Drug Development and Industrial Pharmacy, 23(5), 427-439 (1997)

Statistical Evaluation of Stability Data of Pharmaceutical Products for Specification Setting

Peter Wessels,[1] Martin Holz,[2] Fritz Erni,[1] Kurt Krummen,[1] and Joerg Ogorka[1]

[1]*Sandoz Pharma Ltd., 4002, Basel, Switzerland*
[2]*Nelkenstr. 5, 79395 Neuenburg, Germany*

ABSTRACT

Recent experience has shown that the regulatory authorities try to narrow specifications for a new drug product closely around stability data that have been generated in the course of registration stability testing. This should ensure that the quality of the marketed product is produced within tight limits, very similar to the quality which is registered. A strategy is proposed to set scientifically based, statistically justified specifications. The long-term stability data available at the time point of registration application are extrapolated to a target shelf life (usually 36 months) of the new pharmaceutical product by linear regression. Batch-to-batch heterogeneity is tested as prescribed by the International Conference on Harmonization (ICH) guideline on stability testing of new drug substances and drug products. The data of different batches are combined in appropriate statistical models for further evaluation. The one-sided 99% confidence limit for individuals is used for the determination of release and shelf life specifications. Special attention is given to power calculations by which patients' risk of receiving material which does not fulfill the requirements can be controlled.

INTRODUCTION AND OBJECTIVE

The stability with time under given climatic conditions is a very important property of a new pharmaceutical product. Comprehensive stability studies are routinely conducted by the pharmaceutical industry in order to evaluate any kind of chemical and physical degradation of the product. Dosage units from batches are sampled randomly and stored under controlled temperature and humidity conditions. Individual samples are taken at predetermined time points and analyzed in order to assess, for example, decrease of assay of active ingredient, increase of the content of degradation products, and change of dissolution behavior. From the resulting data the specifications for the new drug product are derived. This means that, on the one hand, the time span during which the product characteristics remain within certain limits under given climatic conditions (shelf life) is determined. On the other hand, the stability data play a key role in determination of the limits for most of the quality characteristics themselves (besides considerations of efficacy, safety, and technical feasibility).

The stability testing requirements for registration application are laid down in the International Conference on Harmonization (ICH) guideline on stability testing of new drug substances and drug products (1). It is stated in the guideline that for registration, at least 12 months long-term stability information has to be provided on a minimum of three batches. The registration stability batches should be representative of the material which is intended for marketing. Therefore, the manufacturing process, the size of the batches, and the packaging are exactly specified in the ICH guideline.

The development target for a new drug product is a shelf life that is significantly longer than 12 months. Currently, a 36-month shelf life is accepted by the regulators as a maximum for unstable products; 60 months is accepted as a maximum for products which do not change with time. This can lead to the situation in which setting of specifications has to be performed based on only 12-month data available at the time point of registration application. This is difficult and needs a special strategy. It is particularly true because it is explicitly stated in the ICH guideline that the stability information obtained from previous development batches with smaller batch sizes, with even slightly different formulation or packaging is considered only as supportive data (1).

A procedure for statistical evaluation of stability data for the purpose of shelf life determination—not for the stipulation of the specifications for the quality characteristics itself—is described in the U.S. Food and Drug Administration (FDA) guidelines (2) and by the International Conference on Harmonization (1). The procedure has also been described in detail in several publications (3–5). It consists of the calculation of individual regression lines for each batch under each experimental condition (temperature, humidity, packaging). An analysis of covariance (ANCOVA) is applied to assess differences between the regressions of the batches. Depending on the results of the ANCOVA calculation, the data of the batches are combined in appropriate statistical models for further evaluation. The expiration dating period (shelf life) is determined as the intersection point of the confidence limit for the regression line with the given specification limit. Some other scientific techniques for the assessment and prediction of the stability of pharmaceutical products have also been reported in the literature. They include principles of reaction kinetics such as the effect of the reaction order and the temperature, Arrhenius kinetics (6–8), and statistical techniques such as the calculation of linear regression lines and confidence intervals (9,10). Also general linear model calculations (11), models with random and fixed effects (12,13), and models with nested error structure (14) are used for the purpose of evaluating stability data.

In this paper we propose a detailed strategy and statistical methodology for the setting and justification of specifications based on limited stability data and based on a target shelf life. In this context special attention is given to the assessment of both the producer's risk of rejecting a batch which fulfills the requirements (α error, error of first kind, type I error) and to the risk of the consumer (patient) receiving medication from a batch which is out of specification (β error, error of second kind, type II error).

STRATEGY FOR SETTING AND JUSTIFICATION OF SPECIFICATIONS

The stability information obtained from batches which have been produced early during development cannot be used as primary source for specification setting (1). Therefore, it is necessary to extrapolate the limited stability information on the large batches available at the time point of registration application to the expected shelf life (usually 36 months). Based on this extrapolation, target specifications are stipulated. Even if the target specifications are not fully supported by real-time data at the time point of registration application, the statistical calculation enables the conclusion that a future individual measurement at the end of the shelf life will be within specification, with a given level of confidence. The target specifications have to be justified by taking into account other factors besides statistical ones which influence the setting of specifications (e.g., safety data, technical feasibility, reproducibility of the process, stability information from previous batches). Also they have to be justified by updating the calculation with new results obtained from the stability testing program which is ongoing during the registration file review period, and which is usually extended to the time after launch of the drug product.

The setting of specifications closely around the stability data which are obtained from the registration stability batches implies the risk of a failure of later production batches because the production batches may vary more than the development batches did. Therefore, the sensitivity of the critical process parameters and the ruggedness of the process has to be assessed during the production scale-up by using a corresponding validation scheme. Also, the ruggedness of the analytical methods should be fully validated prior to the initiation of the registration stability program.

GRAPHICAL EVALUATION OF STABILITY DATA

The listing of stability results in a table format is necessary for documentation, but the tables do not allow easy assessment of key parameters like slopes and intercepts, and the inherent variability of the data. As a consequence the classification of the quality characteristics as critical or noncritical for the product remains sometimes difficult. Therefore, the presentation of the stability results in the registration file should always be supported by visualization of the stability information in appropriate charts with respect to:

- The impact of storage conditions/packaging on the stability
- Differences between batches
- Scattering of the data, for example, due to the variability of the analytical method

Maximum clarity is obtained by presenting the data in separate *xy* scatter charts with standardized symbols for all charts and the same linear scales for the *x* axis and the *y* axis depending on the batch with the longest observation period and the extreme values observed during stability testing. Depending on the effect which should be illustrated, the presentation of the data batch by batch or grouped (e.g., by storage temperature or by packaging) is suggested (Fig. 1).

STATISTICAL EVALUATION OF STABILITY DATA

Single Values Versus Mean Values

Usually, for stability testing two different samples from the material which has been stored in the climate chamber are prepared and measured against the reference standard. In most cases these replicates do not reveal the whole variability of the analytical method. The day-to-day variability is generally more pronounced than the method's repeatability obtained from simultaneously prepared and analyzed replicates. The reason is that for one particular analysis, the conditions are fixed

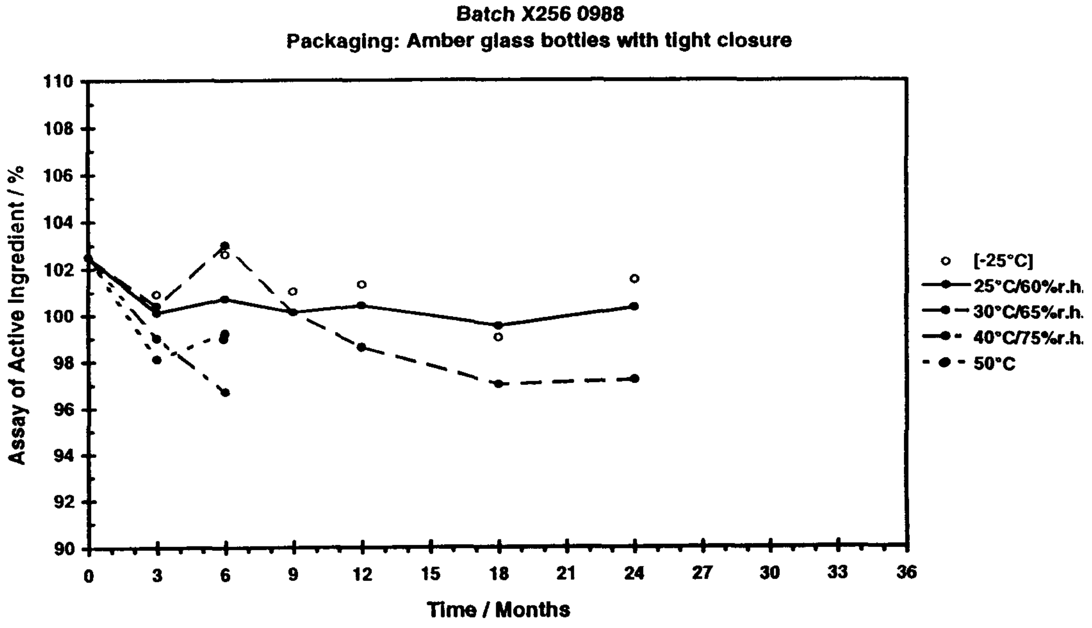

Figure 1. Graphical presentation of assay of active ingredient stability data.

(analyst, reagents, conditions in the lab, analytical instrument, etc.) whereas weeks or months later the conditions may differ significantly. There would be no gain in information if the replicates were included as individual values in the statistical evaluation. Therefore, means of the within-day replicates are used. It was shown (14) that the result of an ANCOVA calculation on the means correctly estimates the common slope (and common intercept) if the within-day variability is clearly smaller than the day-to-day variance component.

For the analysis of chemical parameters (e.g., assay of active ingredient and degradation products by high-performance liquid chromatography—HPLC), the difference between the day-to-day and the within-day variance components can be easily assessed if samples stored in a deep freezer are analyzed together with their counterparts stored under long-term or accelerated conditions (under deep-freezer conditions the drug substance is expected not to change chemically with time). The day-to-day variability can be removed by normalization of the stability data on these data or by using them as an additional factor in the statistical model as described by Langenbucher (11). This leads to more precise stability information because the effect of the day-to-day variability is removed from the residual mean square, which strongly increases the power of subsequent statistical tests. However, in general the whole effective analytical uncertainty should be included in the specification in order not to underestimate the variability of future results contributed by the analyst when the analysis is done in different laboratories, by different operators, with different analytical equipment.

Assessment of the Batch-to-Batch Variability

From a statistical point of view, pooling of data obtained from different batches stored under similar conditions (packaging, storage conditions) into one overall estimate is appropriate only if the data have been proven statistically to be homogeneous. This test is done by an analysis of covariance calculation (ANCOVA) (15) assuming time-independent variance and independently normally distributed residuals. The key numbers in the resulting ANCOVA tabulation are the common slope, the common intercept, and the probabilities of the *F* ratios. The latter have to be considered in a logically sequential manner in order to test the appropriateness of the common regression model (16):*

- *Test 1*: The first hypothesis to test is the equality of slopes of the batches' individual regressions. If the null hypothesis (regression lines of individual batches are parallel) has to be rejected on a given level of significance (e.g., $\alpha = 0.25$), pooling of the data into one overall common slope estimate is statistically not justified. No further information from the ANCOVA should be taken into account and the evaluation has to be restricted to the worst batch alone. Usually this results in a very pessimistic assessment because only few data are available for each individual batch and the worst batch is commonly the one with the fewest time points and the shortest period of observation, which results in the broadest confidence interval. It is therefore recommended that the test be repeated with appropriate subsets of batches (17).
- *Test 2*: The second level of significance to be checked, if test 1 is passed, is the one for the hypothesis of no time dependence (zero slope). It gives the probability that the common slope is indistinguishable from zero. If there is certain evidence against rejecting the null hypothesis, no change of the data with time is detected and the quality characteristic has to be considered as noncritical for the stability of the product. However, the pooling decision still needs statistical support (see next section).
- *Test 3*: The third probability which has to be considered after proving the common regression and the time dependency is that for the hypothesis of equality of the adjusted means (batch means adjusted for the time dependency). This value represents the result of the test of the null hypothesis that the regression lines of the individual batches have the same intercept. In this case the "common slope, common intercept" model is appropriate. If not, a common slope may still be assumed, but the intercept of the worst batch under the common slope model has to be used in place of the common intercept ("common slope, separate intercepts" model) (4).

The ICH guideline (1) prescribes a fixed critical probability of $\alpha = 0.25$ for the tests of equality of slopes and for the equality of adjusted means (the critical probability for the zero slope test is not mentioned in the guideline). The $\alpha = 0.25$ level is much more strict than the more common level of $\alpha = 0.05$ since pooling

*Experience has shown that transformation of data (e.g., on a logarithmic or quadratic scale) is usually not necessary to apply linear regression (4). This is particularly true when relatively small changes within the period of observation are detected.

should only be permitted if not even a mild evidence of heterogeneity is detected. So the permission to pool is a rare event. But the idea behind this strict procedure is the reduction of the β error, the risk of not detecting a difference between the batches which in fact exists (true H_A). It is the intention to keep this risk as small as possible because it relates to safety and efficacy of the drug (consumer's risk). However, minimizing β by simply increasing α is a very coarse method and in addition could lead to wrong decisions: If the analytical method is precise and the data of the individual batches lie on more or less straight lines, it is much more difficult to pass the test than if the variability of the analytical method is high. In fact, the experimental scatter may be so high that not even $\alpha = 0.25$ leads to a sufficiently small consumer's risk (17). In order to prevent these situations and to make most effective use of the data, the consumer's risk β has to be specified explicitly, and the critical significance level α needed to reject the null hypothesis has to be adjusted accordingly.

The relation between α and β for an F test is a complicated function of the experimental scatter, the degrees of freedom, and the critical difference the test should be capable of detecting (noncentral F distribution). However, the approximation of the noncentral F distribution by using the central F distribution (18) allows for easy computation as follows:

$$\alpha_{\text{critical}} = 1 - F\left\{\left[\left(\frac{dF1+\lambda}{dF1}\right) \times F^{-1}\left(\beta, \frac{(dF1+\lambda)^2}{(dF1+2\lambda)}, dF2\right)\right], dF1, dF2\right\}$$

where α_{critical}: critical level of significance, producer's risk; β: probability of second type error, consumer's risk; F: quantile of the central F probability distribution; F^{-1}: inverse of the cumulative distribution function of the central F distribution; *dF1*: degrees of freedom under the null hypothesis = (number of batches −1); *dF2*: degrees of freedom for residual error mean square (*MSE* from the ANCOVA); λ: noncentrality parameter = $SS(H_A)/MSE$ (see next equation).

The noncentrality parameter λ depends on the residual error *MSE* and on the sum of squared deviations $SS(H_A)$ resulting from the constellation as characterized by the alternative hypothesis H_A against which one wants to be protected. The most reasonable choice for H_A in the case of drug stability test situations is a situation where one outlier batch exists in the less favorable direction while the remaining batches are equal. This is the worst-case scenario because it leads to the lowest possible sum of squares deviation between the batches. In this case $SS(H_A)$ and thereby λ can be calculated as follows (17):

$$SS(H_A) = \frac{w_k \cdot \Delta^2 \cdot \left(\sum_{j=1}^{k} \sum_{i=1}^{n_j} (t_{ij} - \bar{t}_{.j}) - w_k\right)}{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (t_{ij} - \bar{t}_{.j})}$$

where $SS(H_A)$: sum of squares deviation corresponding to H_A: w_k: lowest of all weights, typically the weight of the batch with the fewest observations: Δ: minimum detectable critical difference; t_{ij}: time point i within batch j; $\bar{t}_{.j}$: mean of the time values of batch j.

Δ is the smallest possible difference between the outlying batch and the remaining batches, which should be detectable by the ANCOVA F test (e.g., a true difference between slopes of $\Delta = 1\%$/year should be detectable with an error of $\beta = 1\%$). It should be stressed that the choice of the critical values for α, β, and Δ is a pharmaceutical and regulatory issue, and not a statistical one. *MSE* and the degrees of freedom are taken from the ANCOVA table. For the test for equality of slopes, the error mean square under the separate slope model, *MSE(2)*, is used; for the test for zero slope and for equality of adjusted means, the error mean square under the common slope model, *MSE(1)*, is the appropriate error term for the F test.

Stability Data Revealing No Time Dependence

If the level of significance for H_0 = zero slope is above $\alpha_{\text{critical zero slope}}$, no time dependence of this quality characteristic is detected from the statistical point of view (16). If, in addition, the power to detect any meaningful time dependency if great enough, any further evaluation assuming time dependence is not reasonable. The data can be treated as if they were repeated measurements of the same samples, and the unadjusted means can be compared using an analysis of variance (ANOVA) F test. In most cases the result for the equality of the adjusted means test of the ANCOVA calculation represents a good estimate for the result of the ANOVA calculation for the unadjusted means, because the time-dependent variance component is almost negligible.

Assuming the null hypothesis of equal batch means cannot be rejected on a given level of probability $\alpha_{\text{critical equality of adjusted means}}$, and the power to detect a meaningful difference is sufficiently high, different batches can be considered as equal and the specification can be calculated as the one-sided confidence limit for individual data of the one-dimensional population. Under the assumption that future batches will have equal quality compared to the registration stability batches, the setting of specifications by applying a 95% confidence limit will statistically result in a failure of 1 out of 20 successfully prepared batches just by chance. Considering the tremendous costs of a batch failure it is proposed to use the 99% confidence limit instead of the 95% confidence limit. The proposal is also justified because it is not the producer's risk but the consumer's risk, represented by the result of the power calculation (see below), by which the risk of the patient of receiving a batch which is out of specification is correctly assessed.

$$CL = \bar{x}.. \pm CI \qquad \text{with } CI = t_{0.01(1)dFE} \cdot \sqrt{MSE}$$

where CL: confidence limit (for characteristics which are increasing with time CI has to be added; for characteristics which are decreasing with time CI has to be subtracted); $\bar{x}..$: grand mean of the data of all stability batches at all time points (in case of a "common slope, separate intercepts" model $\bar{x}..$ has to be replaced by $\bar{x}.$, which is the mean of the stability data of all time points of the worst batch); $t_{0.01(1)dFE}$: one-sided Student t value for 99% probability and dFE degrees of freedom (dFE: degrees of freedom from the ANOVA); MSE: residual error variance from the ANOVA.

As already mentioned above, for complete characterization of a statistical test, the consumer's risk has to be assessed and specified in addition to the producer's risk and its critical probability level. In this case it is the risk of the patient of receiving medication from a batch which is out of specifications (19). For the Student t test the β error can be calculated as a function of the assumed true difference when the standard error and the number of samples are known (noncentral t distribution). However, it is more advantageous to specify the power 1 − β in advance (e.g., 1 − β = 0.9) and then to compute the minimum detectable real difference between the stability batches and the batch under investigation (20,21), which can analytically be detected with the stated power. For bioequivalence testing it is common to calculate this difference for β = 0.2 probability (power = 80%). For the setting of specifications we propose a much stronger limit β = 0.05 which takes into account the safety of the customer to a larger extent. The resulting minimum detectable difference should be considered with respect to what is known, for example, from safety studies, therapeutic range, etc.

$$\Delta \geq (t_{\alpha(1),dFE} + t_{\beta(1),dFE}) \cdot \sqrt{MSE \cdot \left(\frac{1}{n_{\mathrm{sb}}} + \frac{1}{n_{\mathrm{b}}}\right)}$$

where Δ: minimum detectable difference between the mean of the stability batches and the batch under investigation (in case of a "common slope, separate intercepts" model the minimum detectable difference is calculated under the assumptions described above); n_{sb}: number of data of the stability batches; n_{b}: number of the data of the batch under investigation (usually 1); $t_{\alpha(1),dFE}$: one-sided Student t value for the specified producer's risk (e.g., α = 0.01); $t_{\beta(1),\, dFE}$: one-sided Student t value for the consumer's risk (e.g., β = 0.05); MSE: residual error variance from the ANOVA; dFE: degrees of freedom from the ANOVA.

If the resulting level $\bar{x}.. - \Delta$ (e.g., for content of active ingredient) or $\bar{x}.. + \Delta$ (e.g., for content of degradation products) reveals any risk to the patient (e.g., critical level of a degradation product), the calculation has to be repeated with a tighter confidence limit (e.g., α = 0.05) corresponding to an increased producer's risk but decreased consumer's risk. For critical quality characteristics (e.g., level of a toxic degradation product), keeping the β error under control should have absolute priority.

Example I

The data of the stability batches I to III and the result of the ANCOVA calculation are given in Tables 1 and 2, respectively. When tested against a fixed $\alpha_{\mathrm{critical}}$ = 0.25, the data of the batches are allowed to be combined according to a "common slope, common inter-

Table 1

Stability Data (Content of Degradation Product) for Example I

	Content of Degradation Product (%)		
Time (Months)	Batch I	Batch II	Batch III
0	0.3	0.5	0.3
3	0.1	0.3	0.4
6	0.3	0.3	0.4
9	0.4	0.5	0.6
12	0.2	0.3	0.2
18	0.4		
24	0.1		

Table 2

Results of ANCOVA Calculation Using the Stability Data Given in Table 1

Source	*dF*	*SS*	*MS*	*F*	*p(F)*
Equality of intercepts	2	0.0467	0.0234	1.3206	0.3005
Zero slope	1	0.0033	0.0033	0.1856	0.6736
Error (1)	13	0.2299	0.0177		
Equality of slopes	2	0.0022	0.0011	0.0547	0.9471
Error (2)	11	0.2276	0.0207		

cept" model for further statistical evaluation. The slope of the common regression line is not significantly different from zero [$p(F)_{\text{zero slope}} = 0.67$]. However, a look at the graphical chart reveals an unusual scattering of the values and a difference of the slope of batch II compared to batches I and III (see Fig. 2).

In order to assess the second kind error, not detecting a difference which in fact exists, a power calculation was performed as the next step. According to the available background information (safety data, target population, pharmacological effects of the degradation product, etc.), the critical α values were calculated based on the following assumptions: If the true initial value of one of the batches under investigation is 0.2% higher than the other batches, if the true common slope of the batches is different form zero by more than 0.01%/months (= 0.4% at the end of the expected shelf life), and if the true value of one of batches at the end of the expected shelf life differs by more than 0.4% from the other batches, this should be detected with a probability of 95% (β = 0.05).

According to the result of the power calculation (Table 3), a "zero slope, separate intercepts" model instead of a "zero slope, common intercept" model is appropriate for combination of the data for further statistical evaluation. The calculation of the 99% one-sided upper confidence limit for individuals under the respective model led to 0.73% as the specification limit for the content of this degradation product (see Fig. 3).

Again, the calculation of the confidence limit assesses only the producer's risk (probability of a failure of one of the batches which in fact is within specifications).

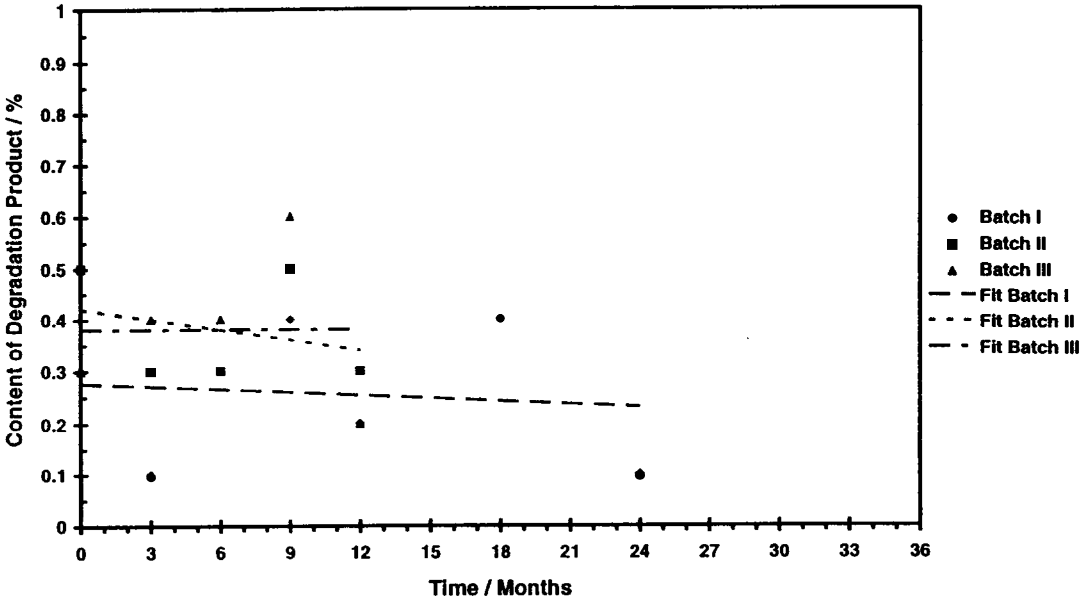

Figure 2. Graphical presentation of the stability data given in Table 1: single values and linear regression fits of the data of the individual batches.

Table 3

Results of the Power Calculation Using the Stability Data Given in Table 1

Source	Minimal Differences to Detect[a]	Probability 1 − β of Detecting the Difference[e]	α_{critical} for Decreasing β to 0.05	Model for Combining the Data
Equality of intercepts	0.2%[b]	0.32	0.92	Common slope
Zero slope	0.01%/month[c]	0.83	0.49	Zero slope
Equality of slopes	0.4%[d]	0.74	0.65	Separate intercepts

[a] To be detected with $\beta = 0.05$.

[b] Difference of one batch in the less favorable direction compared to the other ones at release.

[c] Difference of common slope from zero (= 0.4% at the end of the expected shelf life).

[d] Difference of one batch at the end of the expected shelf life (= 36 months) compared to the others.

[e] Based on $\alpha_{\text{critical}} = 0.25$.

Therefore, an additional power calculation was performed to assess the probability of passing by mistake a batch which in fact is out of specification (consumer's risk). The corresponding calculation led to 1.08%. This means that a batch which truly contains 1.08% of the degradation product is detected as being out of specification with a probability of 95%. In order to address the risk for the patient, the safety margin concerning the content of the degradation product must clearly include this level.

Calculation of Release and Shelf Life Specifications for Quality Characteristics Which Change with Time

If the ANCOVA calculation reveals a time dependency of the stability data (test 2 in the test sequence), further evaluation is based on the common least squares fit line and its confidence interval. The common slope of the regression line is obtained from the ANCOVA. The calculation of the zero time intercept depends on the

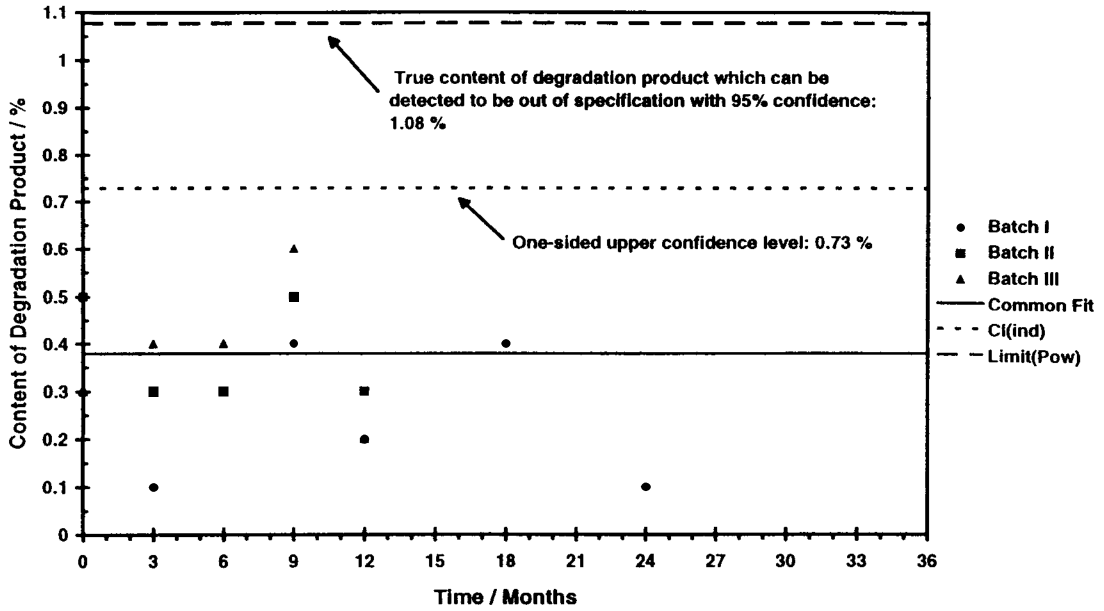

Figure 3. Results of the statistical evaluation of the stability data given in Table 1: single values, common fit with slope = 0 and intercept shifted to intercept of the worst batch, one-sided upper confidence limit [CL(ind)] and true content of degradation product which can be detected as out of specification with 95% probability [Limit(Pow)].

result of the level of significance of H_0 = equality of adjusted means. If it is above $\alpha_{\text{critical equality of adjusted means}}$ the zero time intercept for the common regression line is taken. If not, it is calculated based on the data of the worst batch while retaining the common slope and its confidence interval (4).

Two types of confidence intervals can be computed for linear regression:

• Confidence interval for the mean	Area containing the fit line (mean value) in 100 (1 - α)% of all repetitions of the analysis
• Confidence interval for individuals	Area containing all individual values in 100 (1 - α)% of all repetitions of the analysis

The one-sided confidence limit for individuals is calculated according to the following equation as a function of the time (T) (22):

$$CL(T) = aT + b \pm CI \qquad \text{with } CI = t_{0.01(1),dFE} \cdot \sqrt{A + B + C}$$

$$A = \frac{MSE(T - \bar{t}_{..})^2}{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (t_{ij} - \bar{t}_{.j})^2} \qquad B = \frac{MSE}{N} \qquad C = MSE$$

where CL: confidence limit (for characteristics which are increasing with time CI has to be added; for characteristics which are decreasing with time CI has to be subtracted; a: common slope; b: zero time intercept of the common regression line (in case of a "common slope, separate intercepts" model the zero time intercept of the common regression line is calculated based on the data of the worst batch); T: time for which the calculation is performed (e.g., T = shelf life); A: variance of the slope; B: variance of the intercept; C: variance of the individual values; $t_{0.01(1),\ dFE}$: one-sided Student t value for 1% probability and dFE degrees of freedom (dFE; degrees of freedom from the ANCOVA); $\bar{t}_{..}$: grand mean of all time points of all batches; $\bar{t}_{.j}$: mean of the time values of batch j; t_{ij}: Time point i within batch j; N: number of all data points of all batches; MSE: residual error variance from the ANCOVA.

Both the ICH guideline (1) and the FDA guideline (2) prescribe the use of the confidence interval of the mean for the statistical evaluation of stability results (in order to calculate the confidence interval for the mean, the variance of the individual values is set to $C = 0$; see equation given above). The reason is that the statistical evaluation of stability data, as described in the guidelines, aims at the determination of the shelf life, not at the setting of specifications. For the purpose of shelf-life determination, the use of the confidence interval for the mean is suitable since the expected true value of the characteristic—independent from the analytical error—must be within specification until the expiry date is reached. For the purpose of specification setting it is recommended the the confidence interval for individual values be used, which is broader than the confidence interval for the mean. Only by means of this confidence interval can the likelihood that the value of a future individual observation of the quality characteristics will be within specification be assessed.

For characteristics that are expected to decrease with time (e.g., content of active ingredient) the lower one-sided 99% confidence interval has to be used. For characteristics that are expected to increase with time (e.g., content of degradation products), the upper interval is appropriate. For drug characteristics with both an upper and lower specification limit [e.g., the concentration of an active ingredient in a solution can either increase by loss of the solvent (permeation through the packaging) or can decrease by degradation], the two-sided 99% confidence interval has to be applied. Due to the higher Student t value, the two-sided confidence limits are much broader than the one-sided ones.

The specifications at release and at the projected minimum shelf life (e.g., T = 36 months) are given by the confidence interval for individuals. They define the range in which the data of the release analysis or the analysis at a later stage (recontrol analysis) will fall with a given level of probability (assuming that the future batches equal the stability batches with respect to its initial quality and its stability).

By calculating the 99% one-sided confidence interval, the producer's risk is fixed to 1% (probability with which a compliant batch will wrongly be rejected). The consumer's risk that a noncompliant batch will wrongly pass the specifications is calculated similarly to the time-independent case.

$$\Delta(T) \geq (t_{\alpha(1),dFE} + t_{\beta(1),dFE}) \cdot \sqrt{MSE\left(\frac{1}{n_{sb}} + \frac{1}{n_b} + \frac{(T - \bar{t}_{..})^2}{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (t_{ij} - \bar{t}_{.j})^2}\right)}$$

where $\Delta(T)$: minimum detectable difference; n_{sb}: number of data of the stability batches; n_b: number of the data of the batch under investigation (usually 1); $\bar{t}_{..}$:

grand mean of all time values of all batches; $\bar{t}_{\cdot j}$: mean of the time values of batch j; t_{ij}: time point i within batch j; T: time for which the calculation is performed (e.g., T = shelf life); $t_{\alpha(1),\,dFE}$: one-sided Student t value for the specified producers risk and *dFE* degrees of freedom; $t_{\beta(1),\,dFE}$: one-sided Student t value for the consumers risk (e.g., $\beta = 0.05$) and *dFE* degrees of freedom (*dFE*: degrees of freedom for *MSE*); *MSE*: residual error variance from the ANCOVA.

Example II

The data of batches IV to VI and the result of the analysis of covariance calculation are given in Tables 4 and 5, respectively. When tested against a fixed $\alpha_{\text{critical}} = 0.25$, the slopes of the batches must be considered as different from each other ($p(F)_{\text{equality of slopes}} = 0.06$) and combination of the data of the batches is not allowed. However, a look at the graphical chart reveals a very precise measurement of the data and only a very slight difference of the slopes of the batches (see Fig. 4).

In order to assess the consumer's risk, a power calculation was performed based on the same assumptions as listed in example I. The result of the power calculation (see Table 6) indicates that due to the precise measurement of the data, the probability for detecting a true 0.4% difference of one of the batches at the end of the shelf life is much higher than 99%. For a 95% probability the level can be lowered to $\alpha_{\text{critical equality slopes}} = 0.04$ which is fulfilled [$p(F) = 0.06$]. Thus, a common slope model is in accordance with the allowed slope difference that was specified. The corresponding calculations for the significance of "zero slope" and the "equality of the adjusted means" confirmed what was already clear from the graphical chart. Therefore, the data of the batches can be combined according to a "common slope, separate intercepts" model for further statistical evaluation.

Calculation of the one-sided upper confidence interval leads to 0.26% at release, initial time 0, and 0.72% at the end of the expected shelf life (36 months). The analogous calculation of the probability for passing a batch which is out of specification by mistake (consumer's risk) as described in example I leads to 0.91% (see Fig. 5). This means that if a level of degradation product of 0.91% is present in one of the batches under investigation, the probability of detecting this batch as being out of specification is 95%. In contrast to example I, this level is very close to the shelf life specification, which is a result of the higher precision of the analytical method.

Table 4

Stability Data (Content of Degradation Product) for Example II

	Content of Degradation Product (%)		
Time (Months)	Batch IV	Batch V	Batch VI
0	0.19	0.26	0.05
3	0.24	0.29	0.09
6	0.26	0.35	0.14
9	0.28	0.39	0.18
12	0.36	0.38	0.20
18	0.42		0.22
24	0.55		

DISCUSSION

The requirements for registration stability batches are prescribed in the ICH guideline (1). The formulation, the manufacturing process, and the primary packaging have to be the same as those that are applied to future batches for marketing. The batch sizes should be at least pilot scale (one batch may be smaller) and the quality of the drug product should be representative for the future product that is intended for marketing. Since these batches represent the final stage of the development and thus are usually available towards the end of

Table 5

Results of ANCOVA Calculation Using the Stability Data Given in Table 4

Source	*dF*	*SS*	*MS*	*F*	*p(F)*
Equality of intercepts	2	0.1368	0.0684	106.98	3.30E-09
Zero slope	1	0.1184	0.1184	185.22	1.83E-09
Error (1)	14	0.0089	0.0006		
Equality of slopes	2	0.0034	0.0017	3.6024	0.0595
Error (2)	12	0.0056	0.0005		

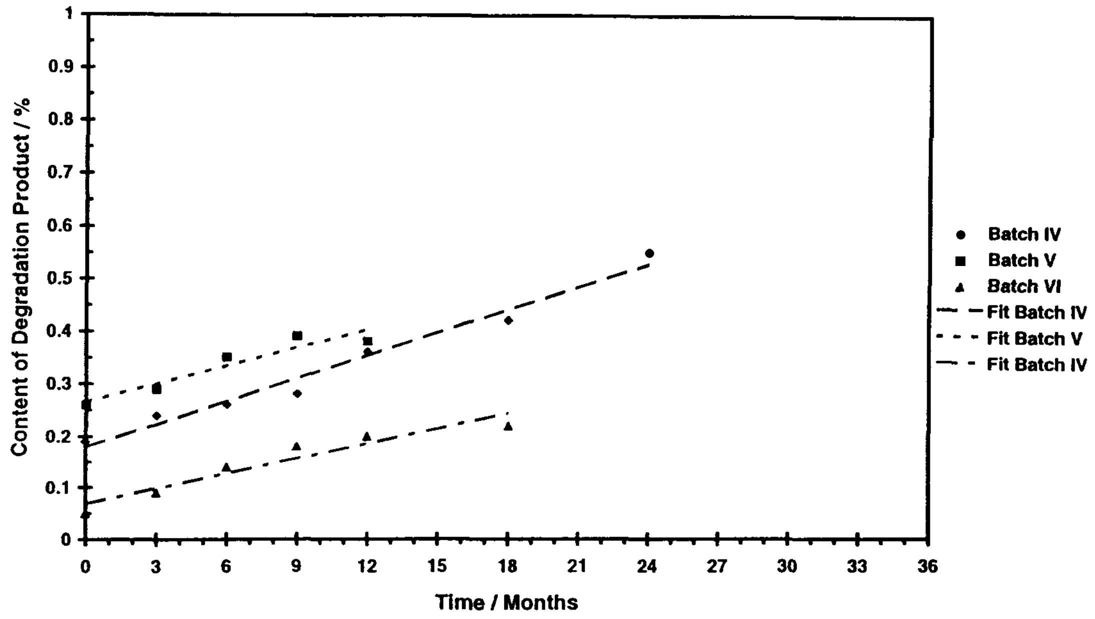

Figure 4. Graphical presentation of the stability data given in Table 4: single values and linear regression fits of the data of the individual batches.

the development, only limited stability data (at least 12 months is required) are available for stipulating specifications. On the other hand, the specifications should allow an acceptable quality of the product for (usually) 36 months. The proposed strategy for setting and justification of specifications successfully solves this problem. First experience with the methodology for specification setting has shown that retrospective analysis of

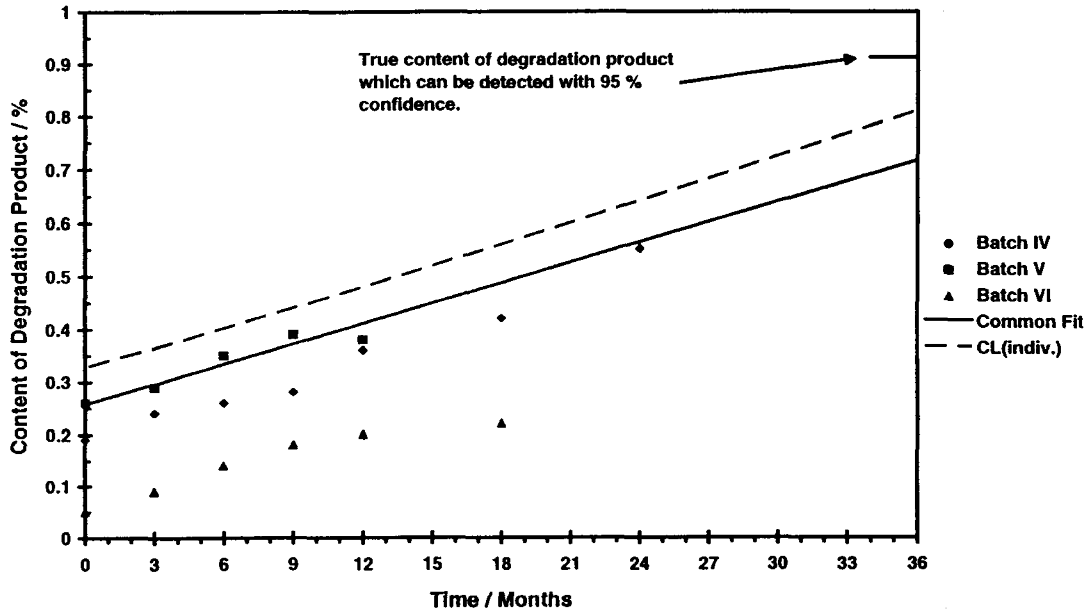

Figure 5. Results of the statistical evaluation of the stability data given in Table 4: single values, common fit with the intercept shifted to the intercept of the worst batch (batch V), one-sided upper confidence limit [CL(ind)] and true content of degradation product which can be detected as out of specification with 95% probability.

Table 6

Results of the Power Calculation Using the Stability Data Given in Table 4

Source	Minimal Differences to Detect[a]	Probability 1 − β of Detecting the Difference[e]	α_{critical} for Decreasing β to 0.05	Model for Combining the Data
Equality of intercepts	0.2%[b]	>0.99	2.22E-16	Common slope
Zero slope	0.01%/month[c]	>0.99	3.83E-08	Slope ≠ zero
Equality of slopes	0.4%[d]	>0.99	0.036515	Separate intercepts

[a] To be detected with β = 0.05.
[b] Difference of one batch in the less favorable direction compared to the others at release.
[c] Difference of common slope from zero (= 0.4% at the end of the expected shelf life).
[d] Difference of one batch at the end of the expected shelf life (= 36 months) compared to the others.
[e] Based on $\alpha_{\text{critical}} = 0.25$.

stability data from already registered products leads to specifications very similar to those agreed upon with the authorities earlier.

The proposed strategy aims at the statistical calculation of specifications for a target shelf life of 36 months by extrapolation of the stability data. This enables the comparison of the quality of future batches intended for marketing with the quality of the registration stability batches and leads to the setting of release as well as shelf-life specifications that are close to the data of the registration stability batches. The methodology is somewhat different from other approaches for calculation of specifications based on the confidence limit of the slope of the regression line (23,24), and from what is described in the ICH guideline (1) as well as in the literature (3–5). The objectives of the latter data evaluation methods are different because they are dealing with the definition of the shelf life for an already given set of specifications.

Since any statistical test is fully described only if both potential errors (producer's risk and consumer's risk) have been taken into account, special attention is given to the calculation of the consumer's risk (second type error) for the ANCOVA calculation as well as for the calculation of the confidence intervals. This forces to define a critical difference for which the probability of detection is calculated or vice versa, a probability is predefined and the critical difference that can be detected is calculated. Especially if specifications for critical quality characteristics (e.g., toxic degradation products, very narrow therapeutic range) are concerned, the assessment of the risk of the consumer to receive material which does not fulfill the specifications should be mandatory.

Of course, safety considerations have the highest priority in the list of factors that affect the setting of specifications. However, fortunately the quality standards in pharmaceutical industry are so high that usually the quality of a drug product is of no safety concern, and thus specification limits are usually not based on safety considerations. Rather, in most cases the observed quality of all registration stability batches throughout their shelf life forms the basis for specification setting. The strategy for specification setting proposed in the present paper should be applicable in the majority of cases. The international quality standards, which are part of the national regulations, provide the frame for this process. Specification limits outside these international standards are usually not accepted by the authorities and require strong justifications.

It should be noted that besides the long-term stability data (the evaluation of which is the subject of the present paper), also the accelerated data have a strong impact on specification setting. According to the ICH guideline (1), 6-month stability data at accelerated conditions have to be within specification for a new pharmaceutical product.

ACKNOWLEDGMENT

The authors wish to thank D. Bates, H. Ludwig, and K. Baumgartner for their ideas and suggestions.

REFERENCES

1. ICH Harmonized Tripartite Guideline, Stability Testing of New Drug Substances and Products, Endorsed by the

ICH Steering Committee at Step 4 of the ICH Process, October 27, 1993.
2. Center for Drugs and Biologics, Food and Drug Administration, U.S. Department of Health and Human Services, Guideline for Submitting Documentation for the Stability of Human Drugs and Biologics, Washington, DC, 1987.
3. J. T. Carstensen, *Drug Stability—Principles and Practices*, Marcel Dekker, New York, 1990.
4. D. J. Schuirmann, in AAPS/FDA Conference, *Proceedings: Stability Guidelines for Testing Pharmaceutical Products: Issues and Alternatives*, Arlington, VA, 1989.
5. N. Delclos, F. Pellerin, M. Aubert, R. Bentejac, C. Bullot, J. L. Colin, A. M. Gallo, P. Grelet, V. Labbe, G. Lenay, F. Malfroid, R. Russotto, and H. Vrinat, STP Pharma Pratiq., 4, 91 (1994).
6. D. L. Bentley, J. Pharm. Sci., 59, 464 (1970).
7. O. L. Davies and H. E. Hundson *Statistics in Pharmacy*, Marcel Dekker, New York, 1981.
8. S. P. King, M. Kung, and H. Fung, J. Pharm. Sci., 73, 657 (1984).
9. V. Hartmann, K. Krummen, G. Schnabel, and H. Bethke, Pharm. Ind., 44, 71 (1982).
10. J. T. Carstensen and E. Nelson, J. Pharm. Sci., 65, 311 (1976).
11. F. Langenbucher, Drug Dev. Ind. Pharm., 17, 165 (1991).
12. S. C. Chow and J. Shao, Stat. Med., 8, 883 (1989).
13. J. J. Chen, J. S. Hwang, and Y. Tsong, J. Biopharm. Stat., 5, 131 (1995).
14. T. E. Norwood, Drug Dev. Ind. Pharm., 12, 553 (1986).
15. O. L. Davies and P. L. Goldsmith, *Statistical Methods in Research and Production*, Longman Scientific & Technical, Essex, 1988.
16. A. A. Afifi and S. P. Azen, *Statistical Analysis: A Computer Oriented* Approach, Academic Press, New York, 1979.
17. S. J. Ruberg and J. W. Stegeman, Biometrics, 47, 1059 (1991).
18. P. B. Patnaik, Biometrika, 36, 202 (1949).
19. H. Peil and V. Haeselbarth, Drug Res., 35, 1489 (1985).
20. J. Zar, *Biostatistical Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
21. W. G. Cochran and G. M. Cox, *Experimental Designs*, Wiley & Sons, New York; Chapman & Hall, London, 1957.
22. N. R. Drapper and H. Smith, *Applied Regression Analysis*, Wiley & Sons, New York, 1966.
23. R. C. Kohberger, in *Biopharmaceutical Statistics for Drug Development* (K. E. Peace, ed.), Marcel Dekker, New York, 1988.
24. P. V. Allen, G. R. Dukes, and M. E. Gerger, Pharm. Res., 8, 1210 (1991).