Journal of Chromatography B, 877 (2009) 2262–2269

Contents lists available at ScienceDirect

Journal of Chromatography B

journal homepage: www.elsevier.com/locate/chromb

# Statistical methods for assessing long-term analyte stability in biological matrices☆

David Hoffman[a,*], Robert Kringle[a], Julia Singer[b], Stuart McDougall[c]

[a] Preclinical and Research Biostatistics, sanofi-aventis, Bridgewater, NJ, USA
[b] Global Clinical Biostatistics, Baxter Bioscience, Vienna, Austria
[c] Global Metabolism and Pharmacokinetics, sanofi-aventis, Alnwick, UK

ARTICLE INFO

*Article history:*
Received 7 May 2008
Accepted 22 August 2008
Available online 29 August 2008

*Keywords:*
Stability
Equivalence
Regression
Bivariate mixed model
Nested errors regression
Bioanalysis

ABSTRACT

The objective of a long-term stability experiment is to confirm analyte stability in a given biological matrix, encompassing the duration of time from sample collection to sample analysis for a clinical or preclinical study. While long-term analyte stability has been identified as a key component of bioanalytical method validation, current regulatory guidance provides no specific recommendations regarding the design and analysis of such experiments. This paper reviews and evaluates various experimental designs, data analysis methods, and acceptance criteria for the assessment of long-term analyte stability. Statistical equivalence tests based on linear regression techniques are advocated. Both a nested errors and bivariate mixed model regression approach are suitable for application to long-term stability assessment, and control the risk of falsely concluding stability.



## 1. Introduction

Bioanalytical methods for the quantitative determination of drugs and their metabolites in biological matrices play a critical role in the evaluation and interpretation of bioequivalence, bioavailability, pharmacokinetic, and toxicokinetic studies. The quality and integrity of these studies is dependent upon the quality and integrity of the underlying bioanalytical data. As such, well-characterized and fully validated bioanalytical methods are essential to ensure the safety and efficacy of pharmaceuticals.

A key aspect in the validation of bioanalytical methods is the assessment of analyte (drug and/or metabolite) stability in biological matrices [1–9]. Various stability evaluations are performed during method validation, typically including: freeze–thaw stability, processed sample stability, stock solution stability, short-term temperature stability, and long-term stability. The proper assessment of long-term analyte stability poses particular difficulties, and is the subject of this paper.

The objective of a long-term stability experiment is to confirm analyte stability in a given biological matrix, encompassing the duration of time from sample collection to sample analysis for a clinical or preclinical study. The stability experiment should mimic the conditions under which the study samples will be collected, stored and processed [1]. Moreover, stability should be assessed in each matrix (e.g. plasma, urine, etc.) and species (e.g. human, dog, etc.) in which the analyte is to be quantified. This assessment is necessary to confirm that degradation after sample collection has not occurred, thus giving credibility to the final study data.

While long-term analyte stability has been identified as a key component of bioanalytical method validation, current regulatory guidance provides no specific recommendations regarding the design and analysis of such experiments. The purpose of this paper is to review and evaluate various experimental designs, data analysis methods, and acceptance criteria for the assessment of long-term analyte stability.

## 2. Experimental design

Long-term analyte stability assessment is performed by preparing stability samples at two or more nominal concentrations [1]. Typically, these stability samples are prepared by spiking control (blank) biological matrix with the analyte of interest. These stability pools are then transferred into individual storage tubes representative of those intended for the long-term storage of study samples, and are stored (frozen) under the conditions that will be used for the study samples. Long-term stability is then assessed by analysis of the stability samples over an appropriate time frame (i.e. suffi-

☆ This paper is part of a special issue entitled "Method Validation, Comparison and Transfer", guest edited by Serge Rudaz and Philippe Hubert.

* Corresponding author.
*E-mail address:* david.hoffman@sanofi-aventis.com (D. Hoffman).

1570-0232/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jchromb.2008.08.015

cient to encompass or exceed the storage time anticipated for study samples).

We consider two proposed experimental designs for long-term stability assessment: a "standard" design and a "concurrent control" design.

The standard design is defined as follows. Stability samples are prepared as described above. Immediately following sample preparation or shortly thereafter, replicate samples are analyzed against a freshly prepared calibration curve. This can assess the accuracy of the spiked sample preparations (i.e. confirm the nominal concentration). The remaining samples are then stored as described above. At pre-specified timepoints, replicate frozen stability samples are thawed and analyzed against freshly prepared calibration curves.

The concurrent control design is identical to that of the standard design, with one modification: at each pre-specified timepoint, replicate "control" samples are analyzed concurrently with the thawed stability samples against the same freshly prepared calibration curve. The concurrent control samples can be prepared in one of two manners:

(i) At each pre-specified timepoint, replicate control samples are freshly prepared at the same nominal concentration and analyzed against the same freshly prepared calibration curve.
(ii) At the time of initial stability sample preparation, the samples are divided into two subsets. The first subset (i.e. stability samples) is stored at the temperature intended for study samples, as described previously. The second subset (i.e. control samples) is stored under temperatures less than −130 °C (e.g. in liquid nitrogen or other suitable freezer). At each pre-specified timepoint, replicates of both the stability samples and control samples are analyzed against the same freshly prepared calibration curve.

With either the standard or concurrent control design, calculated analyte concentrations are subject to both within-run (intra-batch) and between-run (inter-batch) random variability intrinsic to the analytical method. The use of concurrent controls is intended to eliminate or minimize sources of between-run variability (e.g. calibration error) by including control samples in the same analytical run as the stability samples [10].

Theoretical calculations based on the Arrhenius equation, as well as published literature [11,12], indicate storage temperatures of −130 °C or below ensure stability even for unstable analytes. Thus, samples prepared and stored as described in (ii) above may be suitable for use as concurrent controls. It is recognized that the use of concurrent control samples prepared in this manner has been the subject of some debate [6]. This will not be addressed in the current paper, except to note that the stability of such concurrent control samples should be (at minimum) informally verified via graphical inspection and/or descriptive statistics. The use of concurrent controls which exhibit degradation similar to that of stability samples over the storage time is improper and will result in an inflated risk of falsely concluding stability. Freshly prepared control samples (as described in (i) above), by definition, will not exhibit degradation; however, this introduces random variability arising from the preparation of different fresh control samples at each timepoint.

It should also be noted that stability samples could be prepared by collecting or pooling incurred samples from dosed subjects, rather than by spiking control matrix. However, the "nominal" concentration in such incurred samples will be unknown and must be estimated from observed data. Appropriate data analysis procedures for such samples must properly account for this additional source of variability. This is a topic for future investigation and will not be considered further in the current paper.

## 3. Data analysis methods

Specific recommendations for assessing long-term analyte stability are not defined in current regulatory guidance documents. However, a common approach is to evaluate long-term stability using the same criteria typically applied for evaluating accuracy and precision of QC samples [4,7]. This approach ("4–6–15" rule) would require at least two thirds of the individual stability samples at each timepoint to be within, say, 15% of the respective nominal concentration. Another approach ("observed mean" rule) is to require that the observed mean concentration at each timepoint be within 15% of the nominal concentration [7]. While both the 4–6–15 and observed mean rules are easy to implement, they yield unknown and uncontrolled risks of incorrect stability decisions (failing to detect a truly unstable analyte, or falsely concluding instability). The deficiencies of these approaches are well documented [13,14]. As such, neither approach (nor other nonstatistical or ad-hoc rules) will be considered further.

When assessing long-term analyte stability, it is reasonable to assume that the risk of falsely concluding stability should be controlled to a small probability (say, 5%), when the true degradation is bioanalytically relevant (say, 15%). Thus, statistical equivalence tests provide a logical framework for stability assessment [10,11,14]. A statistical equivalence test is based on the set of hypotheses:

$$H_0 : \Delta \leq -D \text{ or } \Delta \geq D$$

$$\text{vs. } H_A : -D < \Delta < D$$

where $\Delta$ is the true analyte degradation and $D$ is the amount of true analyte degradation which is considered bioanalytically relevant (say, 15%). Rejection of the null hypothesis $H_0$ leads to a conclusion of stability. An $\alpha$-level equivalence test is typically conducted by constructing a two-sided $(100 - 2\alpha)$% confidence interval for $\Delta$; if the confidence interval lies entirely within the acceptance limits $(-D, D)$, the null hypothesis $H_0$ is rejected. This equivalence testing approach controls the risk of falsely rejecting $H_0$ (i.e. falsely concluding stability when the true degradation is $D$) at $\alpha$% (say, 5%). All statistical methods considered for further investigation in this paper are based on equivalence tests at the $\alpha = 5$% level.

### 3.1. Simple linear regression

The problem of assessing long-term analyte stability in biological matrices is similar to that of determining shelf-life for drug product. For determination of drug product shelf-life, many quantitative chemical attributes (e.g. assay and degradation product) are assumed to follow zero-order kinetics during long-term storage [15]. Thus, linear regression analysis is generally considered an appropriate approach for evaluating long-term stability data and the performance characteristics of such an approach have been well examined [16,17]. Similar reasoning suggests that the relationship between analyte concentrations in biological matrices and storage time can be represented by a linear function (perhaps after appropriate data transformation, if necessary). Linear regression techniques have also been previously proposed for assessing short-term stability of analytes in biological matrices [14]. As such, the use of linear regression techniques for assessing long-term analyte stability in biological matrices is advocated.

The simple linear regression approach consists of regressing the calculated stability sample analyte concentrations on storage time via the following model:

$$y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon_{ij} \quad (1)$$

where $y_{ij}$ is the calculated analyte concentration for the $j$th stability sample replicate at the $i$th timepoint, $x_i$ is the $i$th timepoint, and $\varepsilon_{ij}$ is the random error for $y_{ij}$, with $\varepsilon_{ij}$ assumed to be independently and normally distributed with mean zero and variance $\sigma_E^2$.

At any fixed timepoint $x$, a 90% two-sided confidence interval for the mean analyte concentration is constructed from the fitted regression model: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. The analyte is considered stable at a given timepoint if the 90% two-sided confidence interval lies entirely within pre-specified acceptance limits (say, ±15% of the nominal concentration).

### 3.2. Nested errors regression

The simple linear regression model above ignores the between-run (inter-batch) random errors induced by assay calibration at each timepoint. The simple linear model assumes that all calculated analyte concentrations are statistically independent of each other. This assumption will not be satisfied with long-term stability data, as calculated concentrations obtained against a common calibration curve will be correlated (i.e. by the between-run random error at each timepoint).

A nested errors linear regression model can appropriately account for both between-run and within-run random errors inherent in long-term stability data [18]. The nested errors regression approach consists of regressing the calculated stability sample analyte concentrations on storage time via the following model:

$$y_{ij} = \beta_0 + \beta_1 x_i + \gamma_i + \varepsilon_{ij} \quad (2)$$

where $y_{ij}$ is the calculated analyte concentration for the $j$th stability sample replicate at the $i$th timepoint, $x_i$ is the $i$th timepoint, $\gamma_i$ is the random error associated with the $i$th timepoint, and $\varepsilon_{ij}$ is the random error for $y_{ij}$. The random errors $\gamma_i$ and $\varepsilon_{ij}$ are assumed to be independently and normally distributed with means zero and variances $\sigma_B^2$ and $\sigma_E^2$, respectively. These variances, $\sigma_B^2$ and $\sigma_E^2$, correspond to the between-run and within-run variability of the analytical method, respectively.

At any fixed timepoint $x$, a 90% two-sided confidence interval for the mean analyte concentration is constructed from the fitted regression model: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. The analyte is considered stable at a given timepoint if the 90% two-sided confidence interval lies entirely within the pre-specified acceptance limits.

### 3.3. Bivariate mixed model regression

Both the simple linear and nested errors regression approaches above are applicable to long-term stability studies utilizing the standard experimental design described previously. Neither approach allows for inclusion of data from concurrent control samples, which may minimize or eliminate the impact of between-run errors in the stability assessment.

One simple approach to incorporate data from concurrent control samples would be to "normalize" the stability sample analyte concentrations by the mean control sample analyte concentration at each timepoint. However, if the between-run and within-run random errors are assumed to follow a normal distribution, then the "normalized" random errors (i.e. ratio of errors) will be non-normally distributed. Furthermore, this simple approach presupposes a high degree of correlation between the stability and control samples at each timepoint. While this should be typically expected (and is the ideal outcome), it is possible that the stability and control samples may exhibit poor correlation (e.g. due to poor precision when spiking fresh control samples at each timepoint or possible matrix effects arising from storage at temperatures below −130 °C). In such cases, simple normalization will be detrimental, causing inflated variability and resulting in poorer precision of stability estimates.

A more flexible approach to incorporate data from concurrent control samples is to jointly model the stability sample and control sample data in a bivariate mixed model. The bivariate mixed model regression approach consists of jointly regressing the calculated stability sample and control sample analyte concentrations on storage time via the following model:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_i + \gamma_i + \varepsilon_{ij} \\ z_{ik} &= \beta_0 + \delta_i + \xi_{ik} \end{aligned} \quad (3)$$

where $y_{ij}$ is the calculated analyte concentration for the $j$th stability sample replicate at the $i$th timepoint, $z_{ik}$ is the calculated analyte concentration for the $k$th control sample replicate at the $i$th timepoint, $x_i$ is the $i$th timepoint, $\gamma_i$ is the random error associated with the $i$th timepoint for the stability samples, $\delta_i$ is the random error associated with the $i$th timepoint for the control samples, $\varepsilon_{ij}$ is the random error for $y_{ij}$, and $\xi_{ik}$ is the random error for $z_{ij}$.

The within-run random errors $\varepsilon_{ij}$ and $\xi_{ik}$ are assumed to be independently and normally distributed with means zero and variances $\sigma_{E1}^2$ and $\sigma_{E2}^2$, respectively. These variances, $\sigma_{E1}^2$ and $\sigma_{E2}^2$, correspond to the within-run variability of the stability and control samples, respectively.

The between-run random errors $\gamma_i$ and $\delta_i$ are assumed to follow a bivariate normal distribution with means zero and covariance matrix $\Sigma$ given by:

$$\Sigma = \begin{pmatrix} \sigma_{B1}^2 & \rho\sigma_{B1}\sigma_{B2} \\ \rho\sigma_{B1}\sigma_{B2} & \sigma_{B2}^2 \end{pmatrix}$$

The variances $\sigma_{B1}^2$ and $\sigma_{B2}^2$ correspond to the between-run variability of the stability and control samples, respectively. The correlation parameter $\rho$ corresponds to the correlation of the between-run random errors for the stability and controls samples analyzed at a given timepoint (i.e. against a common calibration curve).

Note that the model could also be simplified by reasonably assuming the within-run and between-run variances to be identical for both the stability and control samples (i.e. $\sigma_{E1}^2 = \sigma_{E2}^2 = \sigma_E^2$ and $\sigma_{B1}^2 = \sigma_{B2}^2 = \sigma_B^2$).

As with the simple linear and nested error approaches, a 90% two-sided confidence interval for the mean analyte concentration at any fixed timepoint $x$ is constructed from the fitted regression model: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. The analyte is considered stable at a given timepoint if the 90% two-sided confidence interval lies entirely within the pre-specified acceptance limits.

## 4. Results

The performances of the simple linear, nested errors, and bivariate mixed model regression approaches were evaluated via simulation techniques. For all simulations performed, the following conditions were assumed:

(1) sampling timepoints at 0, 3, 6, 9, 12, 18, and 24 months;
(2) true within-run and between-run variances, $\sigma_E^2$ and $\sigma_B^2$, yielding a true total variance ($\sigma_{TOT}^2 = \sigma_B^2 + \sigma_E^2$) equivalent to a 10% total coefficient of variation (%CV);
(3) normally distributed within-run and between-run random errors.

All simulations were performed using SAS software (version 9.1), and all regression models fit with the MIXED procedure.

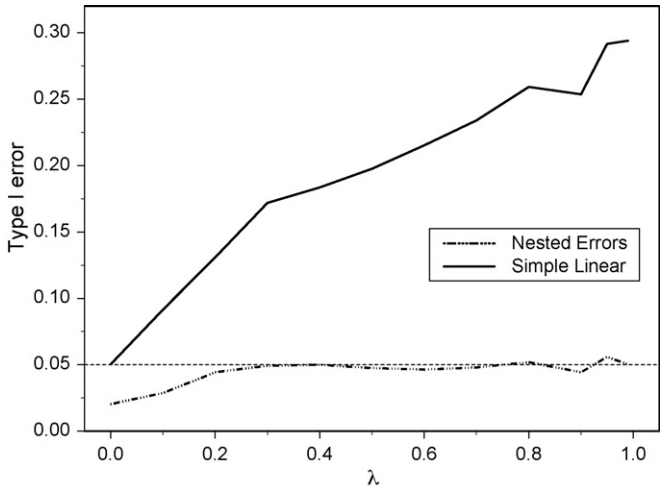


**Fig. 1.** Type I error rate versus $\lambda = \sigma_B^2/(\sigma_B^2 + \sigma_E^2)$ for simple linear and nested errors regression approaches. Reference line at nominal type I error rate of 0.05.

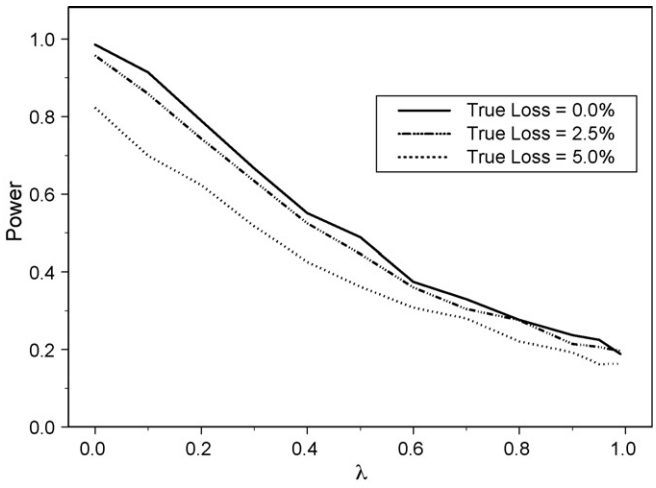


**Fig. 2.** Power to conclude stability versus $\lambda = \sigma_B^2/(\sigma_B^2 + \sigma_E^2)$ for nested errors regression approach, for various true loss.

### 4.1. Simple linear regression approach

A standard experimental design with six stability sample replicates at each timepoint was assumed. Analyte concentrations were assumed to degrade linearly over time, with a true loss of 15% at the 24-month timepoint.

Let the proportion of total variability ($\sigma_{TOT}^2$) due to between-run variability ($\sigma_B^2$) be denoted by $\lambda$. For various values of $\lambda$ (ranging from 0.0 to 0.99), 2500 datasets were simulated. The simple linear regression model shown in Eq. (1) was fit to each simulated dataset and the 90% two-sided confidence interval for the mean analyte concentration at the 24-month timepoint was formed. The type I error of the simple linear regression approach was then estimated as the proportion of confidence intervals (out of 2500 simulated datasets) which were entirely contained within ±15% acceptance limits. Note that the type I error can be viewed as the probability of falsely concluding stability (i.e. the true loss is at the bioanalytically relevant limit of 15%, but the 90% two-sided confidence interval lies entirely within the acceptance limits).

Fig. 1 gives the type I error rate of the simple linear regression approach as a function of $\lambda = \sigma_B^2/(\sigma_B^2 + \sigma_E^2)$.

The results in Fig. 1 clearly indicate the inadequacy of the simple linear regression approach. Recall that the probability of falsely concluding stability should be controlled at 5% for a true loss of 15% when using statistical equivalence tests as described earlier. However, the type I error rate of the simple linear regression approach increases dramatically with $\lambda$, and is nearly 30% when $\lambda$ is close to 1.0. This is because the simple linear regression model ignores the correlated nature of the data. Note that when $\lambda = 0$, the between-run variability is zero and the analyte concentrations are uncorrelated. In this scenario, the simple linear regression approach is appropriate and the type I error rate is maintained at 5% (as shown in Fig. 1). For typical long-term stability studies, $\lambda$ is likely to be quite large (say, $\lambda > 0.50$). Thus, the simple linear regression approach is a poor choice for application in long-term stability analyses.

### 4.2. Nested errors regression approach

The type I error of the nested errors regression approach was estimated using the same simulated data described above. The nested errors regression model shown in Eq. (2) was fit to each simulated dataset and the 90% two-sided confidence interval for the mean analyte concentration at the 24-month timepoint was formed. The type I error of the nested errors regression approach was then estimated as the proportion of confidence intervals (out of 2500 simulated datasets) which were entirely contained within ±15% acceptance limits.

Fig. 1 gives the type I error rate of the nested errors regression approach as a function of $\lambda = \sigma_B^2/(\sigma_B^2 + \sigma_E^2)$.

The results in Fig. 1 indicate that the nested errors regression approach controls the risk of falsely concluding stability, regardless of the value of $\lambda$. Unlike the simple linear regression approach, the nested errors approach appropriately accounts for both between-run and within-run random errors and thus maintains control of the type I error rate.

The nested errors regression approach clearly controls the risk of falsely concluding stability. It is also of interest to assess the power of the nested errors regression approach to correctly conclude stability (i.e. the probability of concluding stability for a truly stable analyte). As before, a standard experimental design with six stability sample replicates at each timepoint was assumed.

Analyte concentrations were assumed to degrade linearly over time, with a true loss of 0%, 2.5%, or 5% at the 24-month timepoint. Various values of $\lambda$, ranging from 0.0 to 0.99, were considered. For each combination of true loss and $\lambda$, 2500 datasets were simulated. The nested errors regression model was fit to each simulated dataset and the 90% two-sided confidence interval for the mean analyte concentration at the 24-month timepoint formed. The power of the nested errors regression approach was then estimated as the proportion of confidence intervals (out of 2500 simulated datasets) which were entirely contained within ±15% acceptance limits.

Fig. 2 gives the power of the nested errors regression approach to conclude stability versus $\lambda = \sigma_B^2/(\sigma_B^2 + \sigma_E^2)$, for true losses of 0, 2.5, and 5%.

The results in Fig. 2 show that the power of the nested errors regression approach to correctly conclude stability decreases dramatically with increasing $\lambda$. Even for analytes with no true loss, the power to conclude stability is poor for $\lambda > 0.20$. The power of the nested errors regression approach could be improved by utilizing a larger sample size (i.e. more timepoints and/or replicates per timepoint). However, the sample size considered in the simulation is reasonably large (six replicates at each of seven timepoints) and much larger sample sizes may be prohibitive in practice. The power will also increase as the true total %CV decreases. Recall the true total %CV is 10% for the simulated data (this is likely quite typical for long-term stability data). Yet regardless of the sample size or true

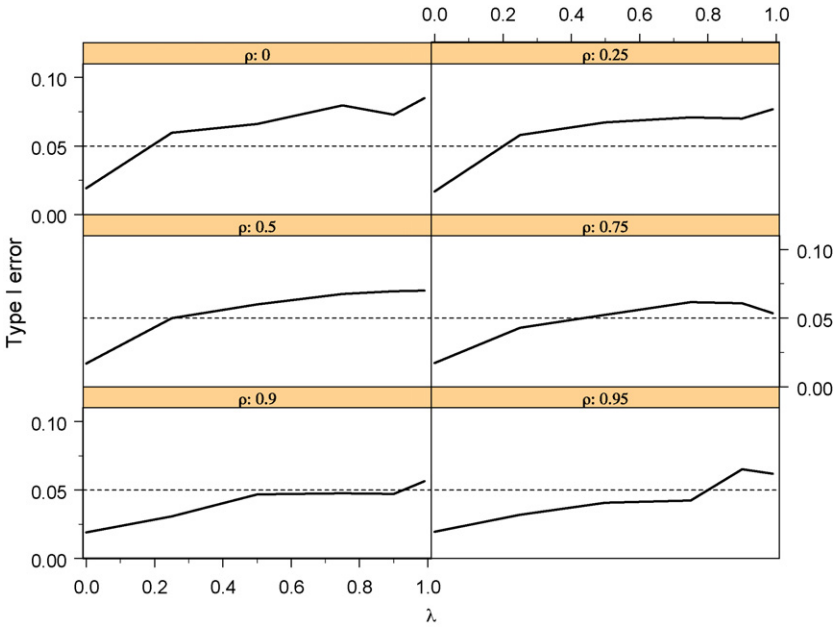


**Fig. 3.** Type I error rate versus $\lambda = \sigma_B^2/(\sigma_B^2 + \sigma_E^2)$ for bivariate mixed model regression approach, for various correlation $\rho$. Reference line at nominal type I error rate of 0.05.

total %CV, the power of the nested errors regression approach will decrease as the relative magnitude of the between-run variability increases (i.e. as $\lambda$ increases).

It should also be noted that the power of the nested errors regression approach may be improved by a modification to the standard experimental design defined previously: at each timepoint, the stability sample replicates are analyzed over multiple analytical runs rather than in a single run. Performing multiple independent runs at each timepoint increases the precision of the stability estimates and thus increases the power to correctly conclude stability. Note that this assumes complete independence of the analytical runs at each timepoint, which may be unrealistic in practice. Failure to meet this assumption (e.g. due to variability induced by assay drift over time, etc.) would mitigate any increases in power.

### 4.3. Bivariate mixed model regression approach

A concurrent control experimental design with six stability replicates and six control sample replicates (at timepoints subsequent to 0 month) at each timepoint was assumed. For simplicity, the stability and control samples were assumed to have identical true within-run variance ($\sigma_{E1}^2 = \sigma_{E2}^2 = \sigma_E^2$) and identical true between-run variance ($\sigma_{B1}^2 = \sigma_{B2}^2 = \sigma_B^2$). Analyte concentrations for the stability samples were assumed to degrade linearly over time, with a true loss of 15% at the 24-month timepoint. Analyte concentrations for the control samples were assumed to have no true degradation over time.

As before, let the proportion of total variability ($\sigma_{TOT}^2$) due to between-run variability ($\sigma_B^2$) be denoted by $\lambda$. Let $\rho$ be the correlation of the between-run random errors for the stability and control samples. Values of $\lambda$ ranging from 0.0 to 0.99 and $\rho$ values of 0, 0.25, 0.50, 0.75, 0.90, and 0.95 were considered. For each combination of $\lambda$ and $\rho$, 2500 datasets were simulated. The bivariate mixed model shown in Eq. (3) was fit to each simulated dataset and the 90% two-sided confidence interval for the mean stability sample analyte concentration at the 24-month timepoint formed. The type I error of the bivariate mixed model regression approach was then estimated as the proportion of confidence intervals (out of 2500 simulated datasets) which were entirely contained within ±15% acceptance limits.

Fig. 3 shows the type I error rate for the bivariate mixed model regression approach as a function of $\lambda$, for various correlation parameter $\rho$.

Fig. 3 indicates that the bivariate mixed model regression approach generally controls the risk of falsely concluding stability. For large values of $\lambda$ (i.e. large between-run variability) and small values of $\rho$ (i.e. poor correlation of stability and control samples), the type I error rate is slightly inflated above the nominal 5%. However, the type I error rate is never more than approximately 8%. For more typical scenarios ($\rho > 0.50$), the type I error rate is at (or below) the nominal 5% level.

As with the nested errors regression approach, the bivariate mixed model regression approach controls the risk of falsely concluding stability. Now it is of interest to assess the power of the bivariate mixed model regression approach to correctly conclude stability.

As above, a concurrent control experimental design with six stability sample replicates and six control sample replicates at each timepoint was assumed. The true loss for both stability and control samples was assumed to be 0%. Various values of $\lambda$ and $\rho$ were considered as previously. For each combination of $\lambda$ and $\rho$, 2500 datasets were simulated and the power of the bivariate mixed model regression approach estimated as described previously. For comparative purposes, the nested errors regression model was also fit to each simulated dataset (ignoring data from the control samples) and the power of the nested errors regression approach estimated as well.

Fig. 4 shows the power of the bivariate mixed model and nested errors regression approaches to conclude stability versus $\lambda$, for various correlation $\rho$.

Fig. 4 illustrates the increase in the power to correctly conclude stability which can be obtained with the concurrent control design. Note that when the correlation $\rho$ is small, the power of the nested errors and bivariate mixed model regression approaches is roughly equal. In these cases, the correlation between the stability and control samples is poor, and the control samples do not reduce the impact of between-run random variability. However, as the correlation $\rho$ increases, the power of the bivariate mixed model regression approach increases accordingly. For values of $\rho \geq 0.50$, the increase in power is substantial, as the control samples reduce the impact of

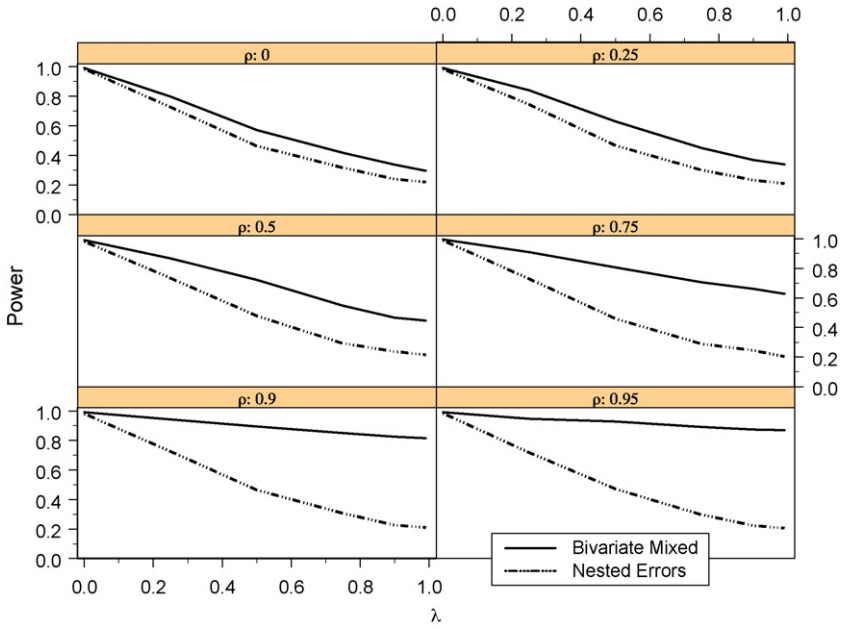


**Fig. 4.** Power to conclude stability versus $\lambda = \sigma_B^2/(\sigma_B^2 + \sigma_E^2)$ for bivariate mixed model and nested errors regression approaches, for various correlation $\rho$. True loss is 0%.

between-run variability and increase the precision of the stability estimates.

## 5. Example

The nested errors and bivariate mixed model regression approaches are illustrated by application to data from an actual long-term stability experiment utilizing a concurrent control experimental design.

A plasma pool was spiked at 200 ng/mL of analyte and six samples were analyzed immediately following pool preparation. The remaining plasma pool was divided into two subsets (stability samples and control samples). Stability samples were stored at −20 °C and control samples at less than −130 °C. Six stability sample replicates and six control sample replicates were then thawed and analyzed against a freshly prepared calibration curve after 1, 3, 6, 9, 12, 18, and 24 months of storage. The raw concentration data are given in Table 1 (note that four observations were missing due to analytical issues and are indicated by '–' in the table).

Both the nested errors and bivariate mixed model regression approaches were applied to the data. The nested errors model was fit using only calculated concentrations from stability samples, while the bivariate mixed model was fit using calculated concentrations from both stability and control samples. Fig. 5 shows the fitted nested errors regression model with two-sided 90% confidence interval for the mean analyte concentration. Fig. 6 shows the fitted bivariate mixed model with two-sided 90% confidence interval for the mean analyte concentration. Note that ±15% acceptance limits correspond to (170, 230) ng/mL.

Fig. 5 indicates substantial between-run variability in the calculated concentrations of the stability samples. This variability is reflected in the width of the two-sided 90% confidence interval about the fitted regression line. The estimated proportion of variability due to between-run variability based on the fitted nested errors regression model is $\hat{\lambda} = 0.85$, with an estimated total %CV of 9.8%. At the 24-month timepoint, the two-sided 90% confidence interval for the mean analyte concentration is (162, 216) ng/mL, which falls slightly outside the acceptance limits of (170, 230) ng/mL. Thus, with the nested errors regression approach, we cannot conclude the analyte is stable at 24 months.

Fig. 6 shows good correlation between the stability and control samples. The estimated correlation of the stability sample and control sample between-run random errors based on the fitted bivariate mixed model is $\hat{\rho} = 0.93$. This strong correlation dramatically reduces the impact of the between-run random variability on the precision of the stability estimates and is reflected in the nar-

**Table 1**
Calculated concentrations (ng/mL)

| Month | Fresh samples | | | | | |
|---|---|---|---|---|---|---|
| | Rep1 | Rep2 | Rep3 | Rep4 | Rep5 | Rep6 |
| 0 | 192 | 204 | 196 | 204 | 208 | 202 |

| Month | Stability samples (−20 °C) | | | | | | Control samples (<−130 °C) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rep1 | Rep2 | Rep3 | Rep4 | Rep5 | Rep6 | Rep1 | Rep2 | Rep3 | Rep4 | Rep5 | Rep6 |
| 1 | 220 | 223 | 214 | 219 | 209 | 217 | 221 | 219 | 222 | 219 | 210 | 215 |
| 3 | 188 | 185 | 192 | 187 | 185 | 194 | 190 | 200 | 194 | 196 | 194 | 191 |
| 6 | 167 | 147 | 141 | 180 | – | – | 172 | 176 | 177 | 174 | 175 | 172 |
| 9 | 188 | 200 | 183 | 183 | 189 | 196 | 198 | 193 | 191 | 194 | 195 | 196 |
| 12 | 179 | 180 | 173 | 197 | 183 | 182 | 189 | 182 | 179 | 176 | 176 | – |
| 18 | 183 | 179 | 188 | 192 | 188 | 193 | 198 | 197 | 195 | 195 | 194 | 201 |
| 24 | 210 | 200 | 199 | 201 | 203 | 207 | 199 | 201 | 198 | 193 | 199 | – |

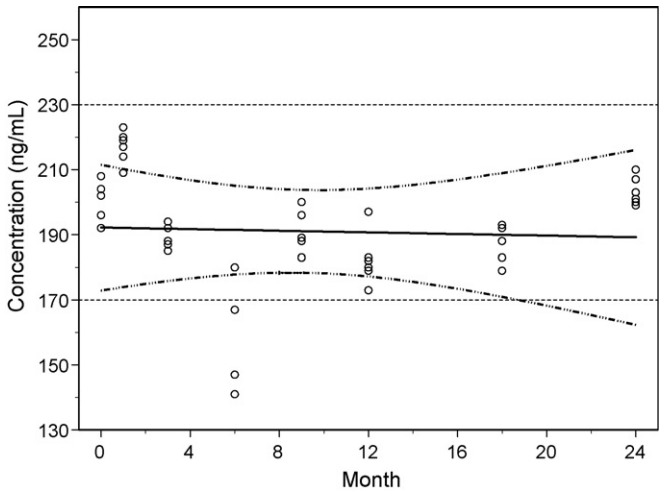


**Fig. 5.** Fitted nested errors regression model with two-sided 90% confidence interval. Calculated concentrations for stability samples given by open circles. Acceptance limits shown at (170, 230) ng/mL.

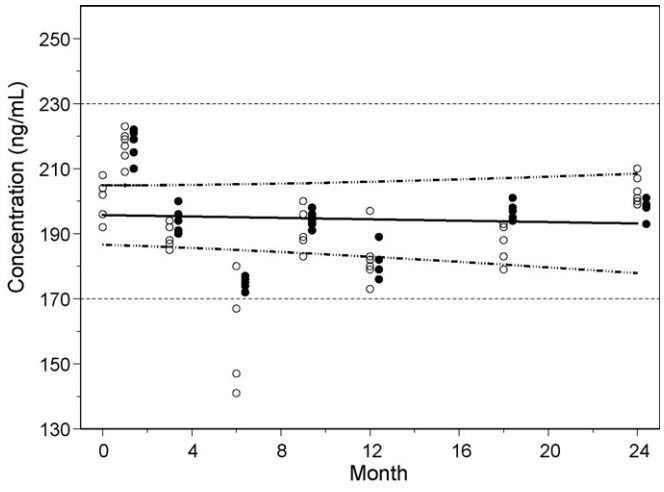


**Fig. 6.** Fitted bivariate mixed model with two-sided 90% confidence interval. Calculated concentrations for stability samples given by open circles and for control samples by closed circles. Acceptance limits shown at (170, 230) ng/mL.

row confidence bounds about the fitted regression line. Note that at the 24-month timepoint, the two-sided 90% confidence interval for the mean analyte concentration is (178, 208) ng/mL. This interval lies entirely within the acceptance limits (170, 230) ng/mL and we can conclude the analyte is stable at 24 months.

## 6. Conclusions

Current regulatory guidance provides no specific recommendations for the design and analysis of long-term stability experiments, and acceptance criteria based on commonly used ad-hoc rules (such as "4–6–15" rule) do not control the risks of incorrect stability decisions. There is a clear need for statistically sound experimental design and data analysis procedures which control the risk of falsely concluding stability (for truly unstable analytes) and provide reasonable power to correctly conclude stability (for truly stable analytes).

Various linear regression techniques for the analysis of long-term stability data were proposed and evaluated. Simple linear regression, commonly used for the determination of drug product shelf-life, is a poor choice for assessing long-term analyte stability in biological matrices. Simple linear regression does not account for between-run sources of variability inherent in long-term stability data and thus fails to control the risk of falsely concluding stability.

Both a nested errors and bivariate mixed model regression approach properly account for between-run and within-run random variability, and control the risk of falsely concluding stability. The nested errors approach can suffer from poor power to correctly conclude stability when the between-run variability is large. However, analyzing stability samples over multiple independent analytical runs (rather than in a single run) at each timepoint may improve the power of the nested errors regression approach. The bivariate mixed model approach incorporates data from control samples analyzed concurrently with stability samples at each timepoint. When the stability and control samples exhibit a high degree of correlation, the bivariate mixed model approach yields stability estimates with increased precision and thus greater power to correctly conclude stability. Both the nested errors and bivariate mixed model regression approaches can be implemented by statistical software packages, such as SAS. Representative SAS code is provided in the Appendix A.

It should be noted that both the nested errors and bivariate mixed model regression approaches offer improved performance (relative to common ad-hoc rules or the simple linear regression approach) for little added cost or resource expenditure. The nested errors regression approach simply utilizes data that is typically generated during a standard long-term stability experiment, but provides an appropriate data analysis model which controls the risk of falsely concluding stability. The bivariate mixed model approach can offer a substantial increase in the power to correctly conclude stability, though it requires the use of a concurrent control experimental design. However, this approach could be implemented with freshly prepared QC samples that are typically included in each run for in-process monitoring. Even if samples stored at less than −130 °C are utilized as concurrent controls, the cost of analyzing these additional samples in each run is likely minimal.

The use of linear regression techniques requires the assumption of zero-order stability kinetics during long-term storage, which may be violated (e.g. enzymatic degradation). The assumption of linearity should be verified by examination of the fitted regression model (e.g. graphical inspection of residuals, etc.). If the relationship between the analyte concentrations and storage time is clearly nonlinear (or cannot be linearized by appropriate data transformation), then nonlinear regression techniques may be more appropriate. This may be a topic for further investigation.

## Appendix A

### A.1. SAS code for nested errors regression model

A useful data format for fitting the nested errors regression model using the SAS MIXED procedure is given in the dataset `nested` (excerpt of raw data shown). The variables `Conc` and `Time` refer to the observed concentrations and timepoints, respectively. The variable `Time_Random` is a duplicate of the `Time` variable, to model the nested structure of the random errors.

```
DATA nested;
INPUT Time Time_Random Conc;
CARDS;
0  0  204
:  :  :
24  24  207
;
```

Sample MIXED code to fit the nested errors regression model is given by:

```
PROC MIXED data=nested;
CLASS Time_Random;
MODEL Conc=Time / ddf=6;
RANDOM Time_Random;
ESTIMATE 'Mean Conc @ Time=24' Int 1 Time 24/ cl alpha=0.10;
```

Note that the degrees of freedom should be directly specified in the MODEL statement, as the MIXED procedure may otherwise overestimate the degrees of freedom. The proper degrees of freedom will generally be equal to the total number of timepoints (or independent analytical runs) minus 2. The MIXED code above assumes an experiment with 8 timepoints, as in the real example given in the text (i.e. timepoints at 0, 1, 3, 6, 9, 12, 18, and 24 months).

### *A.2. SAS code for bivariate mixed model*

A useful data format for fitting the bivariate mixed model using the SAS MIXED procedure is given in the dataset `bivariate` (excerpt of raw data shown). The variables `Conc, Time, Time_Random` are as described before. The variable `Sample_Type` identifies the stability and control samples, and the variable `Indicate` is an indicator variable which takes the value 1 for stability samples and 0 for control samples.

```
DATA bivariate;
INPUT Time Time_Random Sample_Type$ Indicate Conc;
CARDS;
0  0  Stability  1  204
:  :  :  ::
24  24  Stability  1  207
1  1  Control  0  221
:  :  :  ::
24  24  Control  0  199
;
```

Sample MIXED code to fit the bivariate mixed model is given by:

```
PROC MIXED data=bivariate;
CLASS Time_Random Sample_Type;
MODEL Conc=Indicate*Time / ddf=6;
RANDOM Sample_Type / subject=Time_Random type=FA0(2);
REPEATED Time_Random / group=Sample_Type;
ESTIMATE 'Mean Conc @ Time=24' Int 1 Indicate*Time 24/ cl alpha=0.10;
```

## References


[1] FDA Guidance for Industry, Bioanalytical Method Validation, US Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), 2001.
[2] V. Shah, K. Midha, J. Findlay, et al., Pharm. Res. 17 (2001) 1551.
[3] K. Miller, R. Bowsher, A. Celniker, et al., Pharm. Res. 18 (2001) 1373.
[4] B. DeSilva, W. Smith, R. Weiner, et al., Pharm. Res. 20 (2003) 1885.
[5] J. Smolec, B. DeSilva, W. Smith, et al., Pharm. Res. 22 (2005) 1425.
[6] C. Viswanathan, S. Bansal, B. Booth, et al., AAPS J. 9 (2007) E30.
[7] W. Nowatzke, E. Woolf, AAPS J. 9 (2007) E117.
[8] S. Bansal, A. DeStefano, AAPS J. 9 (2007) E109.
[9] M. Kelley, B. DeSilva, AAPS J. 9 (2007) E156.
[10] U. Timm, M. Wall, D. Dell, J. Pharm. Sci. 74 (1985) 972.
[11] D. Dadgar, P. Burnett, J. Pharm. Biomed. Anal. 14 (1995) 23.
[12] D. Dadgar, P. Burnett, M. Choc, et al., J. Pharm. Biomed. Anal. 13 (1995) 89.
[13] R. Kringle, Pharm. Res. 11 (1994) 556.
[14] R. Kringle, D. Hoffman, J. Newton, R. Burton, Drug Inf. J. 35 (2001) 1261.
[15] J.T. Carstensen, Pharmaceutics of Solids and Solid Dosage Forms, Wiley-Interscience, New York, 1977.
[16] FDA Guidance for Industry, Q1E Evaluation of Stability Data. US Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), 2004.
[17] J. Shao, S.C. Chow, Statistica Sinica 11 (2001) 737.
[18] W. Fuller, G. Battese, J. Am. Stat. Assoc. 68 (1973) 626.