Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Pharmaceutical and Biomedical Analysis 41 (2006) 293–298
JOURNAL OF PHARMACEUTICAL AND BIOMEDICAL ANALYSIS
www.elsevier.com/locate/jpba

Short communication

Have more confidence in your stability data: Two points to consider

Karl De Vore *

Bio-Rad Laboratories, 9500 Jeronimo Road, Irvine, CA 92688, USA

Received 21 July 2005; received in revised form 24 October 2005; accepted 25 October 2005
Available online 5 December 2005

Abstract

Simple statistical, mathematical, and chemical arguments are presented that will justify performing all stability studies using a different approach than is currently practiced in the pharmaceutical and IVD industries. The use of multi time point stability studies is in most cases a waste of resources that could be better spent on endpoint studies at less cost and a significant increase in the quality of the data.

Keywords: Stability; Linear regression; Assay precision; Two-point analysis

1. Introduction

Stability testing for pharmaceutical and IVD products is one of the most important and far-reaching process undertaken to insure product quality. In some cases, the cost involved in testing and analysis of stability data far outweighs the cost of any other one function performed in Research and Development. It is therefore essential that any protocol or general procedure written to establish or verify a stability claim be based on sound statistical and mathematical reasoning so the investment is cost effective.

The most significant contributor to the cost of testing is assay imprecision: as imprecision increases, the number of tests required to make sound conclusions also increases. Inter-assay imprecision can be eliminated with batch testing of back-loaded samples against a suitable reference (T_0) samples. Intra-assay imprecision cannot be eliminated but can be minimized through test method selection and randomized testing sequences. However, computer simulations of stability data have shown that not only is assay precision important, but also the way in which the data is collected.

At our facility computer simulations were used to support the argument that real time stability verification could be better assessed using the two-point system. The following demonstrates why this is true and how the same concepts can be applied to open vial and accelerated stability with only some qualifications.

2. Degradation

Although most stability data is analyzed using linear regression, the rate at which an analyte decays is almost never linear (referred to as zero order), but follows a curved path during the process. In almost all cases, the decay is of a higher order, in simple terms: the rate of decay is proportional to the amount of analyte present and therefore the absolute rate decreases as the concentration decreases [1].

For reactions referred to as first order:

$$[\mathrm{A}]_t = [\mathrm{A}]_0 \mathrm{e}^{-kt}$$

where $[\mathrm{A}]_t$ is the concentration of analyte A at time = t; $[\mathrm{A}]_0$, the concentration of analyte A at time = 0; k, the rate of decay.

For higher order reactions (2nd order and above):

$$[\mathrm{A}]_t = \sqrt[n-1]{\frac{1}{(n-1)kt + 1/[\mathrm{A}]_0^{n-1}}}$$

where n, the reaction order.

Whether a chemical reaction follows first or higher order kinetics, curve fit data can only give clues to the actual mechanism. In fact, a very complex process, such as bacterial growth follows first order type kinetics, but the actual mechanism of the growth involves countless steps of DNA

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

0731-7085/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2005.10.038

replication, transcription into RNA, protein synthesis, and so on.

It follows that even if the stability data fits well into a higher order equation, it does not mean that [A] is the only important constituent. In a multi-constituent solution, a single component's decay rate may be influenced by different factors at different times [2]. Therefore, no matter how precise the test method, number of replicates tested, or time points studied, the best line or curve fit to the data is merely a tool to predict real time performance, not a description of the mechanism of decay.

3. Retrospective data review

At our facility a review of R&D archived accelerated stability data for examples of higher order degradation reactions revealed a consistent pattern. When higher temperatures are used and the degradation was allowed to proceed far enough, the best curve fit tended to be higher order, the higher the temperature. This was even the case in lyophilized products, where one would expect minimal interaction between matrix constituents. The less degradation that occurred during the study meant that all rate equations would fit equally well. In the Fig. 1 example, both linear and 2nd order curves fit well for 3-methoxytyramine stability at 35°, but at 47°, the best fit was 2nd order [3].

The forgoing observations could be interpreted in one of two ways:

1. At higher temperatures different chemical reactions come into play that change the kinetics of degradation.
2. If the analyte concentration does not fall to a low enough value during the study, the difference between a linear regression fit and curve fits of higher order are so small that they cannot be measured with the available test methods or sampling protocols.

Although the first interpretation may be valid for some analytes, the second interpretation can account for almost every set of data that was reviewed. Therefore, if analyte degradation is allowed to proceed far enough, all should exhibit higher order (>0 or 1st order) degradation curves.

One interesting exception was uncovered in our review of experimental data. Prostate Specific Antigen (PSA) accelerated

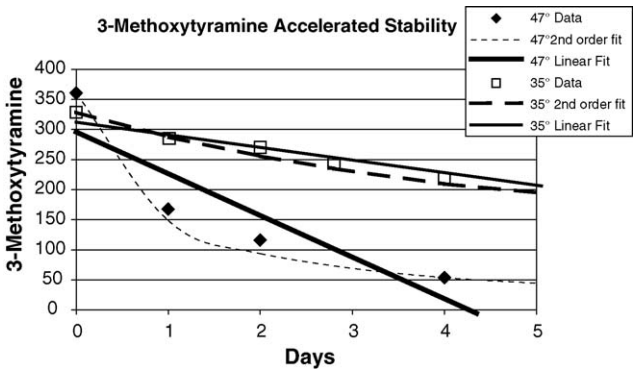

Fig. 1. 3-Methoxytyramine accelerated stability.

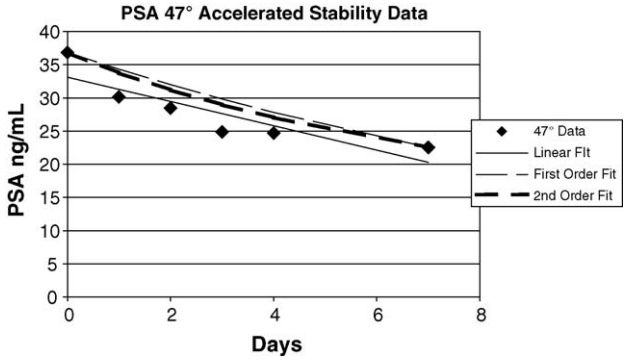

Fig. 2. PSA 47°C accelerated stability data.

stability in one experimental pilot had two study temperatures where the apparent degradation slowed considerably when the PSA concentration reached 60% of its initial concentration. In this case there appeared to be two species of PSA present, one more stable than the other. As shown in Fig. 2, none of the curves fit the data very well, but the change in rate did not occur until 40% of the PSA was eliminated [4].

4. Linear fits of higher order degradation

The data review also demonstrated that the difference in the stability estimate only becomes significant when degradation reaches approximately 20%. In any stability study, whether it be open vial or accelerated, the practice of using linear regression to estimate failure rates is therefore appropriate, provided 20% degradation is not exceeded. In cases where the degradation is greater, the initial time points, that only include the first 20% degradation are regressed.

Figs. 3 and 4 illustrate these differences between the linear and higher order curve fits who's endpoints (T_{final}) have the same results of 20% degradation [5].

In Fig. 3, the maximum difference between the 4th order and linear fits of the data is at the halfway point and is only 2.2% As time progresses the differences becomes more easily discernable (Fig. 4).

Although the differences are small in the early stages of degradation, why are not the higher order fits applied to the data?

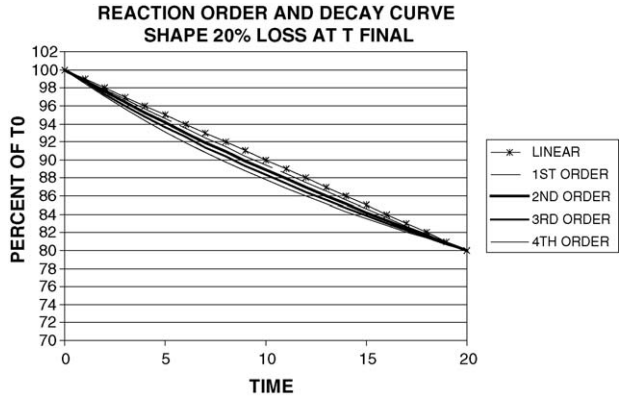

Fig. 3. Reaction order and decay curve shape 20% loss at T_{final}.

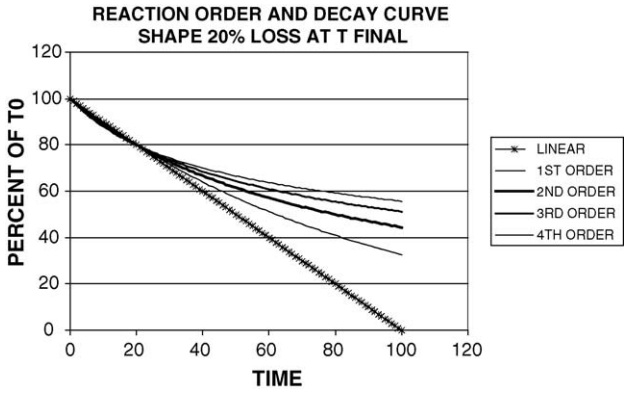

Fig. 4. Reaction order and decay curve shape 20% loss at T_{final}.

1. The commonly available software packages (Excel, Lotus, and Quatropro) do not have the functions to automatically calculate the rate constants.
2. At analyte degradations of ≤20%, applying a higher order rate curves makes an assumption about the data that may be incorrect due to the error. For example: given a method CV of 6% and 3 replicates per time point, one is unlikely to discern a difference of 5.0%, let alone 2.2%.

The ability to determine the difference between two sample means is best demonstrated using the formula for the *t*-test [6].

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\text{se}_1^2 + \text{se}_2^2}}$$

where t is the two-tailed t distribution value at the 95% confidence level, and the numerator is the difference between the two sample means.

se_1^2 = the squared standard error of the mean of sample 1
se_2^2 = the squared standard error of the mean of sample 2

$$\text{Standard error} = \frac{s}{\sqrt{n}}$$

where s is the standard deviation, or for purposes of illustration, the assay CV.

The resulting t from a above calculation must exceed the tabulated value to be statistically significant.

Rearranging we see that:

$$\overline{x_1} - \overline{x_2} = t\sqrt{\text{se}_1^2 + \text{se}_2^2}$$

If three replicates from each sample set are compared the tabulated t-value is 2.776. And, if the assay CV = 6% then

$$\overline{x_1} - \overline{x_2} = 2.776\sqrt{\frac{36}{3} + \frac{36}{3}} = 13.6$$

Therefore, given an average 6% test method CV, you could not detect with 95% confidence a difference of ≤13.6% between two sets of three replicates.

In fact, the assay precision would have to be 1.0% before you could even get close to detecting a 2.2% difference between two sample means.

$$\overline{x_1} - \overline{x_2} = 2.776\sqrt{\frac{1}{3} + \frac{1}{3}} = 2.3\%$$

5. The two-point versus six-point system

The preceding arguments should convince you that a linear fit of the stability data is appropriate if the degradation has not exceeded 20%. The following presents an equally compelling argument for reducing the number of time points in a stability study from six – generally thought of as the minimum number of time points required for regression – to two time points with increased replication.

The intent of using intermediate time points in a stability study is to track the course of decay more closely and therefore more precisely determine the point in time at which failure occurs. Implicit in this approach is the assumption that each point's value, and consequently any underlying error, contributes the same to the overall stability estimate. This is not the case. In fact, with same number of total tests, the use of a two-point system can reduce the error in the decay rate estimate by 30%.

5.1. Slope error

For a standard linear regression, the 95% confidence range of the slope (in our case decay rate) is given as,

$$b_1 = \pm\frac{t(n-2,\ 0.95)s}{\sqrt{\sum (x_i - \overline{x})^2}}$$

where $t(n-2, 0.95)$ is the tabulated t-value for $n-2$ degrees of freedom at the 95% confidence level [7].

$$\sum (x_i - \overline{x})^2$$

is the sum of squares of the values on the x-axis (in our case time), and s is the standard deviation of the residuals (in our case the average difference between the measured analyte recovery and the linear regression estimate for recovery at each time point).

You can see that given the same t value and s, the error in the slope is solely an inverse function of the square root of the sum of squares of the x values.

$$b_1(\text{error}) \propto \frac{1}{\sqrt{\sum (x_i - \overline{x})^2}}$$

So, as the average spread of the x values increases, the 95% confidence range (error) of the determination of the slope, or degradation rate, decreases. Therefore, the minimum amount of error possible is attained when all the x values are at the two extremes (T_0 and T_{final}).

To see how much reduction in error can be achieved, imagine a system used to estimate stability, where six equally spaced time points are used ($n = 6$)

Let $x_1 = 0$ (in our case T_0)

And $x_6 = Q$ (in our case T_{final})

Then starting at x_1, each successive time point's value increases by $Q/(n-1) = 0.2Q$.

The mean:

$$\bar{x} = \frac{0 + 0.2Q + 0.4Q + 0.6Q + 0.8Q + 1.0Q}{6} = 0.5Q$$

The sum of squares

$$\sum (x_i - \bar{x})^2 = (0Q - 0.5Q)^2 + (0.2Q - 0.5Q)^2 \ldots\ldots + (1.0Q. - 0.5Q)^2 = 0.7Q^2$$

and

$$\sqrt{\sum (x_i - \bar{x})^2} = 0.84Q$$

Therefore, the error in the slope estimate with six equally spaced time points is proportional to 1/0.84 = 1.20.

Now, if the same number of x values are used at only two points (T_0 and T_{final}) then $x_1 = x_2 = x_3 = 0$, and $x_4 = x_5 = x_6 = Q$.

And the mean

$$\bar{x} = \frac{0 + 0 + 0 + Q + Q + Q}{6} = 0.5Q$$

The sum of squares

$$\sum (x_i - \bar{x})^2 = 3(0 - 0.5Q)^2 + 3(1.0Q - 0.5Q)^2 = 1.5Q^2$$

and

$$\sqrt{\sum (x_i - \bar{x})^2} = 1.22Q$$

The error in the slope estimate with two time points is proportional to 1/1.22 = 0.82

Therefore the relative improvement in the precision of the slope estimate (95% confidence range) using two points of 3 replicates versus six equally spaced time points [100(1 − (0.82/1.20))] = 31.7%.

5.2. Y-Intercept error

Because the stability estimate is expressed as a percentage of the T_0 concentration, the precision of the *y*-intercept is equally as important as that for the slope. The 95% confidence range for the intercept is calculated using the following:

$$b_0 = \pm t(n-2, 0.95)\left[\frac{\sum x_i^2}{n\sum (x_i - \bar{x})^2}\right]^{1/2} s$$

where $t(n-2,0.95)$ is the tabulated *t*-value for $n-2$ degrees of freedom at the 95% confidence level, and *s* is the standard deviation of the residuals [7].

As shown in the example of the slope confidence range calculation, for six points equally spaced,

$$\sum (x_i - \bar{x})^2 = 0.7Q^2$$

For two points of 3 replicates each the result was $1.5Q^2$.

It can also be shown that for six points

$$\sum x_i^2 = 2.2Q^2$$

and for two points of 3 replicates each, the result is $3.0Q^2$.

Therefore, the results for the middle term of the *y*-intercept error calculations is:

six points $$\left[\frac{\sum x_i^2}{n\sum (x_i - \bar{x})^2}\right]^{1/2} = \sqrt{\frac{2.2}{6 \times 0.70}} = 0.72$$

two points $$\sqrt{\frac{3.0}{6 \times 1.5}} = 0.58$$

So, to use two points of 3 replicates each instead of six points of 1 replicate each, there is a modest reduction in the magnitude of the middle term of the *y*-intercept error calculation. Given the same tabulated *t*-value and standard deviation, the difference is 100(1 − 0.58/0.72) = 19.4%.

6. Standard error of analyte concentrations

In the foregoing demonstration of how the estimates of the slope and *y*-intercept can be improved by the use of two points versus six points, the standard deviation of the residuals (*s*) were assumed to be equal. However, using the same total number of tests, the standard deviation of the residuals is narrower using two points than when six points are used. In fact, it can be shown that that the precision using six points of 3 replicates each (18 tests per study condition) can be further improved by testing two points of 8 replicates each (16 tests per study condition).

s is derived from the sum of squares of the residuals (deviation of the Y values from the estimated regression line) as per the following equation:

$$s = \sqrt{\frac{\sum_{j=1}^{m}\sum_{u=1}^{n}\left(Y_{ju} - \hat{Y}_j\right)^2}{\sum n_{ju} - 2}}$$

where the numerator is the sum or squares of the residuals; *m*, number of *x* groups (time points) from *j* = 1 to *m*; *n*, number of replicates per group (replicates per time point) from *u* = 1 to *n*.

In the denominator $\sum n_{ju} - 2$ is the degrees of freedom.

The sum of squares of the residuals can also be expressed as follows

$$\sum_{j=1}^{m}\sum_{u=1}^{n}\left(Y_{ju} - \hat{Y}_j\right)^2 = \sum_{j=1}^{m}\sum_{u=1}^{n}\left(Y_{ju} - \overline{Y}_j\right)^2 + \sum_{j=1}^{m}\left(\hat{Y}_j - \overline{Y}_j\right)^2$$

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 [7].

Given the same assay precision, the first term will tend to be smaller when *n* = 16 (two points of 8 replicates each) than when *n* = 18 (six points of 3 replicates each). This of course is compensated for in the denominator in the calculation of *s*.

The most reduction in the sum of squares of the residuals with two time points comes form the second term on the right hand side of the equation. For two points $\hat{Y}_j = \overline{Y}_j$ and therefore

Table 1
Calculated parameters from each testing scheme

Parameter	Six points 3 replicates	Two points 8 replicates
s	6.71	6.41
$t(n-2, 0.95)$	2.120	2.145
$\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}$	1.45	2.00
$b_1 = \pm\frac{t(n-2,0.95)\times s}{\sqrt{\sum(x_i-\overline{x})^2}}$ slope error	±9.81	±6.87

$\sum_{j=1}^{m}\left(\hat{Y}_j - \overline{Y}_j\right)^2 = 0$. With six points, because of random error the second term will almost always be ≥0. Therefore, because of random error, the sum of squares of the residuals for six points of 3 replicates will almost always be greater than that for two points of 8 replicates.

7. Overall error reduction

In terms that can be appreciated, how much improvement in the precision of the stability estimate can be expected?

If one assumes a 6% assay CV, six points of 3 replicates will result in an average standard error at each time point of $\%/\sqrt{3} = 3.464$ We should expect that, on average, the sum of squares about the regression line,

$$\sum_{j=1}^{m}\left(\hat{Y}_j - \overline{Y}_j\right)^2 = 6(3.464)^2 = 72,$$

and

$$\sum_{j=1}^{m}\sum_{u=1}^{n}\left(Y_{ju} - \overline{Y_j}\right)^2 + \sum_{j=1}^{m}\left(\hat{Y}_j - \overline{Y}_j\right)^2 = 18 \times 6^2 + 72 = 720.$$

and finally

$$s = \sqrt{\frac{720}{16}} = 6.71$$

For two points of 8 replicates each

$$\sum_{j=1}^{m}\sum_{u=1}^{n}\left(Y_{ju} - \overline{Y_j}\right)^2 + \sum_{j=1}^{m}\left(\hat{Y}_j - \overline{Y}_j\right)^2 = 16 \times 6^2 + 0 = 576$$

and

$$s = \sqrt{\frac{576}{14}} = 6.41$$

The data in Table 1 shows that if given a true slope of 0 (no degradation), the error of that estimate could result in a calculated slope of −9.81, or 9.81% decrease during a six point/3 replicate study, and a calculated slope of −6.87, or 6.87% decrease during a two point/8 replicate study.

In our hypothetical examples where Y_{ju} = % recovery and x_{mn} (T_{final}) = 1, the analyte being tested would just miss failing by 0.19% in the six point/3 replicate study but miss by 3.56% in a two point/8 replicate study. If one is estimating shelf life using a ±10% criteria, the failure estimate for six points/3replicates would be 10/9.81 = 1.02 of T_{final} and for two points/8 replicates would be 10/6.87 = 1.45 of T_{final}. Therefore, on average six points of 3 replicates will shorten the shelf life estimate by 1 − (1.02/1.45) = 0.30, or 30%.

8. Conclusions

The logical conclusion from the arguments presented above is to adopt a new system of stability testing. Moving to a two-point system when the analyte degradation is less than 20% will improve the quality of the data and also result in a modest reduction in reagent cost. For those instances when degradation exceeds 20%, additional intermediate time point vials, set aside for such a contingency, can then be tested as a follow-up. New analytes, for which there is no prior data, would certainly be evaluated with multiple time points during feasibility.

What cannot be quantified from these improvements in the stability estimates is the expected decrease in time required to manage the data, the reduction in stability claim errors, product non-conformances, and any other consequences resulting from less reliable data. It may be difficult for some to appreciate these benefits, but most assuredly, they are substantial.

Because intermediate points do, in some rare cases, provide additional information, there may be some resistance to this testing system. At our facility the exceptions are of course rare and have included the following.

1. Bacterial contamination (biphasic): Analyte stability tracking may follow a linear trend through time, but begin to change exponentially after bacterial growth attains a certain level. Because any exponential change in analyte concentration would exceed 20%, follow-up testing would detect the issue.
2. Incompatibility of testing reagent with a matrix component (biphasic): In this case, an increase in an analyte was discovered in the results reported from the instrument. The increase was not due to the stability of the analyte itself, but with interference of a matrix stabilizer. The increase was actually over 20%, and therefore would have been detected using the proposed system.
3. Analyte stability and oxygen levels: Some analytes are particularly sensitive to oxygen levels—it is improved when oxygen is low. Under high accelerated stability temperatures, oxygen can be consumed and significantly decrease, leading to a decrease in the degradation rate of the analyte, and an overly optimistic estimate of real time stability. In these instances, the degradation was over 20% and therefore would have initiated follow up testing with intermediate time points.

As preparation for this article, a large sample of archived stability data were retrospectively reviewed. The intent was to find exceptional stability patterns in the data that included the first 20% of analyte degradation. 102 analytes and 504 stabilities found no unusual patterns that could be identified as coming from any mechanism other than random assay variation.

With new drugs, diagnostic controls, and reagents being added to the list of products undergoing testing, there may be rare instances in which an unusual pattern is overlooked. The probability would be virtually eliminated by initial feasibility studies that include multiple time points. The benefits of increased confidence in the data, which would apply to every stability study performed using the two-point system, would far outweigh the risk of missing some rare pattern that takes place during the first 20% of degradation. In addition, as implicitly shown previously with the six points/3 replicates per point example, many more than 3 replicates would be required to discern any pattern deviating from a straight line.

9. Recommendation

Based on the preceding arguments, the proposal is that the current stability procedures be revised.

1. For new analytes, the initial feasibility and perhaps the first development pilot stability studies, can include multiple intermediate time points. During the later stages of product development, two time points could be used, with the number of replicates based on method precision.
2. All other stability testing such as non-conforming product and product modification validation can be performed using two time points.
3. Additional, intermediate time point vials would be stressed and set aside if degradation exceeds 20% at T_{final}.

References

[1] Tinoco Jr., K. Sauer, J.C. Wang, Physical Chemistry: Principles and Applications in Biological Sciences, first ed., Prentice Hall, 1978.
[2] O. Levenspiel, Chemical Reaction Engineering, second ed., John Wiley & sons, New York, 1972.
[3] K. De Vore, Personal Experimental Data, 10/7/1998 & 11/2/1998.
[4] K. De Vore, Personal Experimental Data, 12/12/98 & 4/23/98.
[5] K. De Vore, Personal Excel Graphs, 4/14/05.
[6] D.J. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, Chapman and Hall/CRC, 2000.
[7] Draper, Smith, Applied Regression Analysis, third ed., Wiley & Sons, New York, 1998.