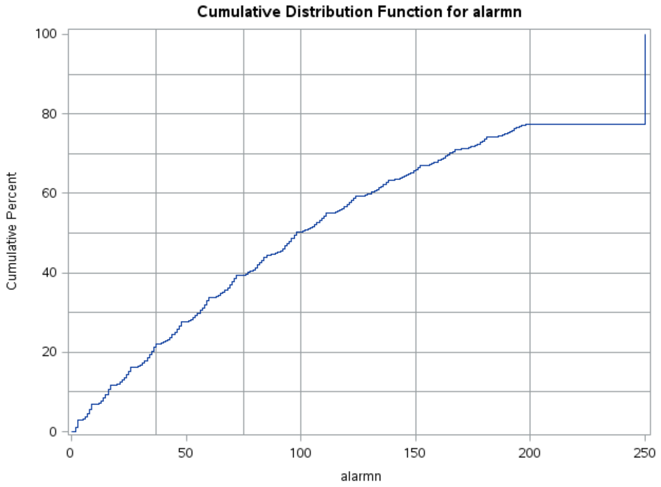

Figure 6 CDF plot for Average Run Length to detect doubling of 15% Error Rate

# of Subjects	Historical Rate			
	1%	**5%**	**10%**	**15%**
37	7.3%	22.2%	41.7%	56.3%
75	11.5%	39.5%	67.7%	86.4%
100	14.0%	50.2%	79.6%	94.2%
150	19.3%	66.1%	92.4%	99.0%
200	24.2%	77.5%	97.3%	99.8%

Table 1 Average Run Length to Detect Doubling of Event Rates

These calculations assume that QTL monitoring can be applied to an entire trial. In early drug development there are small trials that enroll subjects in small cohorts (e.g., single ascending dose (SAD), multiple ascending dose (MAD) and oncology trials to identify a maximum tolerated dose (MAD)), where it may not make sense to treat all of the subjects as belonging to a single trial. In these cases, each cohort may be very small, possibly 3 to 8 patients. These sizes are too small for meaningful use of QTL methods

ARE CONTROL CHARTS WORTH THE EFFORT?

It is tempting to skip the complications of charting the data and set a single QTL based on the final trial results, a simple “bright line” rule. For example, assume a trial team is planning a 200-patient trial and monitor the loss of evaluable subjects. For many situations, a 20% loss would put the validity of the trial in question, which is a 40-patient loss by the end of the trial. Historically the study team has seen a 10% loss. Therefore, they set up a simple monitoring process where if at any point, 25 or more subjects are lost, mitigation actions will be taken, assuming they have a 40-25=15 subject cushion to mitigate the issue on time. How well does that work compare to the O-E method? Since it is a simple rule, it can be coded into the previous simulation code and compared. This time the simulation is run using the historical 10% rate for calculating the action limit and a QTL rate of 20% to simulate the data. A simulation of 10,000 trials resulted in a 97.3% alarm rate using the O-E method and a 99.8% rate for the simple N=25 rule.

At first glance, it appears that the simple rule worked better. Is all the work for the control chart wasted? Not at all. Digging a little deeper, look at how well the rules compare if the process is in control, that is, if there really is a 10% error rate, and calculate the ARL of the last section. If the simulation is rerun with the process in-control, any observed alarms are false alarms; the O-E method has 6.7% false alarm rate while the N=25 rule has over twice that rate, 14.6%. The N=25 rule looks even worse when the ARLs are calculated. The O-E method has an ARL of 58.9, a little over one-quarter of the way through the trial while the ARL for the N=25 rule is 125, over half-way into the trial, leaving far less time to mitigate problems. The fact that the bright line rule produced an alarm later in the clinical trial than the control chart method makes sense in retrospect. The bright line rule was based on finding a percentage that looks concerning towards the end of the trial while the O-E method automatically adapts to how far the trial has progressed and can be triggered by unacceptable events early in the trial.

CONCLUSION

The QTL methodology needs to be implemented by clinicians, data managers, programmers, and other members of the study teams. These are the people closest to the data and science; they require transparent and easily interpretable methods, not black boxes, to define and make use of QTLs. The charting methods and action limits are here to leverage the study teams’ expertise not to replace it. There are many types of control charts, some claiming optimal statistical characteristics but usually at the expense of interpretability. The focus should be on efficient but interpretable methods. In this manuscript three methods were introduced: The O-E difference, O/E ratio, and cumulative probability charts due to their simplicity and ease of understanding for audiences not accustomed to statistical process control methods or surgical surveillance. These approaches are in keeping with the history of SPC, where non-experts implemented methods designed by statisticians and engineers in SPC on the manufacturing lines in their day to day work. These three charts cover many situations seen in clinical trials, but not all. One area for further work is to introduce charts for censored data. An example is patient dropout, in the cases so far, a definite time point early in the trial was defined, say a 30-day assessment to monitor. A dropout could occur any time during a trial. The complication is that when data are monitored there are now three types of subjects instead of Yes/No in earlier examples: subjects who have completed the trial (Yes), subjects who definitely were lost to dropout (No) and those who are still in the trial and may or may not complete (Maybe). Making use of the ‘Maybe’ data along with how long a subject has been monitored is amenable to survival analysis (time to event analysis) and will be an area for further development.

REFERENCES

Blackstone, EH. 2004. “Monitoring surgical performance.” *The Journal of Thoracic and Cardiovascular Surgery*, 128:807-10.

Grunkemeier, GL, JIn R, Wu YX. 2009. “Cumulative Sum Curves and Their Prediction Limits.” *Ann Thoracic Surgery*, 87:361-4.

O’Neill, S, Wigmore SJ, Harrison EM. 2015. “Debate: should we use variable adjusted life displays (VLAD) in identify variations in performance in general surgery?” *BMC Surgery*, 15:102.

Lovegrove J, Valencia O, Treasure T, Sherlaw-Johnson C, Gallivan S. 1997. “Monitoring The Results of Cardiac Surgery by Variable Life-Adjusted Display.” *Lancet*, 350(9085):1128-1130.

Ross, S. 1988. *A First Course in Probability*, Englewood Cliffs, NJ: Prentice Hall

Montgomery, D. 2009, *Introduction to Statistical Quality Control, Sixth Edition*, John Wiley and Sons

TransCelerate . Risk-based quality management: quality tolerance limits and risk reporting. http://www.transceleratebiopharmainc.com/wp-content/uploads/2017/09/Risk-Based-Quality-Managment.pdf. Published 2017. Accessed November 14, 2017.

International Council for Harmonization. E6(R2) Good Clinical Practice: Integrated Addendum to ICH E6(R1): Guidance for Industry. https://www.fda.gov/media/93884/download. Published 2018.

Western Electric History, http://www.porticus.org/bell/doc/western_electric.doc

ACKNOWLEDGMENTS

I would like to acknowledge my colleagues Stacy St. John, Dan Evans, Patti Compton, Rod Junor and Melissa Suprin.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Steven A. Gilbert
Pfizer, Statistical Research and Innovation
Steven.a.gilbert@pfizer.com