AAPS PharmSciTech, Vol. 10, No. 3, September 2009 (© 2009)
DOI: 10.1208/s12249-009-9270-x

Research Article

A Two One-Sided Parametric Tolerance Interval Test for Control of Delivered Dose Uniformity. Part 1—Characterization of FDA Proposed Test

Steven Novick,[1] David Christopher,[2] Monisha Dey,[2] Svetlana Lyapustina,[3,9] Michael Golden,[4] Stefan Leiner,[5] Bruce Wyka,[6] Hans-Joachim Delzeit,[5] Chris Novak,[7] and Gregory Larner[8]

Received 17 January 2009; accepted 27 May 2009; published online 24 June 2009

Abstract. The FDA proposed a parametric tolerance interval (PTI) test at the October 2005 Advisory Committee meeting as a replacement of the attribute (counting) test for delivered dose uniformity (DDU), published in the 1998 draft guidance for metered dose inhalers (MDIs) and dry powder inhalers (DPIs) and the 2002 final guidance for inhalation sprays and intranasal products. This article (first in a series of three) focuses on the test named by the FDA "87.5% coverage." Unlike a typical two-sided PTI test, which controls the proportion of the DDU distribution within a target interval (coverage), this test is comprised of two one-sided tests (TOST) designed to control the maximum amount of DDU values in either tail of the distribution above and below the target interval. Through simulations, this article characterizes the properties and performance of the proposed PTI-TOST under different scenarios. The results show that coverages of 99% or greater are needed for a batch to have acceptance probability 98% or greater with the test named by the FDA "87.5% coverage" (95% confidence level), while batches with 87.5% coverage have less than 1% probability of being accepted. The results also illustrate that with this PTI-TOST, the coverage requirement for a given acceptance probability increases as the batch mean deviates from target. The accompanying articles study the effects of changing test parameters and the test robustness to deviations from normality.

KEY WORDS: delivery dose uniformity; inhaler; parametric tolerance interval PTI; two one-sided test TOST.

[1] Discovery Analytics, GlaxoSmithKline, Research Triangle Park, North Carolina, USA.
[2] Statistics, Schering-Plough Research Institute, Kenilworth, New Jersey, USA.
[3] Pharmaceutical Practice Group, Drinker Biddle & Reath, 1500 K Street NW, Suite 1100, Washington, District of Colombia 20005-1209, USA.
[4] Regulatory Affairs and Quality, Pearl Therapeutics, Raleigh, North Carolina, USA.
[5] Boehringer Ingelheim Pharma GmbH & Co. KG, Ingelheim am Rhein, Germany.
[6] SpiraPharma Consulting, Lincoln Park, New Jersey, USA.
[7] Lab Services Department, Drug Delivery Systems Division, 3M, St. Paul, Minnesota, USA.
[8] Scientific and Laboratory Services, Pfizer, Kalamazoo, Michigan, USA.
[9] To whom correspondence should be addressed. (e-mail: svetlana.lyapustina@dbr.com)

ABBREVIATIONS: ACPS, Advisory Committee for Pharmaceutical Science; BOU, beginning of unit; CMC, chemistry, manufacturing, and controls; DDU, delivered dose uniformity; DPI, dry powder inhaler; EOU, end of unit; IPAC-RS, International Pharmaceutical Aerosol Consortium on Regulation and Science; MDI, metered dose inhaler; OC, operating characteristic; OINDP, orally inhaled and nasal drug products; $Pmax_{TA}$, maximum allowable proportion of doses in a tail area (left or right) outside the target interval; PTI, parametric tolerance interval; TOST, two one-sided tests; α, significance level of the test (maximum type I error); K, PTI test coefficient; L, lower boundary of a target interval; μ, population (batch) mean; N, total sample size for both tiers; N_1, sample size in the first tier; N_2, additional units tested in the second tier; s, sample standard deviation; σ, population (batch) standard deviation; U, upper boundary of a target interval; $\bar{X}$ sample mean.

INTRODUCTION

Uniformity of dose as delivered by an inhalation device or nasal spray has long been viewed as one of the key quality attributes of orally inhaled and nasal drug products (OINDP). Historically, FDA used an attribute (counting) test for control of DDU, which included a "zero-tolerance" component, i.e., a requirement that no observation within a sample be outside pre-set limits. Detailed descriptions of such tests were included in the Agency's draft and final guidances for industry (1,2). Those tests had several counter-productive features, such as penalizing the producer for increased testing irrespective of batch quality (3). With the recent regulatory initiatives to modernize approaches to pharmaceutical manufacturing and controls came the opportunity to improve DDU testing by using PTI methods. After discussions with the industry through a joint working group of the Advisory Committee for Pharmaceutical Science (ACPS), the agency developed and presented at an ACPS meeting a particular PTI test (4), which was recommended as a replacement for the previous non-parametric attribute testing.

In general, a PTI test uses parameter estimates (such as mean and standard deviation) of the underlying data distribution in order to construct a statistical interval to make a decision regarding the batch disposition (e.g., a pass/fail decision) based on the outcome of testing. The exact structure (formulas, coefficients, limiting values, etc.) and protocol of the test (the number of inhalers to be tested, the number of

doses and at what life-stage within multidose inhalers to be collected, the number of inhaler actuations per DDU observation, the number of tiers and rules for going to the next tier, etc.), taken together, define a specific PTI test and determine its performance characteristics. Without specifying all such details, a PTI test description is incomplete and cannot be directly implemented. Each completely defined PTI test is unique in that it behaves differently in response to a given situation. The discussion in this article pertains only to the PTI-TOST and may not be applicable to other PTI tests. The article explores in detail the characteristics of the PTI-TOST identified by FDA as "95% confidence level, 87.5% coverage, target interval=80–120% of the label claim (LC), sample size=20+40 DDU observations in the $1^{st}+2^{nd}$ tier, each inhaler being tested in the beginning and end of container life." The article purposefully does not address the suitability of this test as a standard for inhalation products because such determination would depend on a specific product's data.

Characterizing other types of PTI tests and commenting on the PTI-TOST acceptance criteria are outside the scope of this paper.

MATERIALS AND METHODS

This work is based on statistical simulations using the Monte Carlo technique, by sampling with replacement from a normal distribution and following the sampling scheme specified below.

Assumptions and Notation

DDU measures the amount of drug substance delivered by an OINDP. For multidose products, where the dose and dose uniformity may change as the inhaler unit is being emptied, measurements on the unit are made at the beginning of unit (BOU) and the end of unit (EOU), which are sometimes also referred to as the beginning and end of use, respectively. For example, if a unit is labeled to contain 120 actuations, BOU is typically determined by collecting the first dose after priming (if priming is required) and EOU by collecting the dose near or at the labeled number of actuations. The *dose* is understood to be the number of actuations that comprise the minimum patient dose. For single-dose OINDPs, which do not have life stages, the analysis presented in this paper applies if twice the number of OINDP units is used (to obtain the same total sample size N).

For the purposes of this analysis, it is assumed that the units within a batch and the BOU and EOU measurements, which are collected from each unit, are all independent and come from a univariate normal distribution with mean μ and standard deviation σ, except where specified otherwise. The observed DDU values represent a *sample* from the *population* of all possible DDU values in a given batch. The population parameters μ and σ are *within-batch* parameters and are likely to change from batch to batch; therefore, different combinations of μ and σ are considered throughout the paper. For a particular batch, the estimates for μ and σ are given by $\overline{X}$ and s, the sample mean and sample standard deviation from a sample of size N.

In this article, for a two-tier test, N_1 denotes the sample size (number of DDU observations) in the first tier, and N_2 is the number of additional DDU observations in the second tier, so that $N_1+N_2=N$ gives the total sample size for a complete test. In the PTI-TOST, the number of inhalers used is $N/2$ for multi-dose products because each inhaler is tested twice—at the beginning and end-life stages of the units.

Parametric Tolerance Interval Two One-Sided Tests

A hypothetical normal distribution of doses (x) within a batch with mean μ and standard deviation σ is represented in Fig. 1. The *coverage* of a typical PTI test is represented by the portion of the distribution between L and U, which define the *target interval*. The portions of the distribution below L and above U represent the *tails*. In this paper, we refer to the maximum allowable area in either tail as Pmax_{TA}. Since the proportion in either tail equals Pmax_{TA}, the distribution in Fig. 1 illustrates a batch of *limiting quality*, which has a very small probability of being judged acceptable. For both a typical PTI test and the PTI-TOST, batch quality is determined by first sampling from a batch and then analytically testing the sample. In a typical PTI test, a tolerance interval that is directly related to coverage is calculated using the analytical results and is compared to the target interval. If the calculated tolerance interval is contained within the target interval, the batch passes the test. By contrast, for a PTI-TOST, two one-sided tolerance limits (which are directly related to the tail areas of the distribution) are calculated using the analytical results and compared to L and U, respectively. If both one-sided tolerance limits are inside the target interval (higher than L and lower than U), the batch passes the test.

The Complete FDA DDU Test

The DDU test proposed by the FDA in October 2005 included a two-tier PTI-TOST with $\text{Pmax}_{\text{TA}}=6.25\%$ and overall $\alpha=0.05$ distributed between tiers according to the Lan–DeMets implementation of the Pocock method (see Appendix), and additional non-parametric criteria that the sample mean must fall between 85% and 115% LC (4). In a follow-up presentation (5), FDA clarified that the requirement on sample means must be applied to BOU and EOU

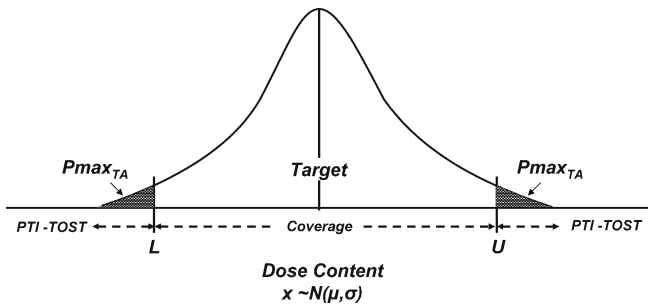

Fig. 1. Illustration of the distribution of doses within a batch and of the maximum allowable proportion of doses in either tail area (Pmax_{TA}) above or below specified dose content limits U and L

separately. Therefore, for a sample size of $N_1=20$ and $N_2=40$, the PTI-TOST as proposed by the FDA can be described as follows.Tier 1:

- Collect 20 doses (from 10 multi-dose OINDP units, a BOU and an EOU measurement from each unit).
- Calculate $\overline{X}_1$ and s_1 as the mean and standard deviation of the N_1 observations.
- For multi-dose OINDPs, calculate means for BOU and EOU ($\overline{X}_{BOU,1}$ and $\overline{X}_{EOU,1}$).

The 20 observations must pass the following criteria.

$$T_{L1} = \overline{X}_1 - K_1\, s_1 \geq 80 \text{ with } \text{Pmax}_{\text{TA}} = 6.25\% \text{ and } \alpha_1 = 0.0226 \text{ where } K_1 = 2.448 \quad (1a)$$

$$T_{U1} = \overline{X}_1 + K_1\, s_1 \leq 120 \text{ with } \text{Pmax}_{\text{TA}} = 6.25\% \text{ and } \alpha_1 = 0.0226 \text{ where } K_1 = 2.448 \quad (1b)$$

$$85 \leq \overline{X}_{BOU,1} \leq 115 \quad (1c)$$

$$85 \leq \overline{X}_{EOU,1} \leq 115 \quad (1d)$$

If the sample fails any of the criteria, the test proceeds to the second tier. Tier 2:

- Collect an additional 40 doses (from 20 multi-dose OINDP units, BOU and EOU measurement from each unit). There are now $(N_1+N_2)=60$ DDU observations.
- Calculate $\overline{X}_2$ and s_2 as the mean and standard deviation of the (N_1+N_2) observations.
- For multi-dose OINDPs, calculate means for BOU and EOU ($\overline{X}_{BOU,2}$ and $\overline{X}_{EOU,2}$).

The 60 observations must pass the following criteria.

$$T_{L2} = \overline{X}_2 - K_2\, s_2 \geq 80 \text{ with } \text{Pmax}_{\text{TA}} = 6.25\% \text{ and } \alpha_2 = 0.0340 \text{ and where } K_2 = 1.940 \quad (2a)$$

$$T_{U2} = \overline{X}_2 + K_2\, s_2 \leq 120 \text{ with } \text{Pmax}_{\text{TA}} = 6.25\% \text{ and } \alpha_2 = 0.0340 \text{ and where } K_2 = 1.940 \quad (2b)$$

$$85 \leq \overline{X}_{BOU,2} \leq 115 \quad (2c)$$

$$85 \leq \overline{X}_{EOU,2} \leq 115 \quad (2d)$$

If the criteria in Eqs. 1a, 1b, 1c, and 1d or 2a, 2b, 2c, and 2d (in case tier 2 is needed) are met, the batch passes the test.

A closed form for calculating K_1 and K_2 is given in the Appendix. From that formula, it is apparent that K_1 and K_2 depend on the overall α, α_1, and α_2, Pmax_{TA}, total sample size N, and sample size in the first tier N_1 but not on the target interval. A number of values for K_1 and K_2 have been calculated for various test protocols and are presented in the second article of this series.

Operating Characteristic Curves

This article uses operating characteristic (OC) curves to investigate various scenarios and to elucidate the test's characteristics. OC curves graphically illustrate the probability of passing a defined test under a given set of conditions (e.g., μ and σ) (6). OC curves are also useful for comparing probabilities of passing different tests, multiple tests, or tests run under different scenarios. In this article, OC curves were generated using standard statistical methods assuming a normal data distribution. The test's performance with non-normal distributions is studied in the third paper of this series.

RESULTS

OC Curves for the PTI-TOST

This section presents the OC curves for the FDA DDU test as introduced above.

For the complete FDA DDU test, the OC curves in this section represent, for different values of batch μ and σ, the following acceptance probability:

$$\Pr\left(T_L \geq 80,\ T_U \leq 120,\ 85 \leq \overline{X}_{BOU} \leq 115,\ 85 \leq \overline{X}_{EOU} \leq 115 | \mu, \sigma\right)$$

Figure 2 presents OC curves for batches with the mean ranging from on-target to 18% off-target. As these curves show, the probability of passing the FDA DDU test is highest when $\mu=100$ ($\Delta=0$) and σ is small. As μ approaches 80 or 120 ($\Delta=20$) or as σ approaches 12, it is nearly impossible to pass the test.

It is of interest to investigate the influence of the non-parametric criteria for the life-stage means (expressed in Eqs. 1c, 1d, 2c, and 2d) on the operating characteristics of the test. The OC curves for testing scenarios with and without the life-stage means component are illustrated by Fig. 3. The thick dashed and solid lines show the complete PTI-TOST, and the thin dashed and solid lines show PTI-TOST without the life-stage means criteria. The two are nearly identical except when the batch is off-target by more than 15% LC. It is clear from Fig. 3 that the failure mode is mostly a function of the PTI portion of the test until the batch life-stage means approach 85% LC or 115% LC. When the batch is off-target by exactly 15% LC, failure by the life-stage means portion of the test occurs at a 50% rate, with the remaining probability due to the PTI-TOST. When the batch is off-target by more than 15% LC, failure is nearly certain.

Coverage Requirements for the FDA DDU Test

In this section, we examine coverage requirements implied by the FDA DDU test in order to provide a more

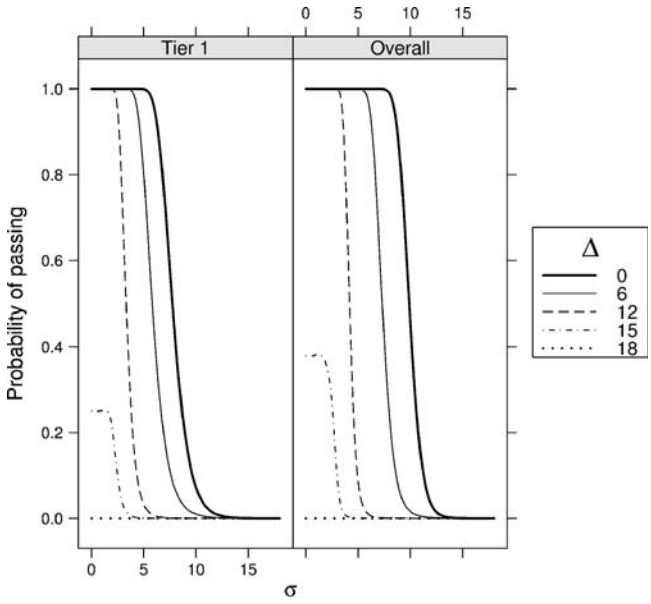

Fig. 2. Tier 1 and overall (both tiers) probability for FDA DDU test. The standard deviation (σ) is given on the horizontal axis. The *left* and *right panels*, respectively, show the tier 1 and overall (both tiers) probability of passing the test for various off-target deviations of the batch mean (Δ from 0% to 18%)

complete understanding of the performance of the test. A comparison of the acceptance probability from the perspective of coverage and tail areas is described in Figs. 4 and 5. Table I lists coverages required to pass the FDA DDU test with a given probability.

Figure 4 illustrates how each one-sided test performs individually with respect to the tail areas (top panel) and with respect to coverage (bottom panel). The top panel in Fig. 4

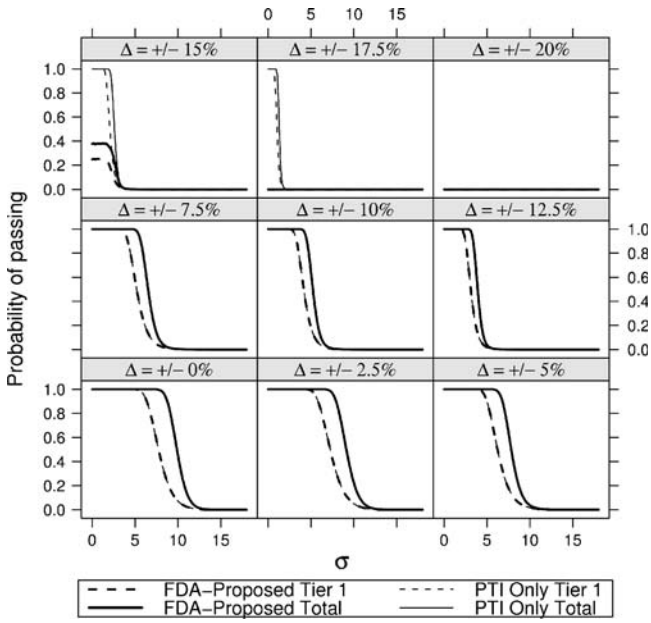

Fig. 3. Tier 1 (*dashed lines*) and total (*solid lines*) probability of passing the PTI-TOST with (*thick lines*) and without (*thin lines*) the life-stage criteria. For most of the panels, the lines overlap. The standard deviation (σ) is given on the *horizontal axis*. The various panels show different off-target deviations of the batch mean (Δ)

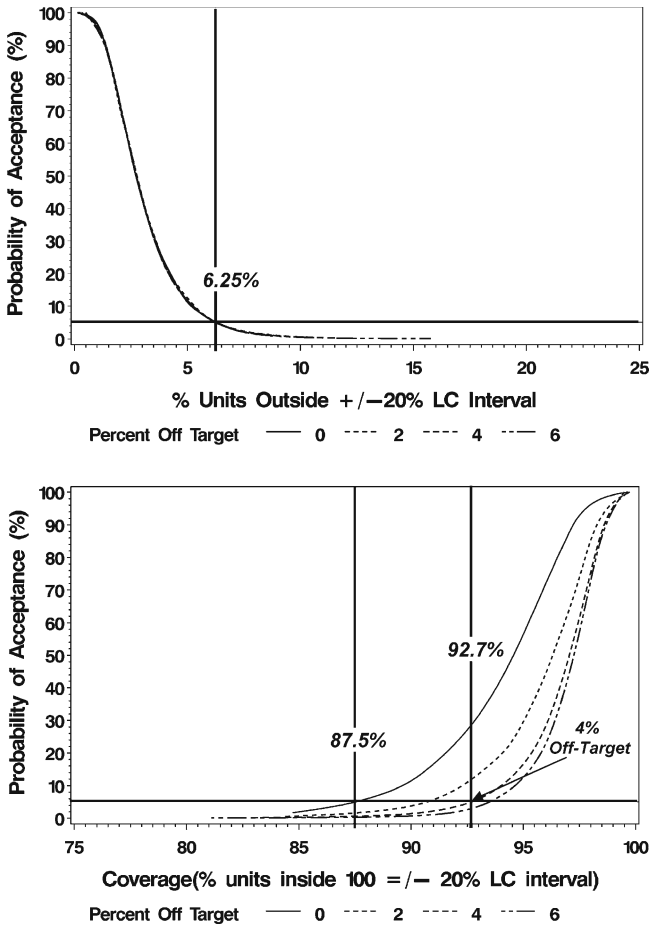

Fig. 4. OC curves for a single one-sided test as a function of the tail area (*top panel*) and of coverage (*bottom panel*). The different batch mean deviations from target are indicated by *solid* and *dashed lines*. All curves in the top panel overlap

confirms that the FDA DDU test controls the tail area as designed. The acceptance rate is 5% when the tail area is 6.25% regardless of the batch mean so that the OC curves exactly overlap. The bottom panel on Fig. 4 presents the probability of acceptance as a function of coverage, which is the proportion of the DDU values within the target interval. As the non-overlapping OC curves demonstrate, the coverage requirement for the one-sided test changes with the batch mean. For instance, when the batch is exactly on target (solid line) and has an 87.5% coverage (vertical line), it will pass the single one-sided (one-tail) test with 5% probability (horizontal line). For off-target batches (dotted lines), higher coverages (on the *x*-axis) are required to pass the single "one-tail" test with the same 5% probability. For example, a batch 4% off-target must have 92.7% coverage to pass the single "one-tail" test with 5% probability.

Similarly, Fig. 5 presents OC curves for the PTI-TOST, composed of TOST. The top panel shows the OC curves as a function of the tail area (%units in the DDU distribution outside 100±20% LC). When two single-tail tests are combined, the OC curves are not overlapping, and the probability of acceptance for batches with tail areas that do not exceed 6.25% is far below the 5% observed for each single-tail test. For a batch on target, each of the tails must be no more than 4.25% in order to have 5% acceptance

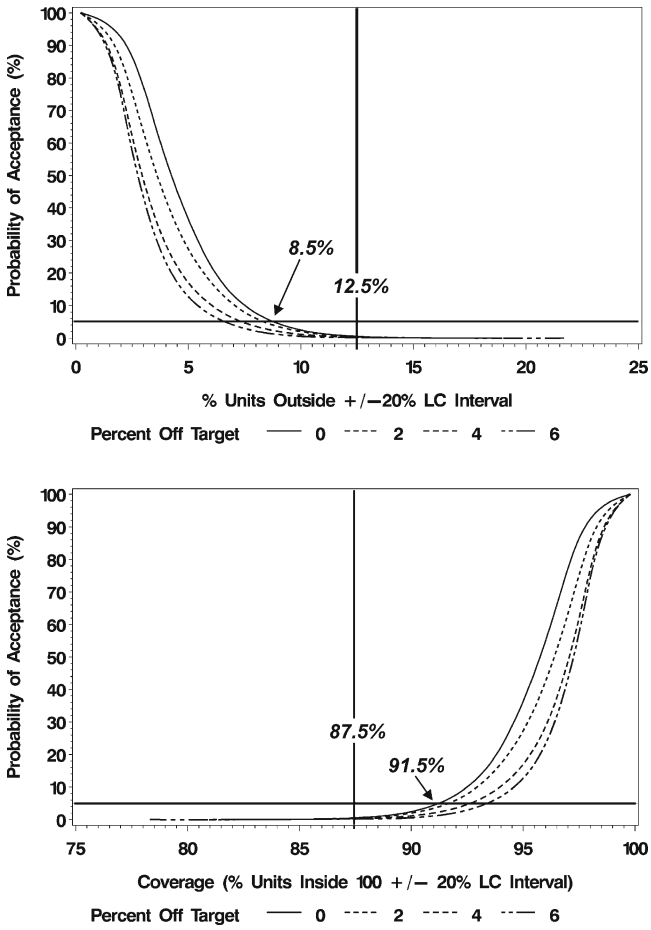

Fig. 5. OC curves for the PTI-TOST as a function of the tail area (*top panel*) and of coverage (*bottom panel*). The different batch mean deviations from target are indicated by *solid and dashed lines*

probability (i.e., 91.5% coverage is required). This effect is due to multiplicity as the testing of each tail independently increases the probability of failure. The bottom panel presents the OC curves as a function of the batch coverage. The coverage required for a 5% acceptance is 91.5% and increases for off-target batches. For example, for a batch 4% off target, 92.7% coverage is needed to pass with 5% rate. There is a hypothetical possibility for the coverage requirement to decrease when the deviation off-target increases (see Appendix).

Table I includes specific numbers for a range of acceptance rates for the complete FDA DDU test and could be used by sponsors as a reference tool when developing target performance criteria and controls for their products.

DISCUSSION

The test characterized in this article was proposed by the FDA after an intense effort spearheaded by the joint working group under the ACPS, comprised of scientists from FDA and the International Pharmaceutical Aerosol Consortium on Regulation & Science (IPAC-RS). During these discussions, IPAC-RS presented to FDA several different PTI approaches, including the Lieberman–Resnikoff method (7), polygon approximations of the Lieberman–Resnikoff method, PTI test with the "zone of indifference" as in USP <905> (8,9), PTI test with an additional criterion on the sample's maximum standard deviation (as in the IPAC-RS 2001 proposal) (10), and others. These options were developed in order to ensure constant batch coverage for a given pass rate, regardless of the batch mean. If a test for coverage is desired, those and other options (11–13) might be considered. In contrast, the FDA chose an approach to control product quality not directly through coverage but through controlling the maximum allowed proportion of the DDU distribution in either tail.

By design, and as the results presented in this paper illustrate, the performance of the PTI-TOST is determined not by the coverage (the proportion of DDU observations

Table I. Coverage Requirements for Given Acceptance Probabilities for PTI-TOST ($Pmax_{TA}$=6.25%, alpha=5%, N=20+40, 80–120% LC Target Interval; K_1=2.448, K_2=1.940; Means for BOU and EOU Within 100±15%)

Acceptance probability (%)	Batch mean deviation from target (% LC)	Batch coverage (%)	Batch standard deviation
99.9	0	99.3	7.4
	2	99.4	6.9
	4	99.6	6.1
	6	99.5	5.3
	8	99.5	4.6
	10	99.6	3.8
99	0	98.8	8.0
	2	99.2	7.3
	4	99.3	6.5
	6	99.3	5.7
	8	99.3	4.9
	10	99.3	4.1
98	0	98.6	8.1
	2	99.0	7.5
	4	99.2	6.7
	6	99.2	5.8
	8	99.2	5.0
	10	99.2	4.2
95	0	98.2	8.5
	2	98.7	7.8
	4	98.9	7.0
	6	98.9	6.1
	8	98.9	5.2
	10	98.9	4.3
50	0	95.7	9.9
	2	96.4	9.3
	4	97.1	8.3
	6	97.2	7.3
	8	97.3	6.2
	10	97.3	5.2
5	0	91.3	11.7
	2	91.8	11.3
	4	92.8	10.4
	6	93.6	9.1
	8	93.7	7.8
	10	93.8	6.5
1	0	88.5	12.7
	2	88.8	12.4
	4	89.7	11.6
	6	90.7	10.3
	8	91.1	8.9
	10	91.2	7.4

PTI-TOST parametric tolerance interval two one-sided tests, *LC* label claim, *BOU* beginning of unit, *EOU* end of unit

inside the target interval) but by the proportion of DDU observations in either tail, tested separately. For clarity, it would be beneficial therefore to refer to the PTI-TOST by its maximum allowable tail area $Pmax_{TA}$ (e.g., 6.25%) rather than by the coverage (e.g., 87.5%). The fact that, for a specific case (a batch with the mean of 100%LC), the overlap of two one-sided intervals ($100-2\times6.25\%=87.5\%$) happens to equal the coverage typically used in PTI tests is coincidental. As the batch moves off-target, the coverage requirement increases (Table I).

Moreover, the results illustrate that very high batch coverages (99% or greater) are needed to comply with the studied test for a product to be commercially viable (acceptance probability over 98%), as shown in Table I. The practical implication is that this test provides an incentive to have batches on target with very low variability, most of the time. In this case, relatively small sample sizes and high pass rates will be the norm. If batches are off-target and/or have high variability (e.g., due to analytical variability for low-dose products), larger sample sizes will be required to accurately characterize the batch quality.

The results (Fig. 3) also indicate that the life-stage mean criteria have little impact on the test outcomes except in extreme cases (i.e., when batches' means consistently deviate from target by close to 15% LC and/or the differences between BOU and EOU means are consistently large). The OC curves for the PTI-TOST with and without the additional life-stage criteria overlap for both tiers until the batch mean's deviation from target reaches 15% LC (upper row in Fig. 3). The impact of a non-zero life-stage mean difference (i.e., a deviation from the basic assumptions addressed in Part 1) will be addressed in Part 3 of this series.

The accompanying papers (Parts 2 and 3) address the effects of changing test parameters and the robustness of the test. In particular, those papers show that increasing sample size, target interval, or maximum allowable tail area increases the probability of acceptance and that the PTI-TOST is reasonably robust to most likely types of deviations from normality.

The presented characterization applies to a single instance of testing. When products are tested on stability, the same batch is tested multiple times, and therefore, its overall probability of acceptance will be lower even if the batch quality does not change. This multiplicity problem (14) is not unique to the PTI test but should be taken into account when proposing an appropriate testing scheme and acceptance criteria.

Characterizing other types of PTI tests and commenting on the PTI-TOST acceptance criteria are outside the scope of this paper.

CONCLUSIONS

The test proposed by FDA for control of delivered dose uniformity in OINDPs is a two one-sided PTI test, which controls the maximum allowable proportion in each tail (areas outside the target interval) of the DDU distribution, tested separately, rather than controlling coverage of the target interval directly. One of the consequences of this test's construction is that the required batch coverages are higher than would be required based on a direct coverage approach. In addition, because the test uses the simplest PTI form, the coverage required for passing with any given probability depends on the batch mean.

ACKNOWLEDGEMENTS

The authors thank the IPAC-RS Board and the IPAC-RS DDU Working Group for the consistent interest in this work and helpful feedback during manuscript preparation. The authors are also grateful to FDA for the opportunity to participate in the joint ACPS subgroup and interact to develop a PTI approach for control of dose uniformity in OINDP. Special thanks go to Bo Olsson and Dennis Sandell, whose work and vision inspired the IPAC-RS explorations of improved DDU tests for OINDP. Finally, we acknowledge Walter Hauck for his original proposal to use parametric approaches for DDU testing, starting with his 1999 presentation at the Management Forum Conference on European/FDA Regulatory Issues in Oral Inhalation and Nasal Delivery. This article presents a factual description and analysis of the test proposed by the FDA and should not be construed as endorsement or advocacy by the authors or organizations with which they are affiliated.

APPENDIX

Parametric Tolerance Intervals

A *tolerance interval* is a statistical confidence interval for a specified proportion of a population. A *tolerance interval test* is a statistical procedure that employs one or more tolerance intervals to compare competing hypotheses in order to examine specific characteristics of a population. The first tolerance interval publication is credited to Wilks (15). Later, Wald and Wolfowitz (16) wrote a pioneering paper on PTIs using normal distribution theory. Hahn and Meeker (17) published an overview of various statistical intervals, including the parametric and non-parametric tolerance interval. For more general methods to compute PTIs (such as a situation involving variance components), see, for example, Wolfinger (18) and Liao *et al.* (19). The PTI test proposed by FDA involves two one-sided tolerance intervals (20,21); the description used in this study follows from normal distribution theory, as given in Section 11.2 of Odeh and Owen (22).

The goal of a tolerance interval test is to determine whether a pre-specified proportion (coverage) of values of an attribute (e.g., DDU) in a certain population (e.g., a manufactured batch) falls inside a target interval $[L, U]$, where the values L and U should be set based on the product's development or historical data for that attribute (e.g., DDU) and clinical goals. The intervals of interest could be specified as sets of inequalities, such as two one-sided "observations$>L$" or "observations$<U$" or a two-sided "$L<$ observations$<$U." The portions of the distribution outside the target interval are typically called *tails*.

To illustrate the tolerance interval test concept, suppose there is a need to determine whether at least 87.5% of DDU values from a batch fall within a pre-specified target interval. If the entire batch were sampled, the proportion of observations falling within the target interval could be determined directly. For a destructive test such as DDU, it is not practical to sample the entire batch. If 20 DDU observations were made on randomly selected units from a batch, a tolerance interval could be constructed, which contains at least 87.5%

of units in the batch. Figure 6 (left panel) shows the density of a normal distribution with μ=100 and σ=5. The area between the dashed vertical lines contains the middle 87.5% of the distribution. A two-sided tolerance interval can be constructed to contain the middle 87.5% (for example, see Wald and Wolfowitz). Alternatively, two one-sided tolerance intervals could be constructed to contain, respectively, the lower and upper 6.25% "tail" quantiles. By containing the lower and upper 6.25% quantiles, the combined endpoints of the two one-sided intervals contain at least 87.5% of the distribution. The right panel of Fig. 6 shows the combined endpoints of two one-sided tolerance intervals for the middle 87.5% of the batch, as derived from a randomly generated sample with 20 observations taken from a normal distribution with μ=100 and σ=5.

Given the data represented in the right-hand panel of Fig. 6, if the target interval was [90, 110], the batch would fail the tolerance interval test since one or more (in this case both) endpoints of the one-sided tolerance intervals are outside of the target interval. On the other hand, given a target interval of [80, 120], the batch would pass since both endpoints of the one-sided tolerance intervals are contained inside the target interval.

Assuming that the data follow a normal distribution, the respective formulae for lower and upper one-sided lower tolerance intervals are given by $\overline{X} - K\,s$ ("observations>L") and $\overline{X} + K\,s$ ("observations<U"), where $\overline{X}$ is the sample mean, s is the sample standard deviation, and K is a function (to be given in the following section) of sample size, confidence level, and coverage.

Typically, confidence and coverage for a tolerance interval are denoted by 100(1−α)% confidence/100p% coverage, and the values for α and p are set during design of the test.

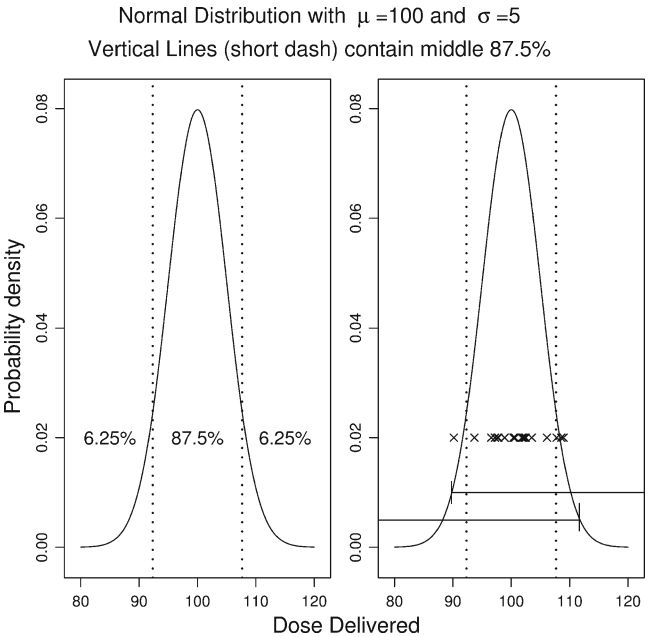

Fig. 6. The probability density function of a normal distribution with mean 100 and standard deviation 5. The area between the *dashed vertical lines* contains the 87.5% of the population. The *right panel* also shows 20 observations sampled from this normal distribution (*crosses*) and two one-sided tolerance intervals (*horizontal segments*) that, combined, contain the middle 87.5% of the population. Each interval was constructed with 95% confidence

In two-sided PTI tests, where the focus of the test is on the middle portion, the proportion p is related to the *limiting quality* [also known as reject quality level (RQL), unacceptable quality level (UQL), lot tolerance percent defective (LTPD), or minimum acceptable quality level (MAQL)].

The PTI-TOST employs two one-sided 100(1−α)%/100(1+p)/2% tolerance intervals for hypothesis testing. In this case, α is the significance level of the test (maximum type I error), which is formally defined as the largest probability that, upon testing a sample of data, the null hypothesis of a given test is rejected when in fact the null hypothesis is true. In the PTI-TOST, there is no direct translation between p and the requirement on coverage. Even though the two tails and the middle portion of a population complement each other, the statistical properties of a two one-sided "tail" test are different from those of a two-sided "coverage" test. A TOST imposes higher coverage requirements on the batch than a two-sided test using the same p. In place of confidence and coverage, therefore, it is more accurate to describe the PTI-TOST by significance level (α) and *maximum allowable tail area* (Pmax_{TA}).

Tolerance Interval Test

This section shows how two one-sided tolerance intervals can be used as the test statistics for a hypothesis test such as the PTI-TOST. The notation and assumptions laid out in the main article will be used throughout the Appendix.

Given a closed target interval [L, U] and a maximum allowable tail-area Pmax_{TA}, consider the following two sets of hypotheses.

$H0_1$: More than 100 Pmax_{TA}% of the population<L
Ha_1: 100 Pmax_{TA}% or less of the population<L

and

$H0_2$: More than 100 Pmax_{TA}% of the population>U
Ha_2: 100 Pmax_{TA}% or less of the population>U

Taken together, these two sets of hypotheses form

H0: Either {more than 100 Pmax_{TA}% of the population <L} or {more than 100 Pmax_{TA}% of the population>U}
Ha: Both {100 Pmax_{TA}% or less of the population<L} and {100 Pmax_{TA}% or less of the population>U}

If both Ha_1 and Ha_2 are true, then reject H0 and conclude that at least 100p% of the population lies in the interval [L, U], where p=(1–2 Pmax_{TA}).

With a significance level α, the hypotheses can be examined via two one-sided 100(1−α)%/100(1−Pmax_{TA})% tolerance intervals. Let the two one-sided tolerance intervals be given by $T_L = \overline{X} - K\,s$ and $T_U = \overline{X} + K\,s$ where $K = T^{-1}\left(1-\alpha, N-1, Z_{1-Pmax_{TA}}\sqrt{N}\right)\frac{1}{\sqrt{N}}$, $T^{-1}(q, df, \text{ncp})$ is the inverse T cumulative distribution function taken at the qth quantile with df degrees of freedom and non-centrality parameter ncp, and Z_a is the inverse standard normal cumulative distribution function taken at the ath quantile. Note that as $N\to\infty$, $T_L \to \mu - Z_{1-Pmax_{TA}}\,\sigma$ and $T_U \to \mu + Z_{1-Pmax_{TA}}\,\sigma$.

If $T_L \geq L$, we reject H_{01}, and if $T_U \leq U$, we reject $H0_2$, and thus conclude that no more than 100 Pmax_{TA}% of the

population lies in either tail (i.e., no more than 100 $Pmax_{TA}$% is $<L$ and no more than 100 $Pmax_{TA}$% of the population is $>U$). This ensures that at least 100(1–2 $Pmax_{TA}$)% of the population lies between L and U. Each of the two sets of hypotheses uses significance level α. Because this testing strategy is an example of an intersection–union test (23), the significance level for H0 is also α.

The set of parameters that satisfy the null hypothesis (H0) is given by $\Theta = \{\mu, \sigma : \mu \le L + Z_{1-Pmax_{TA}}\sigma \text{ or } \mu \ge U - Z_{1-Pmax_{TA}}\sigma\}$. The significance level α is the largest probability of rejecting H0 when μ and σ satisfy Θ, or put into an equation: $\alpha = \sup_{\theta\in\Theta} \Pr(T_L > L, T_U < U|\theta)$. For sufficiently small σ, the probability is maximized when $\mu = L + Z_{1-Pmax_{TA}}\sigma$ or when $\mu = U - Z_{1-Pmax_{TA}}\sigma$. It follows that

$$\begin{aligned}\alpha &= \Pr(T_L > L, T_U < U|\mu = L + Z_{1-Pmax_{TA}}\sigma, \sigma = small)\\ &= \Pr(T_L > L|\mu = L + Z_{1-Pmax_{TA}}\sigma, \sigma = small)\\ &= \Pr(T_U < U|\mu = U - Z_{1-Pmax_{TA}}\sigma, \sigma = small)\end{aligned}$$

Any other set of parameters (μ, σ) in the null hypothesis space will result in a type I error smaller than α.

To examine this concept further, consider two populations and assume that α=0.05, $Pmax_{TA}$=6.25%, [L, U]=[80, 120]. For the first population, μ=85 and σ=3.26 (small relative to [L, U]), so that the area of the left tail is 6.25% and the area of the right tail is approximately 0%. This population is (just barely) contained in Θ, and the probability to reject the null hypothesis is approximately 5%. For the second population, μ=100 and σ=13.04 so that each of the tail areas is 6.25%. This second population is also (just barely) contained in Θ, but the probability to reject the null hypothesis is less than 1%. Even though the two given scenarios (μ=85, σ=3.26) and (μ=100, σ=13.04) are both considered to represent limiting quality under the PTI-TOST, they have drastically different probabilities of being accepted because the first distribution reaches the maximum allowable tail area in only one tail, while the second distribution reaches the maximum allowable tail area in both tails. This feature of the PTI-TOST is illustrated in Fig. 7.

In Fig. 7, the solid and dashed curves illustrate the distributions of two batches of extreme limiting quality. The dashed curve has a total of only $Pmax_{TA}$ in its tails, while the solid curve has a total 2 $Pmax_{TA}$ in its tails. The two curves, respectively, exemplify the maximum (probability=0.05) and minimum (probability=0.0025) type I error from batches of extreme limiting quality. Although the PTI-TOST logically rejects batches with 2 $Pmax_{TA}$ in its tails more often than batches with $Pmax_{TA}$ in its tails, it may be considered an odd feature of the test that, while both batches meet the limiting quality standard, the batch that is greatly off-target has a much lower chance of being rejected than the on-target batch.

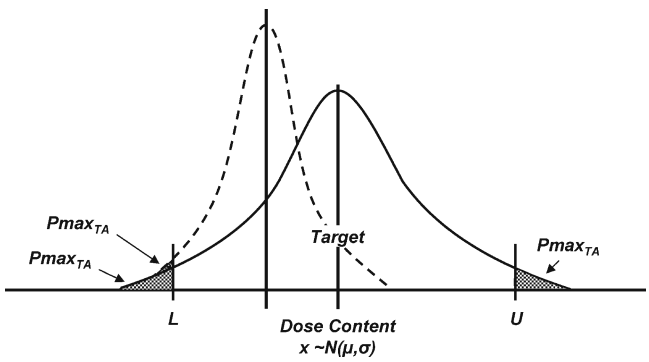

Fig. 7. The *solid and dashed curves* illustrate the distributions of two batches of extreme limiting quality

Significance Level for the PTI-TOST

The PTI-TOST is a "sequential test" in which second-tier statistics are computed on first-tier data plus additional second-tier units. In a two-tier sequential test, a set of hypotheses (a null and an alternative) are compared, first with N_1 observations. Should the null hypothesis fail to be rejected, additional N_2 observations are collected and the hypotheses are compared again with N_1+N_2 observations. To achieve an overall (both-tiers) significance level of α=0.05, an "alpha-spending" scheme must be put into place. The significance levels for each of the tiers 1 and 2 (α_1 and α_2) will be smaller than 0.05, their sum not equal to 0.05, and the exact values will depend on the alpha-spending technique chosen. An overview of alpha-spending methods was published by Li and DeMets, (24) who showed that the first- and second-tier values for significance level generally depend on the overall (both tiers) α and on the ratio of $N_1/(N_1+N_2)$.

For the DDU test, the FDA proposed choosing α_1 and α_2 by the Lan and DeMets (25) approach using the Pocock-alpha spending function. This approach is identical to that used by Hauck and Shaikh (26) who examined multi-stage testing for a two-sided PTI test. The R (27) library "ldbounds" provides a fast way to compute α_1 and α_2 by the Lan and DeMets method. The website http://www.biostat.wisc.edu/landemets/ also provides free software for computing the Lan and DeMets bounds.

As illustrated in the main paper, for the most part, the non-parametric test of the life-stage means has little impact on the PTI-TOST outcomes. In the previous section "Tolerance interval test," the null hypothesis space (without consideration for the life-stage means test) was given by $\Theta = \{\mu, \sigma : \mu \le L + Z_{1-Pmax_{TA}}\sigma \text{ or } \mu \ge U - Z_{1-Pmax_{TA}}\sigma\}$. The significance level for the FDA PTI-TOST will be smaller (though in practice, only negligibly smaller) as can be seen in the follow equation:

Significance level

$$\begin{aligned}&= \sup_{\mu,\sigma} \Pr\left(\begin{array}{l} T_L > 80,\ T_U < 120, 85 < \overline{X}_{BOU} \\ \quad < 115, 85 < \overline{X}_{EOU} < 115 | H0 \text{ is true}\end{array}\right)\\ &\le \sup_{\theta\in\Theta} \Pr(T_L > 80, T_U < 120|\theta)\end{aligned}$$

Because of the non-parametric life-stage means tests, the tier-1 and overall type I errors are not controlled exactly but have respective upper bounds of α_1 and α.

Using the Pocock alpha-spending function, the FDA-proposed method for choosing α_1 and α_2 is easily generalized to different sampling plans (e.g., N_1=20 and N_2=20). Users should keep in mind that different implementations of the Lan–DeMets method may yield slightly different values for α_1

and α_2 (and, consequently, different corresponding K values). For the purposes of this study, for $\alpha=0.05$ and $N_1/(N_1+N_2)=1/3$, the values $\alpha_1=0.0226$ and $\alpha_2=0.0340$ were used.

REFERENCES

1. FDA/CDER (2008) Draft guidance for industry "Metered dose inhaler (MDI) and dry powder inhaler (DPI) drug products chemistry, manufacturing, and controls documentation", October 1998. http://www.fda.gov/cder/guidance/2180dft.pdf. Accessed September 11, 2008.
2. FDA/CDER (2008) Guidance for industry "Nasal spray and inhalation solution, suspension, and spray drug products chemistry, manufacturing, and controls documentation". Draft: May 1999. http://www.fda.gov/ohrms/dockets/ac/00/backgrd/3609b1k.pdf. Accessed September 11, 2008. Final: July 2002. http://www.fda.gov/cder/guidance/4234fnl.pdf. Accessed September 11, 2008.
3. Murphy JR, Griffiths KL. Zero-tolerance criteria do not assure product quality. Pharm Tech 2006;30(1):52–60. http://www.pharmtech.com/pharmtech/content/printContentPopup.jsp?id=283486. Accessed September 11, 2008.
4. M.M. Nasr. Parametric tolerance interval test for delivered dose uniformity (DDU). Presentation to Advisory Committee for Pharmaceutical Science on 25 October 2005. http://www.fda.gov./ohrms/dockets/ac/05/slides/2005-4187S1_13_Nasr.ppt. Accessed September 11, 2008.
5. R. Lostritto (2005) Advisory Committee for Pharmaceutical Science Meeting (in transcripts), 25 October 2005, p. 361. http://www.fda.gov/ohrms/dockets/ac/05/transcripts/2005-4187T1.pdf. Accessed September 11, 2008.
6. Montgomery DC. Introduction to statistical quality control. 5th ed. New York, NY: Wiley; 2004.
7. Lieberman GJ, Resnikoff GJ. Sampling plans for inspection by variables (Corr: p1333). J Am Stat Assoc. 1955;50:457–516.
8. USP <905>Uniformity of dosage units. The sixth interim revision announcement. Pharm Forum. (2006);32(6):1649–1659 (1st supplement to USP 30–NF 25).
9. USP. Explanatory note "USP-NF harmonized chapter <905>uniformity of dosage units." http://www.usp.org/USPNF/notices/generalChapter905.html. Accessed September 11, 2008.
10. IPAC-RS. A parametric tolerance interval test for improved control of delivered dose uniformity of orally inhaled and nasal drug products; 2001. http://ipacrs.com/PDFs/IPAC-RS_DDU_Proposal.PDF. Accessed September 11, 2008.
11. Flann B. Comparison of criteria for content uniformity. J Pharm Sci. 1974;63(2):183–99.
12. Hauck WW, Shaikh R. Modified two-sided normal tolerance intervals for batch acceptance of dose uniformity. Pharm Stat. 2004;3(2):89–97. doi:10.1002/pst.103.
13. Hauck WW, Shaikh R. Sample sizes for batch acceptance from single- and multistage designs using two-sided normal tolerance intervals with specified content. J Biopharm Stat. 2002;11(4):335–46.
14. Foust L, Diener M, Gorko MA, Hofer J, Larner G, LeBlond D, *et al.* Overcoming disincentives to process understanding in the pharmaceutical CMC environment. Pharm Technol. 2007;31(9):106–15.
15. Wilks SS. Determination of sample sizes for setting tolerance limits. Ann Math Stat. 1941;12:91–6.
16. Wald A, Wolfowitz J. Tolerance limits for a normal distribution. Ann Math Stat. 1946;17(2):208–215.
17. Hahn GJ, Meeker WQ. Statistical intervals: a guide for practitioners. New York: Wiley; 1991.
18. Wolfinger RD. Tolerance intervals for variance components models using bayesian simulation. J Qual Technol. 1998;30(1):18–32.
19. Liao CT, Lin TY, Iyer HK. One- and two-sided tolerance intervals for general balanced mixed models and unbalanced one-way random models. Technometrics. 2005;47:323–35.
20. Tsong Y, Shen M. Parametric two-stage sequential quality assurance test of dose content uniformity. J Biopharm Stat. 2007;17(1):143–57. doi:10.1080/10543400601001527.
21. Tsong Y, Shen M, Lostritto RT, Poochikian GK. Parametric two-tier sequential quality assurance test of delivery dose uniformity of multiple-dose inhaler and dry powder inhaler drug products. J Biopharm Stat. 2008;18(5):976–84. doi:10.1080/10543400802287222.
22. Odeh RE, Owen DB. Tables for normal tolerance limits, sampling plans, and screening. New York: Marcel Dekker; 1980.
23. Casella G, Berger RL. Statistical inference. 2nd ed. Belmont, CA: Duxbury; 2001.
24. Li Z, DeMets DL. On the bias of estimation of a Brownian motion drift following group sequential tests. Stat Sin. 1999;9:923–37.
25. Lan KKG. DeMets. DL. Discrete sequential boundaries for clinical trials. Biometrika. 1983;70:659–63.
26. Hauck W, Shaikh R. Sample sizes for batch acceptance from single- and multistage designs using two-sided normal tolerance intervals with specified content. J Biopharm Stat. 2001;11:335–46.
27. The R foundation for statistical computing. R version 2.6.1 (2007-11-26). ISBN 3-900051-07-0.