Available online at www.sciencedirect.com
ScienceDirect
Journal of Pharmaceutical and Biomedical Analysis 47 (2008) 413–421
JOURNAL OF PHARMACEUTICAL AND BIOMEDICAL ANALYSIS
www.elsevier.com/locate/jpba

Short communication

A six-sigma approach to stability testing

Karl De Vore *

Bio-Rad Laboratories, 9500 Jeronimo Road, Irvine, CA 92688, United States

Received 20 June 2007; received in revised form 20 December 2007; accepted 23 December 2007
Available online 4 January 2008

Abstract

The following defines stability testing in the diagnostic and pharmaceutical industries as a process which, depending on the manufacturer's current approach, may contain many opportunities for improvement. Statistical thinking and six sigma concepts will enhance stability testing process capability and lead to higher confidence in the data. Tools for set up and the rationale behind them are provided to assist in establishing appropriate criteria and volume of testing.

Keywords: Six sigma; Stability testing set up; Failure criteria; Test method precision; Testing volume

1. Introduction

Stability testing is certainly one of the most important processes in the manufacture of diagnostic quality control materials and reagents. Although some of the specific requirements may be different than those of the pharmaceutical industry, there is sufficient overlap in the challenges of performing a successful stability evaluation that overcoming the obstacles common to both disciplines can be covered in the same treatment.

The discipline of six sigma views every business activity as a process, that once optimized and controlled, reduces cost. Hence, six sigma itself is a process that is often briefly described by the acronym DMAIC, which stands for define, measure, analyze, improve, and control. First, the stability testing process, or process issue, needs to be defined. Second, since stability testing itself is a measuring process, its capability needs to be measured. Third, the capability of the process needs to be analyzed in order to determine if it is delivering what is required (accurate stability predictions or estimates), and if not, improve. Finally, control the stability testing process by insuring that the improvements that have been implemented are maintained through time.

2. *D*MAIC: defining the process

Because the accelerated stability testing process attempts to obtain a prediction of real time performance, error is magnified in the extrapolation of degradation rates from warmer to colder temperatures. Due to scheduling, commitment of resources, and regulations, verifying real time stability, though challenging in its own right, is less technically demanding. However, the real time stability monitoring and accelerated stability testing processes both have a common need of appropriate failure criteria and scope of testing. These two parameters are closely linked. With pharmaceutical and diagnostics stability testing, the failure criteria may be established prior to the choice of test method and testing volume, but with diagnostics control products it need not be. The criteria can, and should be, tied to the customer's ability to detect a difference. For the pharmaceutical industry this may be a regulatory requirement or the point in which the patient receiving the medication would be aversely affected due to the change in the strength of the active ingredient, excipients, or degradation products. In six-sigma language, this is referred to as the "voice of the customer", while the manufacturer's testing process capability is the "voice of the process". The goal of six sigma is to make the voice of the process more powerful than the voice of the customer, leading to reduced cost of operations and superior product quality.

Statistics is usually appreciated or applied during the data analysis stage only, squandering much of its potential. Set-up is actually more important in obtaining high quality data. How-

* Tel.: +1 949 598 1317; fax: +1 949 598 1553.
E-mail address: karl_devore@bio-rad.com.

0731-7085/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2007.12.038

ever, a review of the literature on the subject revealed a dearth of direct guidance or suggestions on the how to determine the volume of testing for a successful accelerated stability evaluation. The appropriate number of time points, placement of time points, and replicates per time point, if available were likely buried deep in the text and therefore difficult to find [1–8]. The following therefore addresses directly these important considerations based on the failure criteria and test method precision, and provides tools for set up so that the process of stability testing will provide the best quality data.

2.1. General considerations

Before proceeding, some general considerations will be briefly reviewed:

(1) To eliminate the effects of inter-assay variation, all testing involving comparisons of time series data should be performed in the same assay run. This is accomplished by initiating stress at appropriate times so that they completed simultaneously.
(2) Test sequences should be randomized, and whenever different, unrelated comparisons are being performed (different formulations, strengths, container sizes, etc.) in the same run, they should be separated as a group within the run. This will insure that comparisons of apposite test results are generated as close together in time as possible.
(3) Test method precision should be established prior to initiating the stability evaluation. For internal methods, this would be best accomplished through gage R&R (repeatability and reproducibility) studies.
(4) Accelerated stability stress should be adequate to result in degradation that significantly larger than the test method precision. However, if equivalent accelerated stress greatly exceeds worst-case real time stress, excessive degradation of the analyte, matrix components, or loss of Oxygen could lead to exotic phenomenon not likely to occur under real storage conditions. Therefore, accelerated study stress should strive to achieve approximately 10–20% degradation.
(5) Accelerated temperatures should be as close to the anticipated storage temperature as possible, but consistent with experimental need.
(6) The use of 95% confidence limits of a regression to establish the worst-case conditions for product claims should only be used for real time or in-use stability estimates, not for Arrhenius predictions. Because Arrhenius predictions use extrapolation of a regression on plots of log rates versus the inverse of the absolute temperature [9–11], a worst-case prediction would put an unacceptable burden on the manufacturer, and make release of new products nearly impossible. In addition, each individual temperature's mean rate should be used in generating Arrhenius predictions, since using the 95% confidence limit's worst case of each temperature's data would lead to a prediction using a combination of three relatively rare events which would be extraordinarily unlikely ($0.05^3 = 0.000125$ or 0.0125%).

2.2. Time points

When establishing the appropriate number of time points required for regression of accelerated stability data (not to be confused with tests per time point or replicates), the failure criteria and the total expected degradation needs to be considered. If the failure criteria is $\geq \pm 10\%$ and it is approximately equal to the expected degradation, then the number of time points required would certainly be less than 6, which appears to be the default minimum number used by many. In fact, it has been shown that the number of time points can be reduced to two [12]. Therefore, the reason that six points is typically chosen is likely the result of researcher's common use of six point calibration curves for the test methods. However, because degradation does not follow the same path as a dose response curve, it need not be treated the same.

For calibration, six points is a reasonable choice that can capture four inflection points of a dose response curve [1]. With degradation, especially in the first important 20%, curvature is difficult to resolve with typical test method precisions, even if the kinetics of degradation is second or third order [12]. Still, even when these arguments are seriously considered, there is strong resistance to using any number less than six. Therefore, in determining the number of total tests (the product of points and replicates per point) six time points will be covered as well as two points. Regardless of the total number of time points used for each temperature, if they are equally spaced, a halfway point should be avoided. This is because no matter what recovery value is obtained, a linear regression slope, and therefore rate estimate will not be affected. To clarify; the slope of the regression line is defined as

$$b = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$$

If the time point (x_i) is at the halfway point, then the squared difference will equal zero and it will have no impact on the slope of the regression line and hence the rate estimate. Note, this will not be the case with unequally spaced time points, where the mean of the points will not equal the median.

2.3. Number of temperatures

When combining the data from each temperature for Arrhenius plots, three temperatures and the analyte's respective degradation rate is generally accepted as the minimum: two points to determine the line, with one point to confirm. Although additional temperature data will improve the precision of the estimate, the cost can be prohibitive. Each temperature should be as close to the real time storage temperature as possible while providing data in a time frame consistent with experimental need. There are two reasons for this: one is that different mechanisms of degradation are more likely to be involved the further the temperature of the study is from the intended real time storage temperature and perhaps more importantly, the Arrhenius extrapolation error increases the further the accelerated temperature's rate data are from the intended real time storage temperature.

2.4. Replicates

How much degradation and the failure criteria will be key to determining the replicates per time point required. For the purposes of discussion, ≥±10% failure criteria will be assumed. This is a typical criteria used in the diagnostic industry, and is generally accepted as standard. However, the calculations can be performed using any failure criteria. For instance, with diagnostic controls, as the test method imprecision increases, a 10% change becomes more difficult to detect with an acceptable level of confidence. It also means that the both the customer requirements (voice of the customer) and stability testing process performance (voice of the process) may change. As the method imprecision increases, a 10% change in analyte concentration will have less impact on product performance, and the failure criteria specifications may need to be modified. With quality controls a >8% total method CV is the point in which a modification of the criteria would be appropriate. This will be discussed in detail later.

As with the estimate of the number of time points required to determine a statistically significant difference between two samples sets, the number of replicates (Y) per time point (X) required is relatively straight forward. However, it does require us to make the following assumptions:

(1) The linear model is correct.
   (a) In reality, the linear model is usually incorrect, but is sufficient to describe higher order decay within the first 20% of analyte degradation [12].
(2) For ease of calculations, the standard deviation of the Y data replicates will be lower by ∼4% at T_0 (the first time point or reference sample), and higher by 6% at T_{final} (the last time point of the study).
   (a) As the concentration of the analyte decreases the magnitude of a set standard deviation relative to the concentration increases. How this is accounted for in the calculations will become clearer in the following discussion.
(3) The inter-vial variance will be insignificant relative to the method imprecision.
   (a) Although theoretically the vial-to-vial variance should increase with time, the magnitude of this potential increase will be unknown.
(4) During the stability study, the amount of degradation we need to detect is at least 10%. This number can be varied depending on the customer requirements.
   (a) Implicit here is that the failure should be reached by T_{final}.

3. D*M*AIC: measure the process

Determining process capability, or measuring the measuring process, is frequently overlooked when establishing a stability-testing program. Typically, inconclusive results of suspected unstable analytes are indicative of an incapable stability testing process. The capability can be roughly predicted before testing even begins if the average standard error of the method is greater than one-third the criteria. If so, then more replication will be required. At our facility computer simulations were helpful at gaining insight into process capability. How replication and time point place placement impacted the process was demonstrated. Why, was latter determined after extensive statistical analysis.

4. DM*A*IC: analyze

4.1. Hypothetical example

Suppose that during a stability study using six time points, the degradation is 10%, i.e., at T_{final}, the recovery is 90% of the T_0 (beginning time point of the study) value. In addition, suppose that the regression is a perfect fit through these points (Table 1).

In this instance, the slope (degradation rate) of the line is −2, because the change in concentration is −10 units in 5 units of time (−10/5 = −2). The sum of squares of the Y values is 70, and because it is a perfect fit, it equals the sum of squares due to the regression.

$$\sum (Y - \bar{Y})^2 = \sum (\hat{Y} - \bar{Y})^2 = 70$$

where $\hat{Y}$ is the interpolated regressed value at each X. With regression data, the total sum of squares (left hand term in above equation) will usually be greater than the sum of squares due to the regression because of an additional term: the sub of squares of the residuals.

$$\sum (Y - \bar{Y})^2 = \sum (\hat{Y} - \bar{Y})^2 + \sum (Y - \hat{Y})^2$$

The second term on the right had side is the sum of squares of the residuals. But because the regression is a perfect fit, this term equals 0. Shorthand for the above equation is

$$SS_Y = SS_{reg} + SS_{res}$$

Now, suppose that there are 3Y values for each time point X with a CV of approximately 1%. Then, the sum of squares due to the regression, SS_{reg} will equal 3 × 70 = 210, and the sum of squares due to the residuals = 12 (Table 2).

$$SS_Y = SS_{reg} + SS_{res}$$

$$SS_{res} = SS_Y - SS_{reg} = 222 - 210 = 12$$

You can see that if the linear model is perfect, then given a set number of replicates and time points, the total sum of squares will only increase due to an increase in the sum of squares of the

Table 1
Sum of squares of a hypothetical stability data set

X (time)	Y (conc)	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$
0	100	6.25	25
1	98	2.25	9
2	96	0.25	1
3	94	0.25	1
4	92	2.25	9
5	90	6.25	25
Sum		17.5	70

Table 2
Regression statistics for a hypothetical stability data set

X (time)	Y1	Y2	Y3	$(X-\bar{X})^2$	$(Y_1-\bar{Y})^2$	$(Y_2-\bar{Y})^2$	$(Y_3-\bar{Y})^2$	$(\hat{Y}_1-\bar{Y})^2$	$(\hat{Y}_2-\bar{Y})^2$	$(\hat{Y}_3-\bar{Y})^2$
0	100	99	101	6.25	25	16	36	25	25	25
1	98	97	99	2.25	9	4	16	9	9	9
2	96	95	97	0.25	1	0	4	1	1	1
3	94	93	95	0.25	1	4	0	1	1	1
4	92	91	93	2.25	9	16	4	9	9	9
5	90	89	91	6.25	25	36	16	25	25	25
Sum				17.5	70	76	76	70	70	70
						222[a]			210[b]	

[a] SSy.
[b] SSreg.

residuals. The sum of square of the residuals is directly impacted by the precision of the method. We will come back to the sum of squares calculation later.

4.2. Significance of the regression slope

To estimate degradation rates in a stability study, at a minimum the slope of the regression line must be significantly different from a slope of zero. A slope of zero would indicate no correlation of decay with time. The equation for determining the significance of an apparent slope of −10% is

$$t = \left| \frac{B_{10\%} - B_0}{\text{S.E.}_{\text{slope}}} \right|$$

where $B_{10\%}$ is a linear regression slope of 10% degradation, B_0 a slope of zero, and $\text{S.E.}_{\text{slope}}$ is the standard error of the regression slope. $\text{S.E.}_{\text{slope}}$ is defined as

$$\text{S.E.}_{\text{slope}} = \frac{\sqrt{\left(\sum (Y_i - \hat{Y})\right)/(n-2)}}{\sqrt{\sum (X_i - \bar{X})^2}}$$

The numerator in the above equation, the familiar term SS_{res}, is directly impacted by assay precision and number of replicates, the denominator is impacted by the spacing of the time points.

It has already been shown that for a linear model, two points, a T_0 and T_{final}, with multiple replicates each, has less error than the same number of total replicates spread over six time points [12]. However, if the analyte's degradation profile is unknown, assuming a linear model has some risk if intermediate time points are not tested. A compromise between the two extremes is to test less replicates at the intermediate time points, with a greater number at T_0 and T_{final}. In a six-point study, two replicates each for the four intermediate time points should be sufficient to determine if important deviations from linearity are occurring. The number of replicates required for T_0 and T_{final} would vary depending on the method precision.

4.3. Predicting stability study precision

Once the precision of the test method is determined, through direct in-house testing or test method's insert claims, the amount of variation for each time point tested will, by chance, be within a range of values. This is illustrated through the χ^2 equation and table [13].

$$\chi^2 = \frac{(n-1)\text{S.D.}^2}{\sigma^2}$$

The χ^2 table lists two numbers for each degree of freedom $(n-1)$ and confidence level. If the χ^2-result is between these two numbers, then the hypothesis that the sample set, or more precisely, its standard deviation, comes from a population with a standard deviation equal to σ, cannot be rejected. This equation can be rearranged to predict the range of sample set standard deviations one would expect from a population with a set standard deviation (σ) or CV.

$$\chi^2 = \frac{(n-1)\text{S.D.}^2}{\sigma^2} \rightarrow \text{S.D.} = \sqrt{\frac{\chi^2\sigma^2}{n-1}}$$

Assigning the χ^2-variable its upper end value for 95% confidence, the highest expected CV can be estimated. However, since it is assumed that the model is correct and a perfect fit, then the sum of squares of the residuals will essentially be only contributor to error, and will total the sum of squares error about each time point mean for each set of Y replicates.

$$\sum_{j=1}^{6}\sum_{u=1}^{n}(Y_{ju} - \hat{Y}_j)^2 = \sum_{j=1}^{6}\sum_{u=1}^{n}(Y_{ju} - \bar{Y}_j)^2 - \sum_{j=1}^{6}(\hat{Y}_j - \bar{Y}_j)^2$$

where the first term on the right hand side of the equation is the sum of squares of each Y replicate minus the mean Y value associated with the time point. The second term on the right hand side of the equation is the sum of squares of the predicted Y values of the regression minus the mean Y value associated with each time point. In our hypothetical example, this second term = 0, because with a perfect linear fit, the mean of each time point's replicates equals the regressed point on the regression line.

Because the regression line and each regressed time point $(\hat{Y}_j)$ is determined by the slope and intercept parameters, the degrees of freedom associated with the sum of squares of the residuals is $n-2$, not $n-1$. Therefore, the standard deviation

Table 3
Adjustment of hypothetical stability data set imprecision so that average CV per time point equals 1.0%

Uncorrected

X (time)	Y1	Y2	Y3	Mean	SD	CV	
0	100	99	101	100	1.0	1.00%	Average CV 1.05%
1	98	97	99	98	1.0	1.02%	
2	96	95	97	96	1.0	1.04%	
3	94	93	95	94	1.0	1.06%	
4	92	91	93	92	1.0	1.09%	
5	90	89	91	90	1.0	1.11%	

Corrected for degradation

X (time)	Y1	Y2	Y3	Mean	SD	SD * 0.95	CV	
0	100	99	101	100	1.0	0.96	0.960%	Average CV 1.00%
1	98	97	99	98	1.0	0.96	0.980%	
2	96	95	97	96	1.0	0.96	1.000%	
3	94	93	95	94	1.0	0.96	1.021%	
4	92	91	93	92	1.0	0.96	1.043%	
5	90	89	91	90	1.0	0.96	1.067%	

about the regression line is

$$\sqrt{\frac{\sum_{j=1}^{6}\sum_{u=1}^{n}(Y_{ju}-\hat{Y}_j)^2}{(6\times n)-2}}=\sqrt{\frac{\sum_{j=1}^{6}\sum_{u=1}^{n}(Y_{ju}-\bar{Y}_j)^2}{\text{Total\_reps}-2}}$$

By applying the χ^2 equation, we find that

$$\text{S.D.}=\sqrt{\frac{\chi^2\sigma^2}{\text{Total\_reps}-2}},$$

and the upper χ^2 variable inserted in the above equation is the one corresponding to (Total_reps − 2) degrees of freedom.

4.4. Keeping precision in proportion to concentration

Given a constant standard deviation of the Y (concentration) values through time, it will be a larger proportion, or %CV, as the analyte degrades. To ensure that the model's CV averages a certain percentage throughout the 10% drop in concentration, the calculated CV is multiplied by 0.96. This is illustrated in Table 3.

5. DMA*I*C: improve the process

5.1. The model

By applying these concepts, we can now begin to determine the number of replicates required to resolve a statistically significant difference between a slope of 0 versus a slope of −2 (10%).

First, we restate the equation for standard error of the slope

$$\text{S.E.}_{\text{slope}}=\frac{\sqrt{\left(\sum(Y_i-\hat{Y})/(n-2)\right)}}{\sqrt{\sum(X_i-\bar{X})^2}}$$

Convert to our model based on a worst case χ^2 imprecision of a pre-established method CV and factor for the intended average variation through time.

$$\text{S.E.}_{\text{slope}}=\frac{\sqrt{[(0.96\times\sqrt{CV^2\chi^2})^2]/((reps\times pts)-2)}}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}}$$

We can now control the value of the numerator by increasing the number of replicates. The denominator can also be controlled by increasing the replicates at the extreme values of X (T_0 and T_{final}), while maintaining a constant two replicates for each of the 4 intermediate time points.

By adjusting the number of replicates at a given method precision, the effect on the value of t can be determined.

$$t=\left|\frac{B_{10\%}-B_0}{\text{S.E.}_{\text{slope}}}\right|$$

Three t tables were generated using the above model. In Table 4, only T_0 and T_{final} points are included in the regression. In Table 5 the number of replicates was varied by the same amount for each of six time points. In Table 6, the four intermediate time points were held constant at two replicates, while the T_0 and T_{final} were varied.

The resulting t values were compared to the critical t values for significance at the 90, 95 and 99% confidence levels. If the resulting t values are below that required for 90% confidence, they are not highlighted, if they are significant – at or above one of the three levels of confidence – they are highlighted in light gray for 90% confidence, dark gray for 95% confidence, and specked for 99% confidence.

5.2. Establishing appropriate criteria for quality control materials

For most analytes in diagnostic controls the failure criteria of ≥±10% is applied. The 10% criteria does have precedence in the IVD industry, and modifying it for test methods that have acceptable CVs is not usually done, but the criteria is frequently increased when the precision exceeds acceptable levels. "Acceptable levels" is a subjective term, but CVs > 8% are usually considered unacceptable. Though adding the method CV

Table 4
Replicates for two point stability studies: locate the test method precision (CV) in the far left column of the table, then move across the row until a shaded column is reached

Replicates Required for 2 Point Regression of Stability Data or Total Replicates Required for Comparison ot Two Sample Means): Failure Criteria =±10%												
Total Reps	4	6	8	10	12	14	16	18	20	22	24	26
Reps / point	2	3	4	5	6	7	8	9	10	11	12	13
df reg	2	4	6	8	10	12	14	16	18	20	22	24
Chi Square	7.38	11.14	14.45	17.53	20.48	23.34	26.12	28.85	31.53	34.17	36.78	39.36
Critical t (99%)	**9.925**	**4.604**	**3.707**	**3.355**	**3.169**	**3.055**	**2.977**	**2.921**	**2.898**	**2.845**	**2.819**	**2.797**
Critical t (95%)	4.303	2.776	2.447	2.306	2.228	2.179	2.145	2.12	2.101	2.086	2.074	2.064
Cricat t (90%)	2.92	2.132	1.943	1.86	1.812	1.782	1.761	1.746	1.734	1.725	1.717	1.711
CV 1.0	5.423	**7.645**	**9.493**	**11.126**	**12.607**	**13.973**	**15.252**	**16.456**	**17.599**	**18.690**	**19.734**	**20.738**
1.5	3.615	**5.096**	**6.328**	**7.418**	**8.405**	**9.316**	**10.168**	**10.971**	**11.733**	**12.460**	**13.156**	**13.825**
2.0	2.711	3.822	**4.746**	**5.563**	**6.304**	**6.987**	**7.626**	**8.228**	**8.800**	**9.345**	**9.867**	**10.369**
2.5	2.169	3.058	**3.797**	**4.451**	**5.043**	**5.589**	**6.101**	**6.582**	**7.040**	**7.476**	**7.894**	**8.295**
3.0	1.808	2.548	3.164	**3.709**	**4.202**	**4.658**	**5.084**	**5.485**	**5.866**	**6.230**	**6.578**	**6.913**
3.5	1.549	2.184	2.712	3.179	**3.602**	**3.992**	**4.358**	**4.702**	**5.028**	**5.340**	**5.638**	**5.925**
4.0	1.356	1.911	2.373	2.782	3.152	**3.493**	**3.813**	**4.114**	**4.400**	**4.672**	**4.933**	**5.184**
4.5	1.205	1.699	2.109	2.473	2.802	**3.105**	**3.389**	**3.657**	**3.911**	**4.153**	**4.385**	**4.608**
5.0	1.085	1.529	1.899	2.225	2.521	2.795	**3.050**	**3.291**	**3.520**	**3.738**	**3.947**	**4.148**
5.5	0.986	1.390	1.726	2.023	2.292	2.541	2.773	**2.992**	**3.200**	**3.398**	**3.588**	**3.771**
6.0	0.904	1.274	1.582	1.854	2.101	2.329	2.542	2.743	**2.933**	**3.115**	**3.289**	**3.456**
6.5	0.834	1.176	1.460	1.712	1.940	2.150	2.347	2.532	2.708	**2.875**	**3.036**	**3.190**
7.0	0.775	1.092	1.356	1.589	1.801	1.996	2.179	2.351	2.514	2.670	**2.819**	**2.963**
7.5	0.723	1.019	1.266	1.484	1.681	1.863	2.034	2.194	2.347	2.492	2.631	2.765
8.0	0.678	0.956	1.187	1.391	1.576	1.747	1.907	2.057	2.200	2.336	2.467	2.592
8.5	0.638	0.899	1.117	1.309	1.483	1.644	1.794	1.936	2.070	2.199	2.322	2.440
9.0	0.603	0.849	1.055	1.236	1.401	1.553	1.695	1.828	1.955	2.077	2.193	2.304
9.5	0.571	0.805	0.999	1.171	1.327	1.471	1.606	1.732	1.853	1.967	2.077	2.183
10.0	0.542	0.764	0.949	1.113	1.261	1.397	1.525	1.646	1.760	1.869	1.973	2.074
10.5	0.516	0.728	0.904	1.060	1.201	1.331	1.453	1.567	1.676	1.780	1.879	1.975
11.0	0.493	0.695	0.863	1.011	1.146	1.270	1.387	1.496	1.600	1.699	1.794	1.885
11.5	0.472	0.665	0.825	0.968	1.096	1.215	1.326	1.431	1.530	1.625	1.716	1.803
12.0	0.452	0.637	0.791	0.927	1.051	1.164	1.271	1.371	1.467	1.557	1.644	1.728
12.5	0.434	0.612	0.759	0.890	1.009	1.118	1.220	1.316	1.408	1.495	1.579	1.659
13.0	0.417	0.588	0.730	0.856	0.970	1.075	1.173	1.266	1.354	1.438	1.518	1.595
13.5	0.402	0.566	0.703	0.824	0.934	1.035	1.130	1.219	1.304	1.384	1.462	1.536

Light gray, dark gray and speckled shading indicates a *t*-value that represents ≥90%, ≥95%, or ≥99% confidence, respectively, i.e., the probability that a difference of ±10% can be detected. Next, move up the column from the selected confidence level to the second row from the top of the table. This row indicates the number of replicates required per time point. The first row indicates the total tests required, summing all time points and replicates together. For example, a method CV of 4.5% would require four replicates per time point for 90% confidence and five replicates per time point for 95% confidence.

Table 5
Replicates for six point/equal replicate number stability studies

Replicates Required for 6 Point Regression of Stability Data (all points have equal number of replicates): Failure Criteria >=±10%							
Total Reps	6	12	18	24	30	36	42
Reps/point	**1**	**2**	**3**	**4**	**5**	**6**	**7**
DF Regression	4	10	16	22	28	34	40
DF Chi Square	4	10	16	22	28	34	40
Critical chi square	11.1	20.5	28.8	36.8	44.5	52	59.3
Critical t (99%)	**4.604**	**3.169**	**2.921**	**2.819**	**2.763**	**2.725**	**2.704**
Critical t (95%)	2.776	2.223	2.119	2.074	2.049	2.03	2.02
Critical t (90%)	2.132	1.812	1.746	1.717	1.701	1.69	1.684
CVs **1.5**	3.488	**5.739**	**7.501**	**8.985**	**10.306**	**11.508**	**12.625**
2.0	2.616	**4.304**	**5.626**	**6.739**	**7.729**	**8.631**	**9.469**
2.5	2.093	**3.443**	**4.501**	**5.391**	**6.183**	**6.905**	**7.575**
3.0	1.744	2.869	**3.750**	**4.492**	**5.153**	**5.754**	**6.313**
3.5	1.495	2.460	**3.215**	**3.851**	**4.417**	**4.932**	**5.411**
4.0	1.308	2.152	2.813	**3.369**	**3.865**	**4.316**	**4.734**
4.5	1.163	1.913	2.500	**2.995**	**3.435**	**3.836**	**4.208**
5.0	1.046	1.722	2.250	2.695	**3.092**	**3.452**	**3.788**
5.5	0.951	1.565	2.046	2.450	**2.811**	**3.139**	**3.443**
6.0	0.872	1.435	1.875	2.246	2.576	**2.877**	**3.156**
6.5	0.805	1.324	1.731	2.073	2.378	2.656	**2.914**
7.0	0.747	1.230	1.607	1.925	2.208	2.466	**2.705**
7.5	0.698	1.148	1.500	1.797	2.061	2.302	2.525
8.0	0.654	1.076	1.406	1.685	1.932	2.158	2.367
8.5	0.615	1.013	1.324	1.586	1.819	2.031	2.228
9.0	0.581	0.956	1.250	1.497	1.718	1.918	2.104
9.5	0.551	0.906	1.184	1.419	1.627	1.817	1.993
10.0	0.523	0.861	1.125	1.348	1.546	1.726	1.894
10.5	0.498	0.820	1.072	1.284	1.472	1.644	1.804
11.0	0.476	0.783	1.023	1.225	1.405	1.569	1.722
11.5	0.455	0.749	0.978	1.172	1.344	1.501	1.647
12.0	0.436	0.717	0.938	1.123	1.288	1.439	1.578
12.5	0.419	0.689	0.900	1.078	1.237	1.381	1.515
13.0	0.402	0.662	0.865	1.037	1.189	1.328	1.457
13.5	0.388	0.638	0.833	0.998	1.145	1.279	1.403

Instructions: locate the test method precision (CV) in the far left column of the table, then move across the row until a shaded column is reached. Light gray, dark gray, and speckled shading indicates a *t*-value that represents ≥90%, ≥95%, or ≥99% confidence, respectively, i.e., the probability that a difference of ±10% can be detected. Next, move up the column from the selected confidence level to the second row from the top of the table. This row indicates the number of replicates required per time point. The first row indicates the total tests required, summing all time points and replicates together. For example, a method CV of 4.5% would require two replicates per time point for 90% confidence and three replicates per time point for 95% confidence.

to the 10% criteria does, on the face of it, appear reasonable, it is arbitrary, and reduces the overall acceptable level of quality, in essence "lowering the bar" for analytes with less precise test methods.

Since the quality control material is used to monitor the day-to-day or run-to-run performance of an assay, the criteria could be based, at least partially, on Westgard rules and the customer's ability to detect the change in concentration due to instability. In addition, the manufacturer's ability to detect the change, and the volume of testing required to do so should also be considered.

Using Table 1 as a reference, one can see that when a test method's within run CV reaches 8%, the number of replicates required for each time point to determine a statistically significant 10% slope at the 95% confidence is 10 replicates per each time point. When the CV reaches 10%, the number of replicates required is 13 per time point. This amount of testing may not only be expensive, but at this level of imprecision, the customer is unlikely to detect a 10% difference. This is about the point where the CV is usually added to the criteria. But, adding 8–10% at this point and effectively doubling the criteria suddenly increases the chance the customer will detect an issue. Therefore a more gradual approach to adjusting the criteria, which considers the voice of the customer, is more logical.

Not all laboratories are likely to monitor the QC results using all the Westgard rules, but the two most common, the 1_{3S} and the 2_{2S} helps establish a point from which to start. These two rules, when broken invalidate the customers test results [14].

As degradation proceeds, at some point the range of test result values obtained will lead to an unacceptable frequency of 1_{3S} Westgard rule violations. If we assume a ≥5% frequency as unacceptable, we can begin to determine appropriate fail-

Table 6
Replicates for six time point stability studies with intermediate points' replicates = 2

	Replicates Required for 6 Point Regression of Stability Data (Intermediate points = 2 replicates, T0 and Tfinal replicates varied): Failure Criteria >=±10%												
	Total Reps	12	14	16	18	20	22	24	26	28	30	32	34	36
	RepsT0 TF	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
	DF regression	10	12	14	16	18	20	22	24	26	28	30	32	34
	DF Chi Square	10	12	14	16	18	20	22	24	26	28	30	32	34
	Critical chi square	20.5	23.3	26.1	28.8	31.5	34.2	36.8	39.4	41.9	44.5	47	49.5	51.9
	Critical t (99%)	**3.169**	**3.055**	**2.977**	**2.921**	**2.898**	**2.845**	**2.819**	**2.797**	**2.779**	**2.763**	**2.75**	**2.74**	**2.73**
	Critical t (95%)	2.228	2.179	2.145	2.12	2.101	2.086	2.074	2.064	2.056	2.048	2.042	2.035	2.03
	Critical t (90%)	1.812	1.782	1.761	1.746	1.734	1.725	1.717	1.711	1.706	1.701	1.697	1.694	1.692
CVs	**1.5**	**5.739**	**6.870**	**7.879**	**8.815**	**9.680**	**10.487**	**11.263**	**11.998**	**12.712**	**13.380**	**14.036**	**14.667**	**15.290**
	2.0	**4.304**	**5.152**	**5.909**	**6.611**	**7.260**	**7.866**	**8.447**	**8.998**	**9.534**	**10.035**	**10.527**	**11.000**	**11.468**
	2.5	**3.443**	**4.122**	**4.728**	**5.289**	**5.808**	**6.292**	**6.758**	**7.199**	**7.627**	**8.028**	**8.422**	**8.800**	**9.174**
	3.0	2.869	**3.435**	**3.940**	**4.407**	**4.840**	**5.244**	**5.631**	**5.999**	**6.356**	**6.690**	**7.018**	**7.333**	**7.645**
	3.5	2.460	2.944	**3.377**	**3.778**	**4.148**	**4.495**	**4.827**	**5.142**	**5.448**	**5.734**	**6.015**	**6.286**	**6.553**
	4.0	2.152	2.576	2.955	**3.305**	**3.630**	**3.933**	**4.224**	**4.499**	**4.767**	**5.018**	**5.263**	**5.500**	**5.734**
	4.5	1.913	2.290	2.626	**2.938**	**3.227**	**3.496**	**3.754**	**3.999**	**4.237**	**4.460**	**4.679**	**4.889**	**5.097**
	5.0	1.722	2.061	2.364	2.644	**2.904**	**3.146**	**3.379**	**3.599**	**3.814**	**4.014**	**4.211**	**4.400**	**4.587**
	5.5	1.565	1.874	2.149	2.404	2.640	**2.860**	**3.072**	**3.272**	**3.467**	**3.649**	**3.828**	**4.000**	**4.170**
	6.0	1.435	1.717	1.970	2.204	2.420	2.622	2.816	**2.999**	**3.178**	**3.345**	**3.509**	**3.667**	**3.823**
	6.5	1.324	1.585	1.818	2.034	2.234	2.420	2.599	2.769	**2.934**	**3.088**	**3.239**	**3.385**	**3.528**
	7.0	1.230	1.472	1.688	1.889	2.074	2.247	2.413	2.571	2.724	**2.867**	**3.008**	**3.143**	**3.276**
	7.5	1.148	1.374	1.576	1.763	1.936	2.097	2.253	2.400	2.542	2.676	**2.807**	**2.933**	**3.058**
	8.0	1.076	1.288	1.477	1.653	1.815	1.966	2.112	2.250	2.384	2.509	2.632	**2.750**	**2.867**
	8.5	1.013	1.212	1.390	1.556	1.708	1.851	1.988	2.117	2.243	2.361	2.477	2.588	2.698
	9.0	0.956	1.145	1.313	1.469	1.613	1.748	1.877	2.000	2.119	2.230	2.339	2.444	2.548
	9.5	0.906	1.085	1.244	1.392	1.528	1.656	1.778	1.894	2.007	2.113	2.216	2.316	2.414
	10.0	0.861	1.030	1.182	1.322	1.452	1.573	1.689	1.800	1.907	2.007	2.105	2.200	2.294
	10.5	0.820	0.981	1.126	1.259	1.383	1.498	1.609	1.714	1.816	1.911	2.005	2.095	2.184
	11.0	0.783	0.937	1.074	1.202	1.320	1.430	1.536	1.636	1.733	1.825	1.914	2.000	2.085
	11.5	0.749	0.896	1.028	1.150	1.263	1.368	1.469	1.565	1.658	1.745	1.831	1.913	1.994
	12.0	0.717	0.859	0.985	1.102	1.210	1.311	1.408	1.500	1.589	1.673	1.754	1.833	1.911
	12.5	0.689	0.824	0.946	1.058	1.162	1.258	1.352	1.440	1.525	1.606	1.684	1.760	1.835
	13.0	0.662	0.793	0.909	1.017	1.117	1.210	1.300	1.384	1.467	1.544	1.620	1.692	1.764
	13.5	0.638	0.763	0.875	0.979	1.076	1.165	1.251	1.333	1.412	1.487	1.560	1.630	1.699

Locate the test method precision (CV) in the far left column of the table, then move across the row until a shaded column is reached. Light gray, dark gray, and speckled shading indicates a *t*-value that represents ≥90%, ≥95%, or ≥99% confidence, respectively, i.e., the probability that a difference of ±10% can be detected. Next, move up the column from the selected confidence level to the second row from the top of the table. This row indicates the number of replicates required for the T_0 and T_{final} time points. The first row indicates the total tests required, summing all time points and replicates together. For example, a method CV of 4.5% would require two replicates for the T_0 and T_{final} time points for 90% confidence and three replicates for the T_0 and T_{final} time points for 95% confidence.

ure criteria. A 5% frequency of violations will occur when the distribution of T_{final} values has ≥5% of its tail outside the 99% (±3S.D.) of the initial range of pre-established QC values ($T_0\mu \pm 3$S.D.). This is represented graphically in Fig. 1. Using a z table [13] we find that a 5% area corresponds to a z value of 1.65, and by subtraction (3S.D. − 1.65S.D.) we find that this corresponds to 1.35S.D. from the mean of the initial values ($T_0\mu$). Using this process, the failure criteria would be set at 1.35 times the total CV of the method.

By applying this same approach with the Westgard 2_{2s} rule, we choose a probability of two events occurring in succession equal to 5%. In this case, the tail area outside the established con-

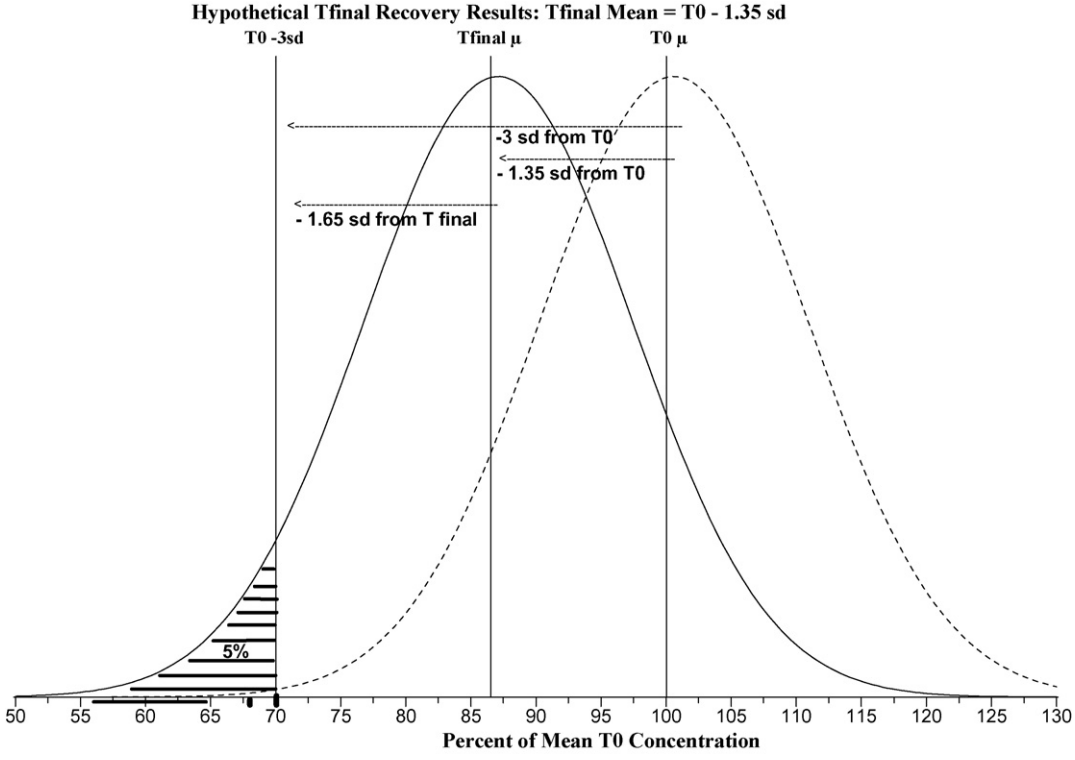

Fig. 1. Average shift down in recovery equal to 1.35 S.D. leading to 5% of results outside 3 sd.

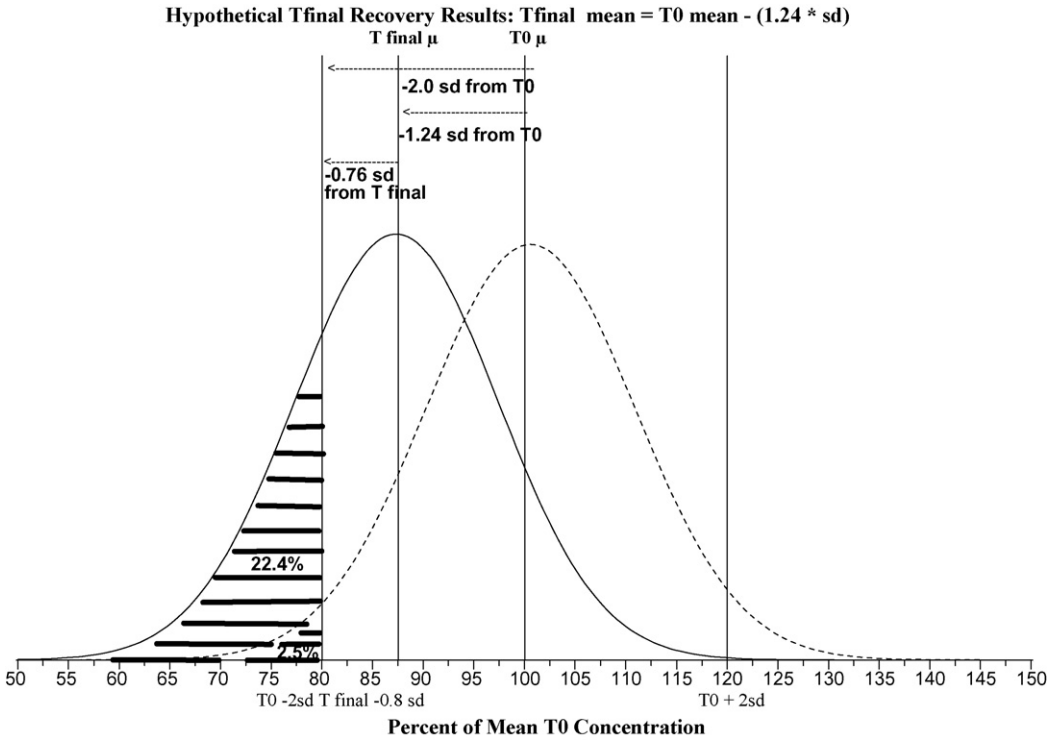

Fig. 2. Average shift down in recovery equal to 1.24 S.D. leading to 22.4 % of results outside 2 sd.

trol Charts 2S.D. range that is under the curve of stressed analyte range of results will equal $0.224 = \sqrt{0.05}$, which corresponds to a z value from the T_{final} distribution of 0.76. The difference between this z value and the $T_0 z$ value of 2 is $1.24 = (2 - 0.76)$ (Fig. 2).

The two z values obtained from the above analysis are relatively close, and either one could be used for the criteria. The smaller z value may be more defensible since it represents less customer risk. The volume of testing required to obtain statistically significant differences at the 95% confidence level will vary depending on the within run CV, which will usually be less than the total CV, for which the criteria is based. If we were to assume that the total CV is less than or equal to the within run CV, then the number of replicates per time point required for a two point comparison would equal 10. Therefore, the total number of replicates required to make this sort of determination would be $\leq$20.

This approach is much less arbitrary and also sets a threshold from which to begin increasing the criteria (CV = 8%). This probably should only be attempted when the within run CV reaches this level of imprecision. Then ideally, the criteria would be changed based on the total CV of the most widely used method.

6. DMAIC: control the process

Once the stability testing process is improved through appropriate criteria, testing volume and strategies, control through monitoring of its capability will be important. Test method performance should be assessed continually by monitoring controls and stability testing results for affects of imprecision, drift, and autocorrelation. Randomized test sequences will certainly reduce the chances that overlooked test method issues will impact results.

7. Conclusions

A successful stability evaluation is viewed by many in the sciences as requiring a strong knowledge of physical chemistry. Statistical tools do not appear to be considered nearly as important; likely because the analysis of the data is usually performed by software after the fact. However, with a six-sigma approach, stability testing is viewed as a process where the precision must be adequate to distinguish between good and bad product with a high degree of confidence. Therefore, establishing appropriate criteria and understanding process capability must be considered equally as important as the science behind the testing.

References

[1] N.R. Draper, H. Smith, Applied Regression Analysis, third ed., Wiley & Sons, New York, 1998.
[2] 21 CFR Parts 210 and 211, Current Good Manufacturing Practice in Manufacturing, Processing, Packing or Holding of Drugs; General and Current Good Manufacturing Practice For Finished Pharmaceuticals. Revisions as of May 2, 2006.
[3] L. Kennon, J. Pharm. Sci. 53 (1964) 815–818.
[4] I. Porterfield, J.J. Capone, Med. Dev. Diag. Ind. (1984) 45–50.
[5] V.J. Stella, J. Parent. Sci. Tech. 40/4 (1986) 142–163.
[6] T.B.L. Kirkwood, J. Biol. Stand. 12 (1984) 215–224.
[7] T.B.L. Kirkwood, Biometrics 33 (1977) 736–742.
[8] G. Anderson, M. Scott, Clin. Chem. 37 (1991) 398–402.

[9] H. Eyring, A. Stearn, Proceedings of the Symposium on the Physical Chemistry of Proteins, Am. Chem. Soc., 1938.
[10] I. Tinoco Jr., K. Sauer, J.C. Wang, Physical Chemistry: Principles and Applications in Biological Sciences, first ed., Prentice Hall, 1978.
[11] O. Levenspiel, Chemical Reaction Engineering, second ed., Wiley & Sons, New York, 1972.
[12] K. De Vore, J. Pharm. Biomed. Anal. 41 (2006) 293–298.
[13] D.J. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, Chapman and Hall/CRC, 2000.
[14] J.O. Westgard, P.L. Barry, M.R. Hunt, T. Groth, Clin. Chem. 27 (1981) 493–501.