MARCEL DEKKER, INC. • 270 MADISON AVENUE • NEW YORK, NY 10016

JOURNAL OF BIOPHARMACEUTICAL STATISTICS
Vol. 12, No. 3, pp. 323–332, 2002

PROFILE ANALYSIS FOR ASSESSING IN VITRO BIOEQUIVALENCE

Bin Cheng* and Jun Shao

Department of Statistics, University of Wisconsin,
1210 W. Dayton Street, Madison, WI 53706

ABSTRACT

For locally acting drug products such as nasal aerosols and nasal sprays, therapeutic equivalence between two drug products may be established by in vitro bioequivalence studies based on measurements intended to reflect the rate and extent to which the active ingredient becomes available at the site of action. For cascade impaction or multistage liquid impinger for particle size distribution, profile analysis is required. However, we find that the analysis procedure described in the 1999 FDA guidance lacks statistical justification. In this article, we explain why FDA's approach is incorrect and propose a correct statistical method for profile analysis using the basic ideas in the FDA guidance.

Key Words: Nasal aerosols and nasal sprays; Lot-to-lot variation; Compound multinomial; Median

INTRODUCTION

Testing bioequivalence between a test drug product and a reference drug product is considered as a surrogate for clinical evaluation of the therapeutic

*Corresponding author. E-mail: bcheng@stat.wisc.edu

DOI: 10.1081/BIP-120014562 1054-3406 (Print); 1520-5711 (Online)

equivalence of the two drug products. For locally acting drug products such as nasal aerosols (e.g., metered-dose inhalers) and nasal sprays (e.g., metered-dose spray pumps) that are not intended to be absorbed into the bloodstream, the U.S. Food and Drug Administration (FDA) indicates that bioequivalence may be assessed, with suitable justification, by in vitro bioequivalence studies based on measurements intended to reflect the rate and extent to which the active ingredient or active moiety becomes available at the site of action. Although it is recognized that in vitro methods are less variable, easier to control, and more likely to detect differences between products if they exist, the clinical relevance of the in vitro tests is not clearly established until a guidance on bioavailability and bioequivalence studies for nasal aerosols and nasal sprays for local action was issued by the FDA (see Ref. [1]).

The FDA classifies assessments of six in vitro bioequivalence tests for nasal aerosols and sprays as either the nonprofile analysis or the profile analysis. The nonprofile analysis can be carried out using an approach similar to those for in vivo bioequivalence testing specified in Refs. [2,3] (see Ref. [4]). In this article, we focus on the profile analysis, which applies to cascade impaction or multistage liquid impinger for particle size distribution. As indicated in Ref. [1], bioequivalence should be assessed by the ratio of the profile variation between the test and reference products over the profile variation within the reference product, where the profile variation is obtained using a chi-square type difference (see "FDA's Approach"). A 95% upper confidence bound for the mean of this ratio is described in Ref. [1]. In vitro bioequivalence can be claimed if and only if the 95% upper confidence bound is lower than a bioequivalence limit set by the FDA.

The 95% upper confidence bound described in Ref. [1], however, lacks statistical justification. In fact, it is incorrect. The purpose of this article is to explain why FDA's approach is incorrect (see "FDA's Approach") and to propose a correct statistical method for profile analysis using the basic ideas in the FDA guidance (see "The Proposed Tests"). Some discussions are provided in the fourth section.

FDA's APPROACH

Droplet size distribution measurements are critical to delivery of drug to the nose, which could affect the efficiency of the nasal absorption. As FDA[1] indicated, sizing of droplets or particles by cascade impaction or multistage liquid impinger measures aerodynamic diameter based on inertial impaction, an important factor in the deposition of drug in the nasal passages. Cascade impaction or multistage liquid impinger drug deposition profile data should be reported in mass units and based on three size range groups. Group 1 includes summation of drug deposition in or on the valve stem, actuator, inlet port, and upper stage, which should have a nominal effective cutoff diameter (e.g., greater than or equal to 9.0, 10.0, 13.0, or 16.0 μm). Group 2 includes drug deposition on the stage

immediately below the upper stage (e.g., greater than or equal to 5.0 μm). Group 3 includes summation of drug deposition below the group 2 stage, including the filter. Further groups may be introduced if necessary. Thus, the observed profile from in vitro bioequivalence testing is a vector of S observations, where S is the number of groups.

According to FDA's guidance, in vitro bioequivalence testing should be carried out by addressing not only the variation among different canisters (bottles) of the drug product, but also the variation among different lots. Let $(X_{T1}, \ldots, X_{TS})$ be the observed profile vector of a sampled canister from a lot of the test product and $(X_{R1}, \ldots, X_{RS})$ and $(X_{R'1}, \ldots, X_{R'S})$ be the observed profile vectors of two canisters from two different lots of the reference product. The FDA[1] considered the following ratio in assessing the difference between the test and reference products:

$$\text{rd} = \frac{d_{TRR'}}{d_{RR'}}, \tag{1}$$

where

$$d_{TRR'} = \sum_i \frac{\left[X_{Ti} - \frac{1}{2}(X_{Ri} + X_{R'i})\right]^2}{X_{Ti} + \frac{1}{2}(X_{Ri} + X_{R'i})} \tag{2}$$

is a chi-square type measure of the profile difference between a test profile vector and two reference profile vectors from two different lots and

$$d_{RR'} = \sum_i \frac{(X_{Ri} - X_{R'i})^2}{\frac{1}{2}(X_{Ri} + X_{R'i})} \tag{3}$$

is a chi-square type measure of the profile difference between the two reference profile vectors from two different lots. The reason for using chi-square type measures is because FDA[1] assumed that profile vectors are distributed as "compound multinomial distributions" (i.e., distributions of multinomial random vectors contaminated by some random errors). To ensure that the quantities in Eqs. (1)–(3) are well defined, we need to set

$$\frac{\left[X_{Ti} - \frac{1}{2}(X_{Ri} + X_{R'i})\right]^2}{X_{Ti} + \frac{1}{2}(X_{Ri} + X_{R'i})}$$

or

$$\frac{(X_{Ri} - X_{R'i})^2}{\frac{1}{2}(X_{Ri} + X_{R'i})}$$

to 0 when $(X_{Ti}, X_{Ri}, X_{R'i}) = (0, 0, 0)$ [or $(X_{Ri}, X_{R'i}) = (0, 0)$] for a given i (which has a positive probability if profile vectors are multinomial or compound

multinomial) and define rd to be ∞ if $d_{RR'} = 0$ (which has a positive but usually small probability).

Assessing bioequivalence involves setting a bioequivalence criterion and constructing a statistical inference procedure to test whether the bioequivalence criterion is satisfied, based on some observed data. FDA[1] considered the expectation $\mathrm{Rd} = E(\mathrm{rd})$ and the criterion

$$\mathrm{Rd} < \theta_{\mathrm{BE}} \tag{4}$$

for bioequivalence [i.e., the test and reference products are in vitro bioequivalent if and only if Eq. (4) holds], where θ_{BE} is a bioequivalence limit determined by the FDA. If this criterion is adopted, then the required statistical procedure is a level 5% test of the hypotheses

$$H_0\colon \mathrm{Rd} \geq \theta_{\mathrm{BE}} \ \text{ vs. } \ H_1\colon \mathrm{Rd} < \theta_{\mathrm{BE}}. \tag{5}$$

Note that if we have a 95% upper confidence bound for Rd, then a level 5% test rejects H_0 (i.e., claims bioequivalence) if and only if the bound is less than θ_{BE}.

FDA[1] recommended the following procedure for testing Eq. (5). The study design consists of 3 lots from the test product, 3 lots from the reference product, and at least 10 canisters from each lot. Thus, a total of 60 canisters are tested and profile vectors are observed. The six lots can be matched into different combinations of lot-triplets, each of which contains one test lot and two different reference lots. Within each lot-triplet, the 30 canisters can be matched into different combinations of canister-triplets, each of which contains one test canister and two reference canisters from different lots. A random sample of 500 canister-triplets are selected from the population of all possible different canister-triplets. Let rd_i be the ratio in Eq. (1) computed based on the ith canister-triplet in the random sample, $i = 1, \ldots, 500$, and $\overline{\mathrm{Rd}}_{95}$ be the 95th percentile of $\mathrm{rd}_1, \ldots, \mathrm{rd}_{500}$. FDA[1] called $\overline{\mathrm{Rd}}_{95}$ a 95% upper bound of Rd and recommended it for testing Eq. (5).

If $\overline{\mathrm{Rd}}_{95}$ is indeed a 95% upper confidence bound for Rd, then FDA's procedure is statistically justified: the probability of rejecting H_0 in Eq. (5) when H_0 is true is smaller than 5%, i.e., there is a 95% statistical assurance when we conclude that the test and reference products are bioequivalent. However, $\overline{\mathrm{Rd}}_{95}$ is not even close to a 95% upper confidence bound.

First of all, $\mathrm{Rd} = E(\mathrm{rd})$ may be always equal to ∞, especially when the profile vectors are multinomial or compound multinomial as suggested in Ref. [1]. If $\mathrm{Rd} = \infty$, then $\overline{\mathrm{Rd}}_{95}$ is certainly not a 95% upper confidence bound for Rd and, in fact, testing Eq. (5) is meaningless. This problem can be fixed by defining Rd in Eq. (5) to be another location characteristic of rd other than the mean of rd, for example, the median or a percentile of rd.

Even if the mean of rd is well defined or Rd is defined to be the median of rd, $\overline{\mathrm{Rd}}_{95}$ is not a 95% upper confidence bound for Rd. At the first glance, one might think that FDA's procedure is closely related to randomization tests or resampling

methods such as the bootstrap. However, FDA's procedure is not a randomization test for the following two main reasons:

1. In a randomization test procedure, a "reference" distribution is constructed based on values having the same distribution under the null hypothesis H_0. In FDA's procedure, however, the reference distribution consists of all rd values from different canister-triplets, which are not identically distributed because of lot-to-lot variation.
2. In a randomization test, an observed value is compared to the "reference" distribution, whereas in Ref. [1], the bioequivalence limit θ_{BE} is compared with an estimated 95th percentile of the reference distribution.

The fact that $\overline{\mathrm{Rd}}_{95}$ is a 95th percentile suggests that FDA's procedure may be similar to the bootstrap percentile method.[5] However, because different lot-triplets share at least one common reference lot, the rd values are not independent. It is well known that any bootstrap method designed for independent data does not work for dependent data. In fact, under the dependence structure of rd values, it is very difficult to derive a valid bootstrap method.

We conducted a simulation study to study the behavior of FDA's procedure. Compound multinomial distributions were used to generate profile vectors. Let $(M_{kj1}, M_{kj2}, M_{kj3})$ be a random vector distributed as multinomial $(100, P_{kj1}, P_{kj2}, P_{kj3})$, and ε_{kji}s be independent random errors having gamma distributions with mean 1 and variances σ_k^2, where $j\ (=1,2,3)$ is the index for lots, $k\ (=T,R)$ is the index for test or reference, and ε_{kji}s and M_{kji}s are independent. The profile vectors are given by

$$(X_{kj1}, X_{kj2}, X_{kj3}) = (\varepsilon_{kj1}M_{kj1}, \varepsilon_{kj2}M_{kj2}, \varepsilon_{kj3}M_{kj3}).$$

The values of P_{kji} were obtained from estimates in a real data set:

$$(P_{T11}, P_{T12}, P_{T13}) = (0.99735, 0.00235, 0.00030),$$

$$(P_{T21}, P_{T22}, P_{T23}) = (0.99780, 0.00150, 0.00070),$$

$$(P_{T31}, P_{T32}, P_{T33}) = (0.99670, 0.00285, 0.00045),$$

$$(P_{R11}, P_{R12}, P_{R13}) = (0.99805, 0.00170, 0.00025),$$

$$(P_{R21}, P_{R22}, P_{R23}) = (0.99760, 0.00155, 0.00085),$$

$$(P_{R31}, P_{R32}, P_{R33}) = (0.99675, 0.00240, 0.00085).$$

Four different values of (σ_T, σ_R) were considered:

$$(0.05, 0.05), (0.05, 0.10), (0.10, 0.05), (0.10, 0.10).$$

Table 1. Simulation Results for FDA's Upper Bound $\overline{\text{Rd}}_{95}$[a]

	(σ_T, σ_R)			
	(0.05, 0.05)	(0.05, 0.10)	(0.10, 0.05)	(0.10, 0.10)
Median of rd	0.6753	0.5378	1.3305	0.7684
95th percentile of rd	117.3226	44.1099	230.5589	70.3797
Frequency of $\overline{\text{Rd}}_{95}$ < median	0	0	0	0
Frequency of $\overline{\text{Rd}}_{95} < 2$	0	0	0	0
Frequency of $\overline{\text{Rd}}_{95}$ < 95th percentile	0.5628	0.0908	0.9626	0.5368

[a]Number of simulations for median and 95th percentile = 10,000,000; Number of simulations for frequencies = 5000.

For each fixed value of (σ_T, σ_R), the random variable rd is defined according to Eqs. (1)–(3) with the triplet T, R, R' randomly selected from the following nine possible lot-triplets:

$$\begin{array}{lll} T1,R1,R2 & T1,R1,R3 & T1,R2,R3 \\ T2,R1,R2 & T2,R1,R3 & T2,R2,R3 \\ T3,R1,R2 & T3,R1,R3 & T3,R2,R3 \end{array} \tag{6}$$

Table 1 lists the median and 95th percentile of rd for each value of (σ_T, σ_R). Also included in Table 1 are simulation frequencies of $\overline{\text{Rd}}_{95}$ < the median, 2, or the 95th percentile. When the median of rd is considered as the parameter Rd in Eq. (5), the frequency of $\overline{\text{Rd}}_{95}$ < the median reported in Table 1 is a type I error probability of FDA's test procedure if θ_{BE} = the median, whereas the frequency of $\overline{\text{Rd}}_{95} < 2$ in Table 1 is a value of the power of FDA's test. Since all values of the median are substantially smaller than 2, the result in Table 1 indicates that FDA's test is very conservative. On the other hand, when the 95th percentile of rd is considered as the Rd in Table 1 and $\theta_{\text{BE}} = \text{Rd}$, the frequency of $\overline{\text{Rd}}_{95}$ < the 95th percentile in the Table is a type I error probability of FDA's test and the result in Table 1 shows that FDA's procedure is incorrect and too liberal. Even if $\overline{\text{Rd}}_{95}$ is viewed as an estimator of the 95th percentile of rd, it is an incorrect estimator since the frequency of $\overline{\text{Rd}}_{95}$ < the 95th percentile may be far away from 0.5, e.g., in the cases where $(\sigma_T, \sigma_R) = (0.05, 0.10)$ and $(\sigma_T, \sigma_R) = (0.10, 0.05)$.

THE PROPOSED TESTS

With three lots from each of the test and reference products, we propose a test procedure for in vitro bioequivalence when Rd in Eq. (5) is considered to be a percentile of the ratio rd defined in Eq. (1). For illustration, we choose the median of rd as the parameter Rd.

Our method requires a two-stage sampling procedure. First, we randomly select a lot-triplet with replacement from the nine possible lot-triplets given in Eq. (6). With the sampled lot-triplet, we randomly select one canister from each lot and calculate the ratio rd according to Eqs. (1)–(3). This process is repeated independently n times to produce independent and identically distributed $\mathrm{rd}_1, \ldots, \mathrm{rd}_n$. Let F be the distribution of rd_j and Rd be the median of F. Then, Eq. (5) is equivalent to

$$H_0\colon 0.5 \geq F(\theta_{\mathrm{BE}}) \ \text{ vs. } \ H_1\colon 0.5 < F(\theta_{\mathrm{BE}}). \tag{7}$$

Let Y be the number of rd_js $\leq \theta_{\mathrm{BE}}$. Then Y has the binomial distribution with size n and probability $F(\theta_{\mathrm{BE}})$. A level α test for Eq. (7) rejects H_0 in Eq. (7) if and only if $Y > b_{n,\alpha}$, where $b_{n,\alpha}$ is an integer satisfying

$$\sum_{k=0}^{b_{n,\alpha}} \binom{n}{k} \frac{1}{2^n} = 1 - \alpha \tag{8}$$

and α is a positive number that is close to 5%. If $n = 18$, for example, we may take $\alpha = 0.0481$, which gives $b_{n,\alpha} = 12$. That is, the bioequivalence can be claimed with 95.19% statistical assurance when at least 12 of 18 canister-triplets have the rd values $\leq \theta_{\mathrm{BE}}$.

Note that this test procedure ensures that the level of the test is exactly α. There are several key differences between our proposed test and FDA's test. First, in our procedure the canister-triplets that produce rd values are nonoverlapped so that $\mathrm{rd}_1, \ldots, \mathrm{rd}_n$ are independent, whereas in FDA's procedure the selected 500 canister-triplets are overlapped so that $\mathrm{rd}_1, \ldots, \mathrm{rd}_{500}$ are dependent. Second, If n in our procedure is 20 (or nearly 20), then there are a total of $3n = 60$ (or nearly 60) sampled canisters, which is the same as that in FDA's procedure. However, in our procedure, the ratio of test canisters over the reference canisters is 1/2, whereas this ratio is 1 in FDA's procedure. Finally, if the total number of sampled canisters is the same, then our procedure requires n computations n rd values, which is much smaller than the 500 required in FDA's procedure.

Since sampling of lot-triplets is random in our procedure, the actual number of sampled canisters from each lot (or lot-triplet) is random. This is also true for FDA's procedure. Although 10 canisters are sampled from each lot in FDA's procedure, the number of canisters from each lot (or lot-triplet) is still random in the 500 randomly selected canister-triplets. To use a more balanced sampling design, we consider the following modification. Assume that in each test lot we sample $3k$ canisters and in each reference lot we sample $6k$ canisters. Then, we randomly group these canisters into $9k$ nonoverlapped canister-triplets, with one test canister and two reference canisters from different lots in each canister-triplet. The rd value of each canister-triplet is calculated, which results in $\mathrm{rd}_1, \ldots, \mathrm{rd}_{9k}$. Note that these rd values are identically distributed, but are slightly dependent. Let Z be the number of rd_js $\leq \theta_{\mathrm{BE}}$. Then, Z is not exactly binomial, but is close to

binomial. If we still reject the null hypothesis H_0 in Eq. (7) when $Z > b_{n,\alpha}$, where $b_{n,\alpha}$ is defined by Eq. (8), then the test is approximately of level α.

Properties of this method (which can be called a random grouping method) and the previous method (which can be called a two-stage random sampling method) are investigated in a simulation study. The same setting as that in the simulation in "FDA's Approach" was used. We considered $n = 9k = 18$ ($k = 2$ and $b_{n,\alpha} = 12$), since this leads to a total of 54 canisters, which is close to the total of 60 canisters in FDA's procedure. Results based on 5000 simulations are given in Table 2. The following is a summary of the results in Table 2.

1. Level of the test based on the RS method. When θ_{BE} is chosen to be the median of rd (given in Table 1), the probabilities in Table 2 are type I error probabilities. The nominal value is $\alpha = 0.0481$ ($n = 18$). Hence, the results in Table 2 are within the nominal value $\pm 0.006 = \pm 2$ estimated standard simulation error.
2. Level of the test based on the RG method. In all cases under consideration, the type I error probabilities are slightly larger than the nominal value.
3. Power of the tests. When $\theta_{\text{BE}} = 1, 1.5$, or 2, the results in Table 2 are values of power except for the case where $\theta_{\text{BE}} = 1$ and $(\sigma_T, \sigma_R) = (0.10, 0.05)$. It can be seen that a reasonable power can be achieved when θ_{BE} is substantially larger than the true median of rd or the test variability (σ_T) is smaller than the reference variability (σ_R). Note that power can also be increased by increasing n or k.

In general, the sampling performance of the test based on the RS method is better than that of the test based on the RG method. The RG method, however, is more balanced in the sense that the number of canisters from each lot-triplet is a constant ($3k$).

Table 2. Simulation Results on the Rejection Probability for the Two-Stage Random Sampling (RS) and Random Grouping (RG) Methods ($n = 9k = 18$)[a]

		(σ_T, σ_R)			
θ_{BE}	Method	(0.05, 0.05)	(0.05, 0.10)	(0.10, 0.05)	(0.10, 0.10)
Median	RS	0.0494	0.0460	0.0538	0.0508
	RG	0.0590	0.0528	0.0648	0.0646
1.0	RS	0.1748	0.3760	0.0132	0.1248
	RG	0.1772	0.3804	0.0308	0.1448
1.5	RS	0.3596	0.6340	0.0752	0.3646
	RG	0.3504	0.6436	0.0906	0.3606
2.0	RS	0.4842	0.7942	0.1560	0.5556
	RG	0.4790	0.7672	0.1918	0.5434

[a]Number of simulations = 5000.

DISCUSSION

To apply our proposed tests, an appropriate percentile of the ratio rd should be chosen as the parameter Rd in Eq. (5) and an appropriate bioequivalence limit θ_{BE} should be specified. These decisions should be made by a regulatory agency (e.g., the FDA) through some empirical studies using historical data. Note that in Ref. [1], a bioequivalence limit θ_{BE} is not provided, due to the fact that using the mean of rd as Rd is inappropriate since the mean of rd is not well defined in many situations.

It is well known that when n is large, a binomial distribution can be approximated by a normal distribution. When Rd is chosen as the median of rd, for example, a test with approximate level 5% rejects H_0 in Eq. (5) if and only if

$$\frac{1}{2} < \frac{Y}{n} - \frac{1.645}{2\sqrt{n}},$$

which is equivalent to

$$Y > \frac{n}{2} + \frac{1.645\sqrt{n}}{2}. \tag{9}$$

When $n = 18$, the right-hand side of Eq. (9) is equal to 12.49. Since Y is integer-valued, this provides the same test given in "The Proposed Tests" section. Hence, we conclude that when Rd is the median of rd, approximation (9) can be used when $n > 18$.

To address lot-to-lot variation, the FDA requires three lots from each of the test and reference products be included in the bioequivalence study. When lot-to-lot variation is large, however, three lots from each of the test and reference products may not be enough. Our proposed tests can be easily extended to the situation where m lots from each of the test and reference products are included in the study. In fact, if there is are large number of lots available, then our procedure can be modified as follows. In the first stage sampling, instead of sampling from the nine overlapped lot-triplets given by Eq. (6), we take n independent (nonoverlapped) lot-triplets from the available lot population. The rest of the procedure remains the same. Since lot-triplets are sampled from the available lot population, the result obtained in the statistical analysis is then applicable to all future lots.

Our test procedures do not require any distributional assumption on the profile data. However, the use of the ratio rd given by Eqs. (1)–(3) is based on the ideas in Ref. [1], i.e., profile vectors are compound multinomial. An example of a compound multinomial distribution is provided in our simulation studies.

REFERENCES

1. FDA, *Guidance for Industry on Bioavailability and Bioequivalence Studies for Nasal Aerosols and Nasal Sprays for Local Action*; Center for Drug Evaluation and Research, Food and Drug Administration: Rockville, Maryland, 1999.

2. FDA, *Guidance for Industry on Bioavailability and Bioequivalence Studies for Orally Administered Drug Products—General Consideration*; Center for Drug Evaluation and Research, Food and Drug Administration: Rockville, Maryland, 2000.
3. FDA, *Guidance for Industry on Statistical Approaches to Establishing Bioequivalence*; Center for Drug Evaluation and Research, Food and Drug Administration: Rockville, Maryland, 2001.
4. Chow, S.C.; Shao, J.; Wang, H. In Vitro Bioequivalence Testing. Stat. Med. **2002**, in press.
5. Efron, B. Bootstrap Methods: Another Look at the Jackknife. Ann. Stat. **1979**, *7*, 1–26.