Identification of Out-of-Trend Stability Results

A Review of the Potential Regulatory Issue and Various Approaches

PhRMA CMC Statistics and Stability Expert Teams

An out-of-trend (OOT) result is a stability result that does not follow the expected trend, either in comparison with other stability batches or with respect to previous results collected during a stability study. This article discusses the regulatory and business basis, possible statistical approaches, and implementation challenges to the identification of OOT stability data. It is intended to begin a dialogue toward achieving clarity about how to address the identification of out-of-trend stability results.

Members of the PhRMA CMC Statistics and Stability Expert Teams are listed in the Acknowledgments section of this article. All correspondence should be addressed to Mary Ann Gorko, principal statistician at AstraZeneca, 1800 Concord Pike, PO Box 15437, Wilmington, DE 19850-5437, tel. 302.886.5883, fax 302.886.5155, maryann.gorko@astrazeneca.com.

Out-of-specification (OOS) regulatory issues have been well documented in the literature (1). Out-of-trend (OOT) stability data identification and investigation is rapidly gaining regulatory interest. An OOT result is a stability result that does not follow the expected trend, either in comparison with other stability batches or with respect to previous results collected during a stability study. The result is not necessarily OOS but does not look like a typical data point. This article discusses the regulatory and business basis, possible statistical approaches, and implementation challenges to the identification of OOT stability data.

Representatives from PhRMA member companies met to consider these topics, review current practices, and summarize various approaches to potentially address this issue. It is noted that the identification of OOT results is a complicated issue and that further research and discussion is needed. This article is not a detailed proposal but is meant to begin the dialogue toward achieving more clarity about how to address the identification of out-of-trend stability results.

Regulatory and business basis

A review of recent Establishment Inspection Reports (EIRs), FDA Form 483s, and FDA Warning Letters indicates the identification of OOT data is becoming a regulatory issue for marketed products. Several companies recently have received 483 observations requesting the development of procedures documenting how OOT stability data will be identified and investigated.

It is important to distinguish between *OOS* and *OOT* results. FDA issued a draft OOS guidance (2) following a 1993 legal ruling from *United States v Barr Laboratories* (3). Much has been written in the scientific literature and discussed at many scientific conferences about OOS results. Although the FDA draft guidance indicates in a footnote that much of the guidance presented for OOS can be used to examine OOT results, there is no clearly established legal or regulatory basis to require consideration of data within specification but not following expected trends.

The 1993 legal ruling from *United States v Barr Laboratories* stated that the history of the product must be considered when evaluating the analytical result and deciding on the disposition of the batch. Also, common sense would indicate that trend

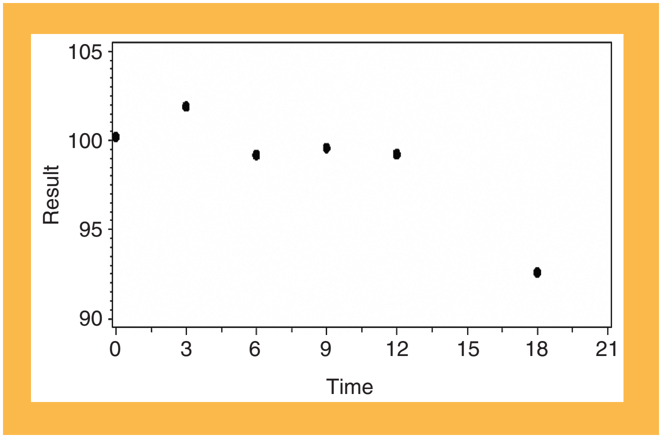

Figure 1: OOT trend within a batch.

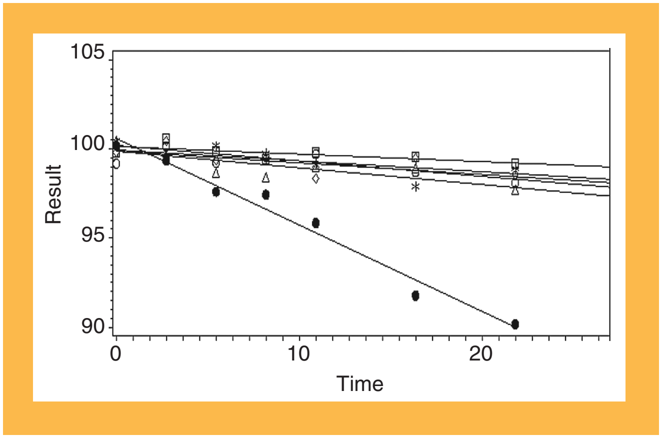

Figure 2: OOT trend across batches.

analysis could predict the likelihood of future OOS results. Avoiding potential issues with marketed product, as well as potential regulatory issues, is a sufficient basis to apply OOT analysis as a best practice in the industry.

The extrapolation of OOT should be limited and scientifically justified, just as the use of extrapolation of stability data is limited in regulatory guidance (ICH, FDA). The identification of an OOT data point only notes that the observation is atypical.

Another strong motivator to identify OOT data is the fact that stability data are used for various business applications. Stability data are provided in regulatory submissions and are used to justify the expiration dating of the product. Moreover, many firms use stability data to calculate internal release limits that a product must meet to ensure that the true means of the analytical property will remain within specifications throughout the dating period. Stability data are used in this process to estimate the amount of product change that will occur during the expiry period, the consistency of the change in the analytical property from lot to lot, and the assay variability. Thus, it is very important that any data outside expectation be identified because these data can have a substantial effect on the calculations performed. Identification methods should discriminate between substantive events and spurious values expected from the inherent randomness.

In summary, the issue of OOT is an important topic both from a regulatory and business point of view. Despite this, little has been discussed in the scientific literature or in regulatory guidance on this topic. This article will introduce some approaches that might be used to identify OOT data and discuss some issues that companies will likely need to address before implementation and during use of an OOT identification procedure. Given the complicated nature of this issue, this article is only intended as a start of the discussion about this topic because many issues are not easily resolved and further research and discussion is definitely warranted among all parties involved.

Statistical approaches

Background. There is a need for efficient and practical statistical approaches to identify OOT stability results to detect when a batch is not behaving as expected. To judge whether a particular result is OOT, one must first decide what is expected and in particular what data comparisons are appropriate. There are two general approaches to identify OOT data. The first is to look within a batch to determine whether that batch is following the same trend across time as indicated by data for earlier time points (see Figure 1). Because observations from several time points are needed to establish a trend, this question can only be answered adequately for later time points. The second is to look across historical stability batches to determine whether the batch under study is following the same trend as other batches of the same product (see Figure 2). The statistical approaches for these two situations differ. Currently, no common agreement exists about which of these situations should be considered or whether both are equally important.

When an “odd-looking” stability pattern occurs, it is common to ask whether the pattern reflects an underlying mechanism (i.e., a “cause”) or is merely a normal process or analytical variation. The authors will not attempt to answer that question but will instead describe some approaches that may be applied to provide guidance on the judgment whether or not a suspect result is OOT from an objective point of view. If a data point is OOT, the nature of the result should determine which steps to take to determine the presence of an underlying cause and if present, the consequence that it has for the batch under consideration. For example, a natural first step in the investigation would be to verify the initial finding by appropriate additional testing.

The procedures described below for detecting OOT results can be viewed as an alarm or alert system, showing that some kind of action is needed. In other words, at each stability time point when a new result is collected, one should determine whether the result is in agreement with what is expected and if not, take the appropriate action. In many instances, the specification limit alone is used as a tool to identify OOT results. Specification limits, however, may not be the most sensitive detector of underlying causes. In addition, specification limits are applied to individual results, whereas trending could involve a series of results or results from a series of batches.

Another important goal for an alarm system would be to answer the question “Do the data obtained so far indicate that the

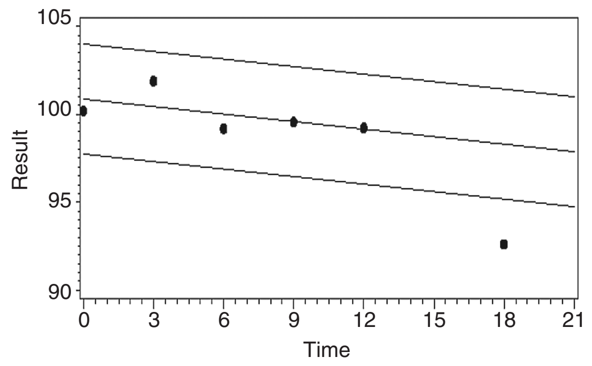

Figure 3: OOT trend within a batch, regression control-chart method. Regression results for example batch: intercept = 100.9, slope = −0.14, s = 1.05, and k = 3.0.

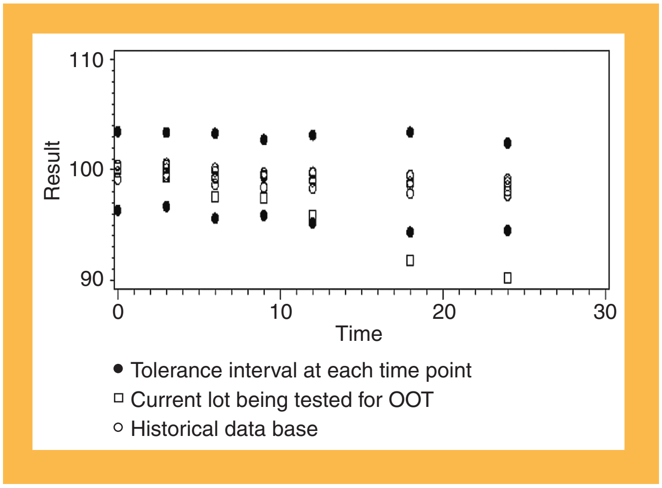

Figure 4: OOT trend across batches by time method, tolerance intervals.

batch will go outside specification during its shelf life?" If so, one may want to introduce a particular type of alarm procedure suitable for this aim. However, this question will not be discussed further in this article.

For all alarm procedures it is important that the procedure provide an alarm when it should but also that the number of false alarms is minimized. Often a compromise is needed to balance the risk of false alarm against the risk of alarm failure: A procedure that is almost certain to provide an alarm when a true OOT result occurs will have a larger risk of false alarms than a procedure that is more selective in giving a signal. It is therefore important to have a suitable balance between these two goals.

The choice of suitable statistical methodology depends on the type of parameter (e.g., property or measurement) under study. For this reason, the two main situations described previously are discussed separately for the two most common types of parameters: single reported values and variability of multiple results. Parameters that need special consideration and will be discussed separately include degradation products and impurities. For these parameters, available data often have low information content as a result of being excessively rounded or truncated when results are less than the limit of quantification or the International Conference on Harmonization (ICH) reporting threshold and/or reported with low precision. Similar issues arise with leachables.

Table I: OOT trend across batches by time method, tolerance intervals.

	Tolerance limit		
Time	Lower	Upper	Results
0	96.4	103.5	100.2
3	96.7	103.4	99.4
6	95.6	103.3	97.6
9	95.9	102.8	97.4
12	95.2	103.2	95.9
18	94.4	103.5	91.8[a]
24	94.5	102.5	90.2[a]

[a] indicates OOT

This article does not attempt to make an exhaustive list of all approaches but briefly presents some possible alternatives. Several other suitable approaches are available and alternative procedures may be more appropriate in special circumstances. Also note that in many situations, slight modifications to the general methods presented may improve the efficiency of the procedure.

Review of current and common approaches. The authors are not aware of an established statistical procedure that is widely used to identify OOT results. However, there are several simple rules of thumb that are sometimes used, and some of these techniques are provided in this section. Various approaches have been used historically for the identification of OOT results, including the following:

- Three consecutive results are outside some limit.
- The difference between consecutive results is outside of half the difference between the prior result and the specification.
- The result is outside ± 5% of initial result.
- The result is outside ± 3% of previous result.
- The result is outside ± 5% of the mean of all previous results.

The advantages of these approaches are that they are easily implemented, easily understood, and usually do not require different limits for each time point. However, the major disadvantage is that these approaches do not have a statistical basis, which makes their performance properties vary depending upon the variability of the data in a given situation. This means that for parameters with high variability, finding a false-positive result will be more likely, but OOT results may be missed for parameters with low variability. In addition, some of these approaches compare the current result to only one other result. If the comparator result is inaccurate (either high or low) purely by chance, the comparison may not accurately reflect whether the current result is OOT or not. Furthermore, if two results are at odds with each other, how can one judge which of them is OOT without the use of additional information?

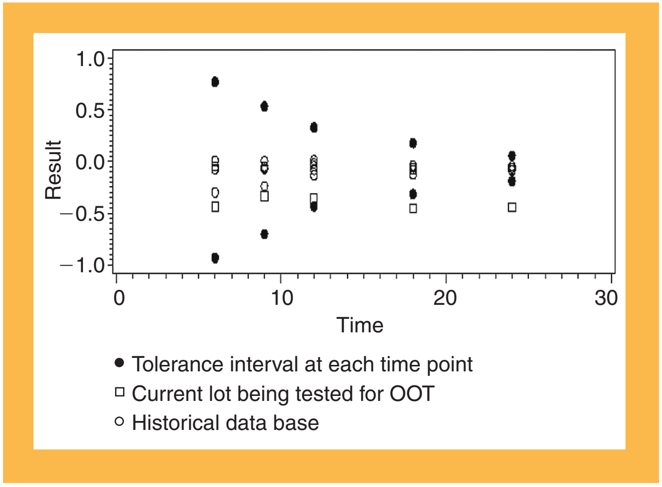

Figure 5: OOT trend across batches, slope control chart method.

To address the disadvantage of not allowing for differences in the variability of the data, one could review past data and calculate the distribution of differences between consecutive results (on some standardized scale such as change per month) for each data set of interest. The allowable difference between consecutive results could be defined as either a stated percentile from the observed distribution or by generating a prediction interval on the basis of the observed distribution.

Note that all the above approaches focus on the issue of detecting OOT results within a given stability batch.

Review of data issues

As previously stated, three types of data can be collected during stability studies. The first type is data reported as a single result such as potency, assay, pH, average dissolution, and average delivered dose. The second type is data with multiple results such as dissolution testing and delivered dose uniformity for inhalation products in which the variability parameter of the multiple results is the end point of interest. Several statistics can be used to quantify the data variability for a time point, including, for example, the relative standard deviation (RSD). The third type is degradation product and impurities data. Degradation product and impurities data also are reported as a single result; however, they are discussed separately because of the nature of the collection and the reporting of data.

For all three types of data, to determine the criteria for OOT at a specific stability time point, one must make assumptions about the underlying distribution of the data. Once a distribution is assumed, then an OOT observation would be an observation that is at an extreme in the distribution.

The following statistical procedures formally require data to follow a normal distribution: The first type of data usually is expected to be approximately normal, and variability-type data typically have a skewed distribution. If the data do not meet the requirement for normality, one solution is to transform the data, for example, by a log or square-root transformation. Provided the transform normalizes the data, limits can be calculated on the basis of the transformed data and finally converted back to the original scale. For data of the second type, a good normalizing transformation may be difficult to find, so the resulting OOT procedures are only approximate. In these cases, it is important to study the properties of the procedure carefully before implementing it.

Identifying an OOT signal for a single result or for variability parameters

These approaches all are based on statistical approaches that take the variability of the data into account when setting limits. One advantage of these approaches is that as long as the assumptions are met, the rate of false positives can be set when one calculates the limits. However, a disadvantage is that a historical database is needed to set up the limits. For products with limited data, the appropriate limits may be difficult to determine. This can lead to wrongly centered, too narrow, or too wide OOT limits. One possible alternative approach when limited data are available is to assume that there is no change over time and regard a significant change as an OOT result. Furthermore, undetected atypical data in the historical database can cause limits to be too wide to identify an atypical result.

Three approaches are described for identifying an OOT stability result: the regression control chart method, the by time point method, and the slope control chart method. The first method is appropriate for both comparisons within batch and comparisons with other batches. The second and third approaches are appropriate only for comparisons with other batches.

Regression control chart method. The first approach is to calculate a regression control chart from data within a batch or data among batches. The control-chart limits bracket the regression line along the length of the stability study. This method requires one to assume data are normally and independently distributed with a constant variability across all time points. A common linear slope for all batches also is required for this method.

A least-squares regression-line is fit to the data. For comparisons within a batch, a regression-line is fit to the data for that batch. For comparisons among batches, a regression line is fit to the historical database for the product assuming a common slope but allowing for various batch intercepts. This fit will provide an estimate of the intercepts, the slope, and the square root of the mean square error. Alternatively, a common slope estimate and standard error from the regression from historical records can be used. An estimate of the expected result at any given time point for a given batch is specified by the following relation:

$$\text{expected result} = \text{intercept} + (\text{slope} \times \text{time})$$

To find the control limits at a given time point, calculate the expected result ± $(k \times s)$, in which k is a multiplier chosen from a table of normal quantiles to give the desired protection

Table II: OOT trend across batches: slope control chart.

Time	Slope tolerance limit Lower	Slope tolerance limit Upper	Slope Estimates
6	−0.93	0.78	−0.43
9	−0.70	0.54	−0.34
12	−0.43	0.33	−0.35
18	−0.31	0.18	−0.45[a]
24	−0.19	0.06	−0.44[a]

[a] indicates OOT

level, and s is the square root of the mean square error from the regression.

Stability data points within the control limits at a given time point are in control and would not be considered OOT. Stability data points outside the control limits at a given time point would be considered OOT and should be investigated further.

Based on the multiplier (k) chosen for the control chart, it is possible to control the confidence level and thus the rate of false alarms. With a basic statistical package, these computations are straightforward.

A better, but more complex, approach would be to use a prediction interval (i.e., an interval that contains the future observation with a certain confidence) or tolerance interval (i.e., an interval that contains a certain percentage of future observations with a given confidence) because these intervals reflect the number of values going into the estimate, the variation in the data, and the amount of extrapolation being performed (4). This additional complexity may be worthwhile depending upon the specific situation. Note that for tolerance intervals with small sample sizes the intervals will be wide and may not be sufficiently discriminatory. If data are not tested at the standard test times of 0, 3, 6, 9, 12 . . . months, limits for the actual test time could be calculated.

Data from Figure 1 were used to demonstrate this method. Data for a 12-month time period were used to set the limits for the regression control chart. Results are shown in Figure 3. The 18-month time point is OOT because it is outside of the calculated regression control limits.

By time point method. The by-time-point approach is used to determine whether a result is within expectations on the basis of experiences from other batches measured at the same stability time point. This method assumes a normal distribution and that all observations at a given time point are independent.

In this approach, historic data are used to compute a tolerance interval for each stability time point. The tolerance interval can be based on the stability results themselves or on the difference from the initial stability result for the lot to minimize the effect of time zero differences among lots. To calculate a tolerance interval, one must calculate the mean $\overline{x}$ and standard deviation (s) for each time point. From tables or using approximations, a multiplier k can be found. The interval can be calculated as $\overline{x} \pm ks$. The width of the interval primarily depends on the number of batches in the historic database and on the choices of confidence and coverage desired. If the current result is outside these limits, the result is considered OOT.

The advantages of this method are that no assumptions about the shape of the degradation curve are needed and it can be used when variability differs for various time points. Also, the level of confidence and the coverage can be chosen to meet the needs for the particular product.

One challenge with this approach arises when current data are not tested at the nominal time points. In that case, limits calculated for the previous or following nominal time point may be used as an approximation. The suitability of this approximation and the choice between limits for previous or following time point depend on the rate of change.

Data from Figure 2 were used to demonstrate this method. Five lots with similar slopes were used as the historical data (see Table I and Figure 4).

Slope control chart method. A third approach for detecting OOT results for single results is to construct a control chart for the slope at each time point. This method is useful for comparison among batches. For each time point, a least squares regression is fit that includes all data up to that time point. The slope estimate for each batch is used to find an overall slope estimate and control limits. Because the slope is normally distributed, OOT limits for the slopes at each time point are obtained from the tolerance interval, in which k is chosen to obtain the desired coverage and $\overline{x}$ and s denote the mean and standard deviation of the historical slope estimates.

The advantage of this method is that slopes are compared so that one can determine whether all batches behave the same. The limits are wider at earlier time points because the slope estimate is more variable when fewer data points are included in the regression. A disadvantage is that to determine whether a data point is OOT, a slope calculation must be performed. Often this calculation is not routinely done after each time point, and the responsible analyst may not have previous data easily available. If data are not tested at the nominal time points, the limits may not be appropriate. Slight differences between the actual test age and the nominal age, however, should not have large effects.

Data from Figure 2 were used to demonstrate this method. Five lots with similar slopes were used as the historical data (see Table II and Figure 5).

Degradation products and impurities

Stability batches are assayed so that one can measure degradation product and impurity levels. The unit of measurement usually is percent area unless a standard curve is used, in which case percent is the typical unit of measurement. To formally determine what is expected requires knowledge about the shape of the underlying trend and the distribution of the results at each stability time point.

When determining potency, one usually assumes that the trend is either linear or can be linearized by transformation of the observed results and/or the time scale. The variance of results at each stability time point also is assumed to be constant. Neither of these assumptions may hold for degradation products or impurities; for example, the variability often increases with time when the level of the degradant increases.

The assay method for degradation products and impurities specifies a limit of quantification (LOQ). Most analytical laboratories will not quantify the result if it falls below the LOQ. The value usually is reported as "< LOQ." In such cases, all that is known about the result is that it is between zero and LOQ. The International Conference on Harmonization (ICH) reporting threshold would have similar implications. The remainder of the discussion in this article uses the LOQ value as the truncation threshold, although one recognizes that similar issues may arise as a result of the ICH reporting threshold as well. By truncating the data, laboratories add variability and lose valuable information.

A special situation arises when a new peak forms during the analysis. When a new peak forms during a stability study, one may expect that it should not exist and hence it would constitute a type of OOT. As discussed previously, a new data point can be compared with previous results from the same batch or with data from other batches. Each of these situations is described in the following paragraphs.

Comparison of a new value to previous values from the same batch. If the degradant or impurity values all are above the LOQ value with a linear relationship over time and the assumption of normality is reasonable, then the techniques used for single results can be applied.

If some of the results are below the LOQ value, if the assumption of normality is not reasonable, or if linearity cannot be assumed, then an attempt to identify OOT results using data from the same batch is not recommended for the following reason: A new data point is OOT when it deviates from what is expected. However, for a given batch, if any of the above situations apply, the expected result usually is not possible to determine. For example, suppose that the data for a batch are below the LOQ value at initial, 3, and 6 months but above LOQ at 9 months. The result at 9 months may be different from the previous results but not OOT because there may have been an underlying increasing trend between initial and 9 months that is first detected above LOQ at 9 months. Therefore, if some results are below LOQ, then a comparison of the new value to values from other batches, described in the following section, is recommended.

Comparison of a new value to values from other batches. Data from other batches can be useful for the identification of an OOT degradation or impurity observation. Three possibilities exist for data obtained from previous batches: all values are above LOQ, all values are below LOQ, and portions of the data are below LOQ.

All values are above LOQ. If all of the data are above LOQ, then the by time point method usually is applicable. To compute tolerance intervals, one assumes normality. However, the distribution of a degradant or impurity may be skewed. One solution is to transform the data by taking the log or square root. After the transformation, the tolerance interval is computed and transformed back to the original scale. If the number of data points is small, then the tolerance interval may be too wide. In such case, a compromise, such as a lower confidence level, may be necessary. The regression control chart and slope control chart methods also can be used in situations with linear (or linearizable) trend and constant variance.

All values are below LOQ. When all of the data are below LOQ, the easiest OOT criteria may be to use the LOQ value. Any result above LOQ is an OOT result. This method may be too conservative if the sample size is small. If seven or more previous results are below LOQ, then a new result above LOQ may be OOT; however, if only two or three previous results are available and below LOQ, a new result above LOQ may not be that unusual.

A portion of the data are below LOQ. If some of the values are below LOQ and some are above, then one must decide what to do with the values less than LOQ. One strategy would be to set all of these values equal to either LOQ, LOQ/2, or zero before calculating the tolerance interval. Several statistical techniques can be used to estimate the mean and standard deviation of a normal distribution when some of the observations are censored (5,6). Using these estimates, one could use the mean plus some multiple of the standard deviation (e.g., 3*s*) to establish OOT criteria.

Additional approaches

The approaches presented are not exhaustive. Many other valid statistical techniques are available. For example, one could develop a diagnostic method using the residuals from a fitted model that is not necessarily a linear model. Another example would be nonparametric approaches such as rank tests. This is an area with many opportunities for further research and discussion.

Implementation challenges

The purpose of developing a criterion for OOT assessments is to identify the quantitative analytical results during a stability study that are atypical enough to warrant a follow-up investigation. Numerous challenges exist that a company must overcome to implement an OOT procedure for commercial stability batches.

Identifying an unusual result is more difficult in a stability setting than in batch-release testing. Stability studies are run less frequently. Once the registration and validation batches are complete, a single batch may be placed on stability each year. Unlike batch-release results, which represent one point in time for a batch, stability results may change over the shelf life of the batch. With adequate experience, an analyst can identify a result that is not typical; however, it takes a long time for an analyst to accumulate experience with a product and its distinct properties. This method also assumes that only a few analysts run the applicable approaches and will remain in the laboratory to perform them over an extended period of time. To further complicate matters, some companies hire third-party contract organizations to evaluate stability data performance.

Current computer systems also can present a challenge. Computer systems typically are designed to treat the stability time point as an independent event. At given time point *T,* samples are pulled and sent to laboratories. The analyst runs a specific test, and that result is entered into the laboratory data system and compared with a single specification. Previous results for that property from that stability study usually are not easily available to visually evaluate the trend over time. Moreover, past results for that property and product obtained at the same time point are not readily available. Thus, for many current computer systems, the historical data needed to visually identify OOT results are not available to the laboratory analyst because the result is being generated.

Before implementing an OOT procedure, one must decide the type of OOT of interest. The approach for identifying an OOT depends on this definition. As discussed previously, two main types of OOT definitions exist: a result is OOT if it is at odds with previous test results for that batch (comparison within batch) or if the result is not similar to the results that past stability studies generated at that same time point (comparison with other batches). There appears to be no common agreement on which of these situations should be considered or if both are equally important. Once a definition is agreed upon, a decision procedure to identify OOT results must be determined. In most situations, it is appropriate to use some of the statistically based approaches briefly outlined above. However, several additional topics must be considered before the detailed decision procedure is developed:

- Is the OOT procedure intended for NDA stability studies, for commercial stability studies, or for both?
- What is the minimum amount of data required for computation and to obtain reasonable estimates?
- What does the change over time look like (linear, nonlinear, etc.)? If data are nonlinear, what adjustments must be made to the analysis?
- Properties may have various stability profiles and distributions, thus possibly requiring different approaches to determine an OOT. A one-size-fits-all approach for all analytical properties may be inappropriate.
- What is the analytical method precision? For properties with results close to zero, these results may have a chopped-off distribution (censored data) because some of the actual values are too small to be adequately measured and reported by the analytical method.
- For impurities, there may be an impact of ICH reporting thresholds. If the true amount of a degradant or impurity is near the threshold, it may appear or disappear from time point to time point, or it may not show up at first then appear at later time points.
- For degradation products and impurities, data often are rounded to one digit past the decimal point. This practice limits the OOT investigation tools and the power of those tools. Degradation products or impurities should be provided to at least two or three digits past the decimal for a statistical evaluation of the data to be conducted. If OOT identification is considered desirable for degradation products and impurities, rounding practices must be changed.
- What is the effect of container–closure on the stability profile (e.g., if tablets are stored in blisters and bottles, is the same change expected over time)?
- Are data from different stability studies independent? When studies are assayed together the results may be dependent and not encompass all sources of analytical variability that will be seen long term. An example is the testing of validation and registration batches, which often are grouped when analytical testing is performed.
- Is the OOT criterion updated after each new data point is collected or revised according to some review schedule (e.g., yearly), or is it a fixed limit?
- The integrity of the data used is critical. If atypical data are included in a data set used to establish the OOT criterion, then the criterion may not be sensitive enough to pick up future OOT results. Therefore, if a statistically based method for identifying an OOT result is used, historical data must be carefully examined for each property of each product.
- As more history is available for a product, time intervals may change for the stability protocol.
- The scope of setting up limits for each product at each stability time point can be overwhelming because a multitude of dosages and pack sizes exist within a company. For example, 20 products with three packages per product, seven tests performed per stability time point, and a typical stability study with seven stability time points could require calculation and maintenance of as many as 2940 sets of limits.
- When looking for OOT results within a batch, a previous time point may turn OOT in light of later measurements (despite being within expectations at the time of collection). Should OOT analysis look back before the current time point to recognize previously unidentified potential OOT results? (During any inspection, it could be evident that a "once within-trend" result has turned OOT).
- The problem of multiplicity in testing raises questions about adjustment in the p-value level for individual comparisons. The adjustment may be necessary because statistical evaluation of the same parameter across repeated time points on the same lot or the testing of numerous parameters at the same time point raises the likelihood of finding significant results in one or more tests just as a result of random chance. An adjustment for multiplicity is recommended, keeping in mind that any adjustment also decreases the likelihood of finding significance for any individual test.
- Previous stability data may not be available on-line for some data systems (e.g., third-party contract organizations), which means that data must be reentered into a new system as it is collected. This process makes routine OOT analysis difficult.

The method to determine the OOT criterion ideally would not be too complex, yet something too simplistic and arbitrary (for example, a maximum *X*% change rule for all properties) may not be sensitive enough to identify a true OOT or may give a high rate of false signals. Thus, any method considered for identifying an OOT should be challenged using real data (and simulation, when appropriate) to "verify" the effectiveness. Once the method is selected, if the analysis is to be performed by a computer, then the process must be integrated into a validated laboratory information management system. Data extraction and the code used to compute the criterion must be validated per 21 *CFR* Part 11 (7), which currently is a labor-intensive issue. Given recent GMP inspection trends, it may not be worth the effort given the potential computer system inspection risks.

The process for determining an OOT stability result should be documented in a standard operating procedure (SOP) and should involve the following:

- What statistical approaches are used to determine OOT criterion? What data are used to determine OOT limits?
- What are the minimum data requirements? What evaluation is performed if the minimum data requirement is not met?
- What data should be used to update limits?
- The investigation requirements (i.e., who is responsible, what is the timeline, how is it documented, who should be notified) must be clearly defined.
- Who is responsible for comparing the result with the OOT criterion? To be effective, the identification of an OOT would occur immediately in the laboratory environment to promptly initiate the appropriate investigations needed.
- How is an OOT result confirmed? What additional analytical testing or statistical analyses are appropriate?
- What actions should be taken if an OOT result is confirmed as an unusual result? How may those actions differ depend-

ing upon the result and the corresponding specification? For example, is an atypical potency result with a two-sided specification treated the same as an atypically low impurity result or would an atypically low RSD result need to be investigated further?

- How are OOT investigations incorporated into the annual product review?

Conclusion

Identifying OOT stability results is a growing concern for FDA and the pharmaceutical industry. Ideally, the method to determine an OOT alarm should not be too complex. However, something too simplistic (for example, a maximum *X*% change rule for all properties) may not be sensitive enough to identify a true OOT or may give a high rate of false signals. The procedure should be chosen to best suit the parameter that is measured. This article outlines various approaches, including methods to detect an atypical single result or atypical variation. The large number of tests and time points requiring OOT limits can make OOT detection a complex problem. For degradation products and impurities, it is difficult to identify OOT unless the data reporting routines are changed. The within-batch methods are more difficult to implement than the between-batch methods because of the sparse data within a batch, especially at early time points. There appears to be no common agreement about which of these situations should be considered or whether both are equally important. Many issues, technical and practical, are not easily resolved and further research and discussion is warranted among all parties involved.

Acknowledgments

The following members of the PhRMA CMC Statistics and Stability Expert Teams assisted in developing this article: Mark Alasandro, Merck & Company, Inc; James Bergum, Bristol Myers Squibb Company; Laura Foust, Eli Lilly and Company; Mary Ann Gorko, AstraZeneca Pharmaceuticals; Jeffrey Hofer, Eli Lilly and Company; Charles Ireland, Sanofi-Synthelabo; Dennis Sandell, AstraZeneca AB; Edith Senderak, Merck & Company, Inc; Andrew Sopirak, AstraZeneca Pharmaceuticals.

References

1. A.M. Hoinowski et al., "Investigation of Out-of-Specification Results," *Pharm. Technol.* **26** (1), 40–50 (2002).
2. FDA Guidance Document, Investigating Out of Specification (OOS) Test Results for Pharmaceutical Production (draft).
3. *US v Barr Laboratories,* 812 F. Supp. 458 (District Court of New Jersey 1993)
4. G. Hahn and W. Meeker, *Statistical Intervals: A Guide for Practitioners* (John Wiley & Sons, New York, 1991).
5. H. Schneider, *Truncated and Censored Samples from Normal Populations* (Marcel Dekker, New York, 1986).
6. J.F. Lawless, *Statistical Models and Methods for Lifetime Data* (John Wiley & Sons, New York, 1982).
7. FDA, 21 *CFR* Part 11: Electronic Records, Electronic Signatures. **PT**