Journal of Biopharmaceutical Statistics, 16: 1–14, 2006

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543400500406421

Taylor & Francis
Taylor & Francis Group

AN ALTERNATIVE APPROACH TO EVALUATION OF POOLABILITY FOR STABILITY STUDIES

Jen-pei Liu
Department of Agronomy, Division of Biometry, National Taiwan University, Taipei, Taiwan
Division of Biostatistics and Bioinformatics, National Health Research Institutes Zhunan Town, Miaoli County, Taiwan

Sheng-Che Tung and Yun-Ming Pong
Department of Agronomy, Division of Biometry, National Taiwan University, Taipei, Taiwan

The current method for pooling the data from different batches or factors, suggested by ICH Q1E guidance, is to use analysis of covariance (ANCOVA) for test interaction between slopes and intercepts and factors. Failure to reject the null hypothesis of equality of slopes and equality of intercepts, however, does not prove that slopes and intercepts from different levels of factors are the same, and the data can be pooled for estimation of shelf life. In addition, the ANCOVA approach uses indirect parameters of intercepts and slopes in the regression model for assessment of poolability. The hypothesis for poolability is then formulated on the basis of the concept of equivalence for the means among the distributions of the quantitative attributes at a particular time point. Methods based on the intersection-union procedure are proposed to test the hypothesis of equivalence. A large simulation study was conducted to empirically investigate the size and power of the proposed method for the bracketing and matrixing designs given in the ICH QID guidance. Simulation results show that the proposed method can adequately control the size and provides sufficient power when the number of factors considered is fewer than three. A numerical example using the published data illustrates the proposed method.

Key Words: Bracketing Design; Equivalence; Matrixing Design; Poolability.

1. INTRODUCTION

A stability study usually considers design factors such as batch, strength, container size, and storage conditions. The ICH Q1D guidance on *Bracketing and Matrixing Designs for Stability Testing of New Drug Substances and Products* (ICH, 2003) provides examples of bracketing and matrixing designs with considerations of design factors: batch, strength, container size, and time points. On the other hand, as indicated by the ICH Q1E guidance *Stability Data Evaluation*, when establishing

Received August 12, 2004; Accepted July 1, 2005
Address correspondence to Jen-pei Liu, Department of Agronomy, Division of Biometry, National Taiwan University, 1 Section 4, Roosevelt Road, Taipei, Taiwan; Fax: 886-2-3366-4791; E-mail: jpliu@ntu.edu.tw

a shelf life (or an expiration dating period) for a drug product across all design factors, it is important to test poolability of these factors to see if the stability data can be pooled for estimating a single expiration dating period for the drug product. The ICH Q1E guidance requires that the method of analysis of covariance (ANCOVA) as the test for poolability be specified to determine whether there are statistically significant differences among factors and factor combinations. In addition, the ICH Q1E guidance recommends that a significance level of 0.25 be used for any terms involving batch and a significance level of 0.05 be used for terms not involving batch.

However, as pointed out by Ruberg and Stegemen (1991) and Chow and Liu (1995), a well-designed and well-conducted stability study provides high-precision data to detect miniature differences of no practice importance so that estimation of shelf life cannot use the common intercepts and slopes estimated with low variability and high efficiency. As a result, one of the major drawbacks of the ANCOVA approach is that good stability studies are in fact penalized for their low variability. Therefore, rejection of equal intercepts and equal slopes does not imply that the data from different factors cannot be pooled for estimation of a common shelf life. On the other hand, failure to reject the null hypothesis of equal intercepts and equal slopes does not prove that the data can be pooled. The reason is simply that the hypothesis of equality for intercepts and slopes is not the correct hypothesis for poolability across design factors.

Currently, the methods for testing poolability (ICH Q1D, 2003; Ruberg and Stegemen, 1991; Yoshioka et al., 1996) are based on the hypothesis formulated in indirect measures of intercepts and slopes of the regression models. As indicated in Yoshioka et al. (1997) and Tsong et al. (2003), poolability can be evaluated directly on similarity among the quantitative chemical attributes of a drug product at a particular time point across different levels of the batch. In the next section, the ANCOVA approach is reviewed, and the hypothesis for poolability based on an alternative representation of equivalence for the means among the distributions of the direct quantitative chemical attributes at time point T_0 is provided. Methods based on the intersection-union principle (Berger, 1982) for testing the hypothesis of equivalence are proposed in Section 3. A numerical example using the published data is provided in Section 4. The results of a simulation study to empirically investigate and compare the size and power of the proposed method for the bracketing and matrixing designs suggested in the ICH QID guidance are given in Section 5. Final remarks and discussion are provided in the final section.

2. EQUIVALENCE HYPOTHESIS FOR POOLABILITY

Assume that for a particular drug product, the relationship between certain quantitative attributes and time can be satisfactorily described by the following decreasing linear regression model:

$$Y_k = \alpha + \beta x_k + \varepsilon_k, \quad k = 1, \ldots, K, \tag{1}$$

where Y_k is the quantitative chemical attribute measured at time x_k, α and β are the intercept and slope of the regression equation, and ε_k is the random error

associated with Y_k and is assumed independently, identically distributed as a normal distribution with mean 0 and variance σ^2.

The current method for estimation of a shelf life recommended by the ICH Q1A (R2) and ICH Q1E is to determine the earliest time, say x_0, at which the 95% lower one-sided confidence limit for the mean intersects the lower acceptance criterion η. The hypothesis corresponding to the above confidence limit approach is given as

$$H_0 : (\eta - \alpha)/\beta \leq x_0 \quad \text{vs.} \quad H_a : (\eta - \alpha)/\beta > x_0, \tag{2}$$

which can be rewritten as

$$H_0 : \alpha + \beta x_0 \geq \eta \quad \text{vs.} \quad H_a : \alpha + \beta x_0 < \eta \tag{3}$$

From Eqs. (2) and (3), the 95% lower confidence limit approach provides a 95% confidence that the average quantitative attribute of the drug product is above η up to the end of shelf life x_0.

For illustration, we first consider the situation of a single factor, say batch, and Eq. (1) can be modified as follows.

$$Y_{jk} = \alpha_j + \beta_j x_k + \varepsilon_{jk}, \quad j = 1, \ldots, J; \quad k = 1, \ldots, K, \tag{4}$$

where Y_{jk} is the quantitative chemical attribute measured at time x_k for batch j, α_j and β_j are the intercept and slope of the regression equation for batch j, and ε_{jk} is the random error associated with Y_{jk} and is assumed independently, identically distributed as a normal distribution with mean 0 and variance σ^2.

Model (4) assumes that the time points for determination of quantitative attribute are the same for each of the J batches. In addition, model (4) can be re-expressed in terms of traditional ANCOVA model as

$$Y_{jk} = \mu + \tau_j + \beta x_k + (\beta_j - \beta) x_k + \varepsilon_{jk}, \quad j = 1, \ldots, J; \quad k = 1, \ldots, K, \tag{5}$$

where μ is the overall mean, τ_j is the fixed effect of batch, and β is the common slope of all batches.

The hypothesis for pooling using the ANCOVA approach based on model (5) is given as (Chow and Liu, 1995):

$$H_0 : \beta_j = \beta \text{ for all } j, \quad \text{vs.} \quad H_a : \beta_j \neq \beta \text{ for some } j \tag{6}$$

and

$$H_0 : \tau_j = \tau \text{ for all } j, \quad \text{vs.} \quad H_a : \tau_j \neq \tau \text{ for some } j, \tag{7}$$

where $j = 1, \ldots, J$.

If the hypotheses of equal intercepts and equal slopes are not rejected at the 0.25 level of significance for batch, a single expiration dating period can be estimated by fitting a single degradation curve based on pooled stability data of all batches.

However, failure to reject the null hypotheses provides no evidence to supporting pooling, rather than failure to show difference. On the other hand, rejection of the null hypothesis for a practically insignificant difference does not imply that the data cannot be pooled either. As a result, Ruberg and Stegemen (1991) and Chow and Liu (1995) proposed testing for equivalence of slopes and intercept as an alternative for batch poolability test. Their approaches are to test the following hypotheses:

$$H_0 : |\beta_j - \beta_{j'}| \geq \delta_\beta \text{ for some } j \neq j', \quad \text{vs.} \quad H_a : |\beta_j - \beta_{j'}| < \delta_\beta \text{ for all } j, j' \tag{8}$$

and

$$H_0 : |\alpha_j - \alpha_{j'}| \geq \delta_\alpha \text{ for some } j \neq j', \quad \text{vs.} \quad H_a : |\alpha_j - \alpha_{j'}| < \delta_\alpha \text{ for all } j, j', \tag{9}$$

where $j = 1, \ldots, J$; δ_β and δ_α are the equivalence limits for slope and intercept respectively.

The hypotheses of both ANCOVA approach and equivalence procedure are based on indirect measures of the intercepts and slopes for the regression model, which are really the effects of factors or the interactions between factors and time. For the examples of bracketing and matrixing designs given in the ICH Q1D guidance, the effects of factors and the interactions between factors and time might not be estimable and may be aliased with other factors. On the other hand, it is not feasible to follow each dosage unit until the time at which the quantitative attribute of the unit degrades to the lower acceptance criterion. A stability study is in fact an indirect assay, and the shelf life of the drug product is estimated from the quantitative attributes measured at the preselected time points. To overcome these issues, the evaluation of poolability can be assessed through the concept of equivalence for the means of the distributions of the quantitative attributes among batches at the prespecified time point T_0. Tsong et al. (2003) suggest testing whether the mean attribute at T_0 of each individual regression line is equivalent to that of the pooled regression line. On the basis of an alternative representation of the equivalence concept, we proposed another method to directly test equivalence in the means of the distributions of the quantitative attributes at T_0 among individual batches. Let θ_j denote the mean of the distribution of the quantitative attribute for batch j at time point T_0, i.e., $\theta_j = \alpha_j + \beta_j T_0$, $j = 1, \ldots, K$. The hypothesis for testing poolability of batches based on equivalence among the means of the quantitative attributes at T_0 is then given as

$$\begin{aligned} H_0 &: |\theta_j(T_0) - \theta_{j'}(T_0)| \geq \delta_{T_0}, \quad \text{for some } 1 \leq j \neq j' \leq J \\ \text{vs.} \quad H_0 &: |\theta_j(T_0) - \theta_{j'}(T_0)| < \delta_{T_0}, \quad \text{for all } 1 \leq j \neq j' \leq J. \end{aligned} \tag{10}$$

Equivalence limit δ_{T_0} in Eq. (10) is a function of time, drug class, and storage conditions. It should be jointly determined by experts of quality control/quality assurance for cGMP, medicinal chemist, pharmacist, clinician, and biostatistician. Furthermore, it should be specified in the stability protocol before the study is conducted. Once the null hypothesis [Eq. (10)] is rejected at the α significant level, the stability data from different batches or different levels of design factors can be pooled for estimation of a common expiration dating period.

3. PROPOSED PROCEDURES

The regression models in Eq. (4) can be represented in a matrix form as

$$\mathbf{Y} = \mathbf{AB} + \boldsymbol{\varepsilon}, \tag{11}$$

where $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_J)'$ is a $JK \times 1$ vector with $\mathbf{Y}_j$ being a $K \times 1$ vector of the K observed quantitative attributes from batch j; $\mathbf{A}$ is a $JK \times 2J$ block diagonal matrix formed by the design matrix $\mathbf{X}$ with on its diagonal for each batch; $\mathbf{X} = (\mathbf{1}, \mathbf{x})$, $\mathbf{1}$ is an $k \times 1$ vector of $\mathbf{1}$ and $\mathbf{x} = (x_1, \ldots, x_K)'$; $\mathbf{B_I} = (\alpha_1, \beta_1, \alpha_2, \beta_2, \ldots, \alpha_J, \beta_J)'$ is a $2J \times 1$ vector with the intercepts and slopes for each batch, $\boldsymbol{\varepsilon}$ is a $JK \times 1$ vector of random errors, which is assumed to follow a multivariate normal distribution with mean 0 and covariance matrix $\sigma^2\mathbf{I}$, and $\mathbf{I}$ is an $JK \times JK$ identity matrix.

It follows that $\mathbf{Y}$ is distributed as a $JK \times 1$ multivariate normal distribution with mean $\mathbf{AB}$ and covariance matrix $\sigma^2\mathbf{I}$.

Because the quantitative attributes from different batches are independent, the least squares estimator (LSE) of the intercept and slope for batch j can be obtained from the K observed quantitative attributes of batch j,

$$\widehat{\mathbf{B}}_j = \left(a_j, b_j\right)' = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}_j,$$

and the LSE of θ_j is given as

$$\begin{aligned}\hat{\theta}_j(T_0) &= a_j + b_j T_0 \\ &= \mathbf{x}_j(T_0)'\widehat{\mathbf{B}}_j, \quad j = 1, \ldots J\end{aligned} \tag{12}$$

where $\mathbf{x}_j(T_0)'$ is the $1 \times JK$ row vector such that the jth entry is 1, the $(j+1)$th entry is T_0, and 0 for the rest entries.

It is straightforward to see that $\hat{\theta}_j(T_0)$ is independently distributed as a normal distribution with mean $\theta_j(T_0)$ and variance $\sigma^2[\hat{\theta}_j(T_0)] = \mathbf{x}_j(T_0)'(\mathbf{A}'\mathbf{A})^{-1}\mathbf{x}_j(T_0)\sigma^2$.

An unbiased estimator of the variance of $\hat{\theta}_j(T_0)$ is given as

$$s^2\left[\hat{\theta}_j(T_0)\right] = \mathbf{x}_j(T_0)'(\mathbf{A}'\mathbf{A})^{-1}\mathbf{x}_j(T_0)\hat{\sigma}^2, \quad j = 1, \ldots, J$$

where $\hat{\sigma}^2 = \mathbf{Y}(\mathbf{I} - \mathbf{A}'(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A})\mathbf{Y}/df$, and $df = J(K-2)$.

It follows that the $(1-2\alpha)100\%$ confidence interval for pairwise differences $\theta_j(T_0) - \theta_{j'}(T_0)$ is give as

$$[\hat{\theta}_j(T_0) - \hat{\theta}_{j'}(T_0)] + t(\alpha, J(K-2))\sqrt{s^2[\hat{\theta}_j(T_0)] + s^2[\hat{\theta}_{j'}(T_0)]}, \quad 1 \le j \ne j' \le J \tag{13}$$

where $t(\alpha, J(K-2))$is the αth upper quantile of a central t-distribution with $J(K-2)$ degrees of freedom.

Under the intersection-union principle (Berger, 1982), the null hypothesis in Eq. (10) is rejected, and the quantitative attributes from batches can be pooled at the α significance level if all $(1-2\alpha)100\%$ confidence intervals for the $J(J-1)/2$ pairwise differences $\theta_j(T_0) - \theta_{j'}(T_0)$ are totally within $(-\delta_{T_0}, \delta_{T_0})$. Because there are J

batches, in addition to the t-statistic method proposed in Eq. (13) a more conservative approach is to employ either the Tukey-Kramer procedure or Bonferroni's correction to construct simultaneous confidence intervals for $\theta_j(T_0) - \theta_{j'}(T_0)$.

4. EXAMPLES

Two datasets given in Tsong et al. (2003) are used to illustrate the proposed procedure. Table 1 presents the first data set consisting of percents of label claim for a certain quantitative attribute from three batches measured at 0, 3, 6, 9, 12, 15, and 18 months. Tsong et al. showed that a target shelf life of 24 months is supported with the ANCOVA approach using the 0.25 significance level recommended by the ICH Q1E guidance. Table 2 gives the 90% confidence intervals at 24 months using the t-statistic method, Tukey-Kramer procedure, and Bonferroni's correction. As mentioned before, equivalence limit should be determined jointly by the project team and specified in the stability protocol. Although an equivalence margin of 3% seems neither too liberal nor too stringent if the lower acceptable limit for percent of label claim is 90%, it is used solely for the purpose of illustration. It follows that the 90% confidence intervals for $\theta_1(24) - \theta_3(24)$ and $\theta_2(24) - \theta_3(24)$ are not contained within $(-3.0\%, 3.0\%)$ by all three procedures. This implies that if the individual linear regression model for each batch holds at 24 months, the means of the distributions of the quantitative attributes are not equivalent at the 0.05 significance level. Therefore, at the 0.05 significance level, the data of quantitative attributes from the three batches cannot be pooled at 24 months for estimation of a common shelf life. From Table 2, the confidence interval by the t-statistic method produces the shortest length. On the other hand, the length of the confidence interval by the Bonferroni's correction is wider than that by the Tukey-Kramer procedure.

Table 3 presents the second dataset of percents of label claim for a certain quantitative attribute from batches. In Tsong et al. (2003), there are three replicates at each time point. For the purpose of illustration, three replicates are modified as low (L), median (M), and high (H) strength. Because there are two factors for this dataset, a two-step procedure is performed. The drug products of the same batch were made of the same raw materials at the same period using the same manufacturing process. Therefore, it is more important to verify whether the data of different strengths of the same batch can be pooled. As a result, the poolability of strength within each batch is tested first. If the data of three strengths can be pooled

Table 1 Stability data in percent of label claim from three batches

	Time (months)						
Batch	0	3	6	9	12	15	18
1	100	101	101	100	99	98	98
2	99	100	100	99	99	98	97
3	100	100	99	99	98	97	97

Source: From Tsong et al. (2003).

Table 2 Summary of 90% confidence intervals for batch poolability for the data in Table 1

Comparison	*t*-Statistic method	Tukey-Kramer procedure	Bonferroni's correction
1 vs. 2	(−1.34, 1.75)	(−1.81, 2.16)	(−1.91, 2.26)
1 vs. 3	(0.18, 3.32)	(−0.23, 3.73)	(−0.34, 3.84)
2 vs. 3	(0.005, 3.14)	(−0.41, 3.56)	(−0.51, 3.66)

within each batch, the poolability of batch is then tested by using the combined data from strength. Table 4 provides the 90% confidence intervals to test the poolability of strengths within batch at 30 months using the *t*-statistic method, Tukey-Kramer procedure, and Bonferroni's correction. Again an equivalence limit of 3% is used to illustrate the example. From Table 4, all nine 90% confidence intervals are totally contained within (−3.0%, 3.0%) by three procedures. This indicates that for each batch, the means of the distributions of quantitative attributes for three strengths are equivalent at the 0.05 significance level. As a result, the data of three strengths can be pooled for estimation of a common shelf life for each batch. Table 5 provides the 90% confidence intervals to test the poolability of batches at 30 months after the data from strengths are combined for each batch. All three 90% confidence intervals for pairwise mean batch differences are also contained within (−3.0%, 3.0%) by *t*-statistic method, the Tukey-Kramer procedure, and the Bonferroni's correction. This concludes that at the 0.05 significance level, the data of the quantitative attributes can be pooled from both batches and strength for estimation of a common shelf life at 30 months.

5. SIMULATION STUDY

A simulation study was conducted to investigate and compare performance of the empirical size and power for testing poolability by the proposed procedures. R package and its subroutines were used in the simulation study. We investigate the impact of factors, the differences between means, the magnitude of variability,

Table 3 Stability data in percent of label claim by batch and strength

		Time (months)						
Batch	Strength	0	3	6	9	12	18	24
1	L	100.6	100.3	99.5	99.1	98.7	97.2	96.6
	M	100.7	99.9	99.4	98.9	98.7	97.2	96.5
	H	100.3	100.1	99.4	99.1	98.5	97.3	96.7
2	L	101.0	100.3	99.6	99.2	98.7	97.3	96.7
	M	101.0	100.3	99.7	99.3	99.0	97.4	96.5
	H	100.4	100.1	99.5	99.2	98.4	97.1	96.4
3	L	100.5	99.8	99.4	98.7	98.3	97.1	96.0
	M	100.6	99.7	99.2	98.8	98.3	97.4	96.2
	H	100.3	99.8	99.5	99.1	98.0	97.0	96.0

Source: Modified from Tsong et al. (2003).

Table 4 Summary of 90% confidence intervals for strength poolability by batch for the data in Table 2

Batch	Strength	*t*-Statistic method	Tukey-Kramer procedure	Bonferroni's correction
1	L vs. M	(−0.39, 0.48)	(−0.51, 0.60)	(−0.54, 0.63)
	L vs. H	(−0.66, 0.21)	(−0.78, 0.32)	(−0.81, 0.36)
	M vs. H	(−0.71, 0.17)	(−0.82, 0.28)	(−0.85, 0.31)
2	L vs. M	(−0.40, 0.47)	(−0.52, 0.59)	(−0.55, 0.62)
	L vs. H	(−0.28, 0.59)	(−0.40, 0.71)	(−0.43, 0.74)
	M vs. H	(−0.32, 0.56)	(−0.43, 0.67)	(−0.46, 0.70)
3	L vs. M	(−0.75, 0.13)	(−0.86, 0.24)	(−0.89, 0.27)
	L vs. H	(−0.41, 0.46)	(−0.53, 0.58)	(−0.56, 0.60)
	M vs. H	(−0.11, 0.77)	(−0.22, 0.88)	(−0.25, 0.91)

time points, and the designs on the size and power. Table 6 presents the full time points and the matrixing design for one-third reduction in time suggested in the ICH Q1D guidance. For one (batch) and two factors (batch, strength), only the matrixing designs on time points were considered and are given in Table 7. Table 8 gives the designs for three factors (batch, strength, and container size) considered in the simulation. These designs include the full design, the matrix design on time points, and the matrix design on both time points and factors recommended in the ICH Q1D guidance. In addition, only the situation of three levels for each factor was considered.

For the simulation, the equivalence limit is set to be 3% again. Different values for the means of the distributions of percents of label claim at 24 months of the three levels of one factor are specified, whereas the means of all levels for other factors are the same. The empirical size and power were then examined at various values between 0% and 3.4% by an increment of 0.2% for the maximum mean difference among three levels of that factor. For the size, the means of the distributions of percents of label claim for the three levels are specified as 92.167, 93.667, and 95.167. The error variances of 0.16, 0.36, and 0.64 were chosen to examine the impact of variability on the size and power. For each of the combinations, 2000 random samples were independently generated from normal distribution. For a 0.05 nominal significance level, a simulation study with 2000 random samples implies that 95% of empirical sizes evaluated at the equivalence margins will be within 0.04045 and 0.05955.

Table 9 presents the empirical sizes of the full design and matrixing design for one factor (batch). For Table 9, the empirical size decreases as the variability

Table 5 Summary of 90% confidence intervals for batch poolability for the data in Table 2

Comparison	*t*-Statistic method	Tukey-Kramer procedure	Bonferroni's correction
1 vs. 2	(−0.12, 0.37)	(−0.18, 0.43)	(−0.19, 0.44)
1 vs. 3	(0.22, 0.71)	(0.16, 0.79)	(0.15, 0.78)
2 vs. 3	(0.10, 0.58)	(0.04, 0.64)	(0.03, 0.65)

Table 6 Design for time points

Design	Time points (month)
Full time points (F)	0, 3, 6, 9, 12, 18, 24, 36
Matrixing on time points (MT)	
T1	0, 6, 9, 12, 18, 24, 36
T2	0, 3, 9, 12, 24, 36
T3	0, 3, 6, 12, 18, 36

increases. On the other hand, the empirical size of the matrixing design is also smaller than that of the full design. The empirical sizes of the t-statistic method range from 0.042 to 0.054, which are within the 95% confidence interval above. This implies that the t-statistic method adequately controls the size for testing the hypothesis of poolability in Eq. (10) for both the full design and matrixing design. However, the empirical size of the pairwise equivalence by Tukey-Kramer procedure and Bonferroni's correction ranges from 0.012 to 0.026 which are smaller than 0.04045. Therefore, these two methods are quite conservative.

For one factor (batch), Fig. 1 provides empirical power curves of the t-statistic method, Tukey-Kramer procedure and Bonferroni's correction, respectively. The magnitude of difference on the horizontal axis is the difference between the largest mean and the smallest mean of the distributions of percents of label claim at 24 months. A horizontal reference line is set at 0.05, and a vertical reference line is set at 3.0 for comparison of size. All empirical power curves are monotonic decreasing function of mean difference. For the t-statistic method, the empirical power function at the maximum mean difference of 3.0 is about 0.05. Its empirical power is greater than 0.05 when mean difference is larger than 3.0 and is smaller than 0.05 when mean difference is less than 3.0. This indicates that the t-statistic method may be an unbiased test for one factor. In addition, for one factor, the t-statistic method is uniformly more powerful than both the Tukey-Kramer procedure and Bonferroni's correction.

For two factors (strength and batch), different values for the means of the distributions of percents of label claim at 24 months for the three levels of strength are specified for each level of batches. Table 10 presents the empirical sizes for two

Table 7 Matrixing design used in simulation for one and two factors

A. Matrixing design on time points (MT) for one factor

Batch 1	T1
Batch 2	T2
Batch 3	T3

B. Matrixing design on time points (MT) for two factors

	Strength		
Batch	S1	S2	S3
1	T3	T1	T3
2	T2	T3	T1
3	T1	T3	T2

Table 8 Matrixing design used in simulation for three factors

Strength	S1			S2			S3		
container size	A	B	C	A	B	C	A	B	C
A. Matrixing design on time points (MT) for three factors									
Batch 1	T1	T2	T3	T2	T3	T1	T3	T1	T2
Batch 2	T2	T3	T1	T3	T1	T2	T1	T2	T3
Batch 3	T3	T1	T2	T1	T2	T3	T2	T3	T1
B. Matrixing design on factors and time points (MFT)									
Batch 1	T1	T2		T2		T1		T1	T2
Batch 2		T3	T1	T3	T1		T1		T3
Batch 3	T3		T2		T2	T3	T2	T3	

factors. Therefore, we first test poolability of strength (batch) within each level of batch (strength) and then proceed to test poolability of batch (strength) if strength (batch) can be pooled. The empirical sizes of testing poolability of strength within each level of batches are presented in Table 10. In general, the performance of all methods for two factors is similar to that for one factor. The *t*-statistic method still outperforms the other methods. However, the empirical sizes of two factors are smaller that those observed for one factor. For example, the empirical size of the *t*-statistic method is below 0.04055 for the matrixing design when the variability is 0.64.

For three factors (strength, batch, and container size), different values for the means of the distributions of percents of label claim at 24 months for the three levels of strength are specified for each level of for each combination of batch and container size. Table 11 presents the empirical sizes for three factors. For three factors, to shorten the number of testing steps and to examine the impact of the number of means, a top-down strategy was used. For example, for each level of batch, poolability is tested for the nine means of strength by container size. If the null hypothesis of inequivalence is rejected, then the data of quantitative attributes can be pooled over all combinations. Table 10 shows that all empirical sizes of three factors under full design, matrixing design on time, and matrixing design on factor and time are smaller than 0.001 for all methods. Therefore, all methods using the top-down strategy are very conservative.

Table 9 Empirical size for one factor

Variation	Design	*t*	Tukey	B
0.16	F	0.054	0.026	0.020
	MT	0.052	0.022	0.017
0.36	F	0.047	0.025	0.018
	MT	0.045	0.020	0.015
0.64	F	0.044	0.018	0.014
	MT	0.042	0.016	0.013

t: *t*-statistics; Tukey: Tukey-Kramer procedure; B: Bonferroni's correction; F: full time points; MT: matrixing design on time.

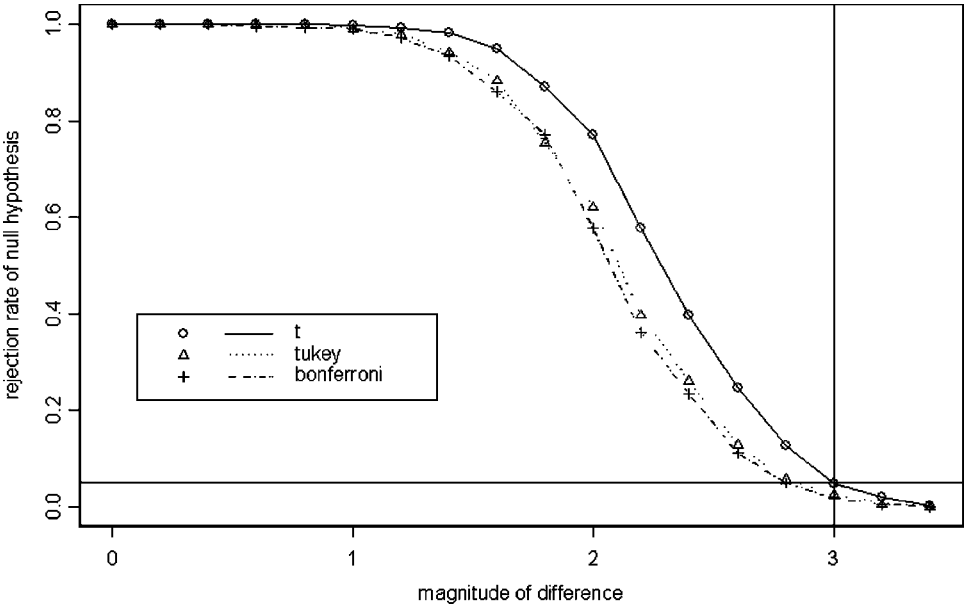

Figure 1 Power curve of one-factor combination.

6. DISCUSSION

Testing poolability for design factors of stability studies has been a very important issue in determination of shelf lives for drug products. Considerable coverage has been devoted to this issue in Appendix B of the recently released ICH Q1E guidance (2004). The ICH Q1E guidance recommends the ANCOVA approach to test poolability based on the null hypotheses of equality of intercepts and equality of slopes at the 0.25 level for batch-related terms and at the 0.05 level for non-batch-related terms. However, failure to reject the null hypotheses of equality of intercepts and slopes does not prove that regression equations for all combination have the same intercepts and slopes. On the other hand, the reason provided in the ICH Q1E guidance for poolability for batch-related terms being tested at the 0.25 significance level is the relatively limited sample size in a typical formal stability study. However, this is also true for non-batch-related terms. The ICH Q1E guidance does not provide the reason for testing the non-batch-related terms at the 0.05 significance level.

The ICH Q1D guidance provides examples of bracketing and matrixing designs for stability studies. Although the reasons behind the bracketing and matrixing designs are different, they are in fact fractional factorial designs. Therefore, certain effects of the matrixing design and bracketing designs are not estimable or might be aliased with other effects and interactions. For example, the matrixing design on factors and time points given in B of Table 8 is a combination of two 3^{3-1} fractional factorial designs with resolution III. However, it is not known which effects or interactions of this parallel flats design can be estimated and their alias patterns (Liao, 1996). However, on the other hand, the primary goal of any stability study is to estimate the shelf life of the drug product. Overemphasis

Table 10 Empirical size for poolability of strength for two factors

Variation	Design	Batch	t	Tukey	B
0.16	F	1	0.052	0.020	0.016
		2	0.048	0.019	0.014
		3	0.037	0.016	0.013
	MT	1	0.045	0.017	0.014
		2	0.047	0.017	0.013
		3	0.043	0.018	0.013
0.36	F	1	0.044	0.017	0.012
		2	0.046	0.019	0.013
		3	0.042	0.016	0.017
	MT	1	0.040	0.017	0.012
		2	0.039	0.015	0.013
		3	0.042	0.018	0.013
0.64	F	1	0.044	0.016	0.011
		2	0.045	0.016	0.013
		3	0.040	0.017	0.014
	MT	1	0.037	0.015	0.010
		2	0.039	0.015	0.011
		3	0.036	0.016	0.013

t: t-statistics; Tukey: Tukey-Kramer procedure; B: Bonferroni's correction; F: full time points; MT: matrixing design on time.

of testing the interactions between design effects and time may obscure the true objective of stability studies. On the other hand, a shelf life should guarantee, at the 95% confidence level, that the mean quantitative attributes is above (or within) the acceptable criteria up to the end of the shelf life. Therefore, it is more reasonable to use the mean of the distributions of the quantitative attributes at the proposed shelf life as the parameter of interest in conjunction of the concept of equivalence for testing the poolability. Not only this approach is intuitively appealing but also it can overcome the estimable issue of the effects and interactions suffered by the ANCOVA approach. In addition, rejection of the null hypothesis of inequivalence concludes the equivalence among means of the distributions at the proposed shelf life. Following the ICH Q1E guidance, if the poolability is proved by the proposed equivalence approach, then the data can be pooled for estimation of a common shelf life as described in Appendix B.1 of ICH Q1E guidance. On the other hand, if the poolability cannot be proved, then the shortest shelf life among all levels of factors is chosen as the shelf life for the drug product as suggested in Appendix B.2.2 an Appendix B.3.2.

Batch should be considered as a design factor for a stability study, and each factor has relatively the same sample size if the levels of all factors are approximately close to each other. As a result, it is recommended that hypothesis (10) be test at the 0.05 significance level for all factors including batch. However, without any prior information about impact of different factors on degradation pattern and rate of the drug product, it seems reasonable to assume a common equivalence margin for all factors. The hypothesis for equivalence proposed by Tsong et al. (2003) is based on the difference between $\theta_j(T_0)$ and its average, whereas our approach is based on the paired difference between $\theta_j(T_0)$ and $\theta_{j'}(T_0)$.

Table 11 Empirical size for poolability of strength and container size for three factors

Variation	Design	Batch	t	Tukey	B
0.16	F	1	<0.001	<0.001	<0.001
		2	<0.001	<0.001	<0.001
		3	<0.001	<0.001	<0.001
	MT	1	<0.001	<0.001	<0.001
		2	<0.001	<0.001	<0.001
		3	<0.001	<0.001	<0.001
	MFT	1	<0.001	<0.001	<0.001
		2	<0.001	<0.001	<0.001
		3	<0.001	<0.001	<0.001
0.36	F	1	<0.001	<0.001	<0.001
		2	<0.001	<0.001	<0.001
		3	<0.001	<0.001	<0.001
	MT	1	<0.001	<0.001	<0.001
		2	<0.001	<0.001	<0.001
		3	<0.001	<0.001	<0.001
	MFT	1	<0.001	<0.001	<0.001
		2	<0.001	<0.001	<0.001
		3	<0.001	<0.001	<0.001
0.64	F	1	<0.001	<0.001	<0.001
		2	<0.001	<0.001	<0.001
		3	<0.001	<0.001	<0.001
	MT	1	<0.001	<0.001	<0.001
		2	<0.001	<0.001	<0.001
		3	<0.001	<0.001	<0.001
	MFT	1	<0.001	<0.001	<0.001
		2	<0.001	<0.001	<0.001
		3	<0.001	<0.001	<0.001

t: t-statistics; Tukey: Tukey-Kramer procedure; B: Bonferroni's correction; F: full time points, MT: matrixing design on time; MFT: matrixing design on factor and time.

The method proposed by Tsong et al. (2003) requires additional estimation of the average of $\theta_j(T_0)$ from the common regression line from the pooled data for their testing procedure. On the other hand, our testing procedure is only based on the estimates of the paired mean differences of the quantitative attributes at the specified time point. For our approach, the common regression line based on the pooled data is fitted only after the poolability of the data is concluded at the α significance level. On the other hand, the equivalence margin for Tsong's procedure is determined by the maximum difference between $\theta_j(T_0)$ and its average. The equivalence margin for our approach is based on the allowable maximum difference between $\theta_j(T_0)$ and $\theta_{j'}(T_0)$ which has a more direct intuitive meaning and interpretation.

The simulation shows that the t-statistic method not only can control the size but also is uniformly more powerful than other methods. The simulation results provided in Table 10 presents a situation of a total of nine means: 3 with mean of 92.167, 3 with mean of 93.667, and 3 with mean of 95.167. Even in this situation, the empirical sizes for all methods under all designs is smaller than 0.001. Therefore, if the number of means increases, say greater than 4, all methods becomes very conservative. As a result, a bottom-up stepwise procedure from the highest level

combinations is recommended. More research is urgently needed to investigate the impact of multifactor matrixing and bracketing designs on poolability of stability for estimating a common shelf life.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers and Associate Editor for their careful, thoughtful, and thorough review and comments, which greatly improved the content and presentation of our work. This work is partially supported by the Taiwan National Science Grant: NSC 93-2118-M-006-002.

REFERENCES

Berger, R. L. (1982). Multiparameter hypothesis testing in acceptance sampling. *Technometrics* 34:295–300.

Chow, S. C., Liu, J. P. (1995). *Statistical Design and Analysis in Pharmaceutical Sciences*. New York: Marcel Dekker.

Guidance for Industry: ICH Q1A (R2). (2003). *Stability Testing of New Drug Substances and Products*. Rockville, Maryland: Food and Drug Administration, Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research.

Guidance for Industry: ICH Q1D (R2). (2003). *Bracketing and Matrixing Designs for Stability Testing of New Drug Substances and Products*. Rockville, Maryland: Food and Drug Administration, Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research.

Guidance for Industry: ICH Q1E. (2004). *Stability Data Evaluation*. Rockville, Maryland: Food and Drug Administration, Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research.

Liao, C. T. (1996). Orthogonal three-level parallel flats designs for user-specified resolution. *J. Comm. Statist. Theory Methods* 28:1945–1960.

Ruberg, S. J., Stegemen, J. W. (1991). Pooling data from stability studies: testing the equality of batch degradation slopes. *Biometrics* 47:1059–1069.

Tsong, Y., Chen, W. J., Lin, T. Y. D., Chen C. W. (2003). Shelf life determination based on equivalence assessment. *J. Biopharm. Stat.* 13(3):431–449.

Yoshioka, K., Aso, Y., Kojima, S., Po, A. L. W. (1996). Power of analysis of variance for assessing batch-variation of stability data. *Chem. Pharm. Bull.* 44(10):1948–1950.

Yoshioka, K., Aso, Y., Kojima, S. (1997). Assessment of shelf-life equivalence of pharmaceutical products. *Chem. Pharm. Bull.* 45(9):1482–1484.