Intersection-Union Tests in Dissolution Profile Testing

Jorge H. Sierra-Cavazos
Departamento de Matemáticas
ITESM Campus Monterrey
Ave. Eugenio Garza Sada 2501 Sur
Monterrey, N. L., México, CP 64849

Roger L. Berger
Department of Statistics
North Carolina State University
Raleigh, NC 27695-8203

November, 1999
NCSU Institute of Statistics Mimeo Series No. 2521

Abstract

For $k = 1, \ldots, p$, let t_k denote the time points at which dissolution profiles are collected, and let X_{ik} and Y_{jk} represent the proportion of drug dissolved from dosage unit i for the reference drug and for dosage unit j for the test drug respectively, $i = 1, \ldots, n, j = 1, \ldots, m$. We assume that the samples are drawn from independent multivariate p normal populations with mean vectors $\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_Y$ respectively and the same variance covariance matrix $\boldsymbol{\Sigma}$. We consider the problem of testing $\mathrm{H}_0 : \frac{\mu_{Xk}}{\mu_{Yk}} \leq \frac{1}{\delta}$ or $\frac{\mu_{Xk}}{\mu_{Yk}} \geq \delta$ for some $k = 1, \ldots, p$, versus $\mathrm{H}_1 : \frac{1}{\delta} < \frac{\mu_{Xk}}{\mu_{Yk}} < \delta$, for all $k = 1, \ldots, p$, where δ is a fixed constant. An easy to implement size α intersection-union test (IUT) is presented for this problem and power comparisons are made by means of simulations between the procedures recommended by the FDA and the IUT given. A method of sample size calculation for the IUT is presented.

Keywords: Intersection-union test, dissolution profile testing, normal population, t distribution, power, sample size.

1 Introduction

The dissolution profile of a drug is constructed by plotting the proportion of tablet dissolved against time in an *in vitro* experiment. A comparison of dissolution profiles between a test drug and a reference drug is one of the procedures required by the United States Food and Drug Administration (FDA) to evaluate the equivalence of the two drugs. The procedures used to compare dissolution profiles are classified in two types, model dependent and model independent. For the model dependent methods, a specific dissolution curve is fitted to the data and the comparison is made by comparing the parameters of the two fitted curves. The model independent methods compare the profiles only at given time points. We will discuss only model independent methods here.

Let t_k, $k = 1, \ldots, p$, denote the time points at which dissolution profiles are being measured and denote by X_{ik} the proportion of tablet i of the reference drug dissolved at time t_k, for $i = 1, \ldots, n$. Similarly, Y_{jk} denotes the proportion of tablet j from the test drug dissolved at the same time point t_k for $j = 1, \ldots, m$. Denote by $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$ and $\boldsymbol{Y}_j = (Y_{j1}, Y_{j2}, \ldots, Y_{jp})'$ the observation vectors for tablet i of the reference and tablet j of the test drugs over time. Some of the model independent methods that are used to compare dissolution profiles are:

1. The f_2 measure of similarity recommended by the FDA.
2. Hotelling's T^2.
3. Linear Mixed Effects models.

O'Hara et al. (1997) compared these three methods by simulations. However, we believe that special care needs to be taken when comparing these methods, since they are addressing different questions. This will be explained in more detail in what follows.

The f_2 method was proposed by Moore and Flanner (1996) and consists of computing the number

$$f_2 = 50 \times \log_{10}\left(\left[1 + \frac{1}{p}\sum_{k=1}^{p} \omega_k(\overline{X}'_k - \overline{Y}'_k)^2\right]^{-\frac{1}{2}} \times 100\right),$$

where ω_k is an optional weight, $\overline{X}'_k$ is the mean percentage (rather than proportion) dissolved of the reference drug and $\overline{Y}'_k$ the mean percentage dissolved of the test drug at the same time point t_k. If the computed number f_2 lies between 50 and 100, then the profiles are declared to be equivalent. After some algebra, the inequality

$$50 \leq f_2 \leq 100$$

is shown to be equivalent to

$$\frac{1}{p}\sum_{k=1}^{p} \omega_k(\overline{X}'_k - \overline{Y}'_k)^2 \leq 99.$$

In other words, the f_2 measure of similarity is just a weighted distance squared between two multivariate means, and when $\omega_k = 1$ for all k, it corresponds to the Euclidean distance squared. In this procedure no assumptions are made about the distribution of the data. The FDA recommends the use of this measure for testing dissolution profile equivalence; see, for example, FDA (1995, 1997a, 1997b).

The last two methods, Hotelling's T^2 and linear mixed effects models, test the statistical hypotheses

$$\begin{array}{c} \mathrm{H}_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y \\ \text{versus} \\ \mathrm{H}_1 : \boldsymbol{\mu}_X \neq \boldsymbol{\mu}_Y, \end{array} \tag{1}$$

where $\boldsymbol{\mu}_X = (\mu_{X1}, \mu_{X2}, \ldots, \mu_{Xp})'$ is the $p \times 1$ mean vector for the population from which the $\boldsymbol{X}_i$'s are sampled. Similarly for $\boldsymbol{\mu}_Y$. The hypotheses in (1) are equivalent to

$$\begin{array}{ll} \mathrm{H}_0 : \frac{\mu_{Xk}}{\mu_{Yk}} = 1 & \text{for all } k = 1, \ldots, p, \\ \text{versus} & \\ \mathrm{H}_1 : \frac{\mu_{Xk}}{\mu_{Yk}} \neq 1 & \text{for some } k = 1, \ldots, p, \end{array} \tag{2}$$

when none of the means μ_{Yk} is zero.

The two statistical methods mentioned, although widely used for assessing equivalence of dissolution profiles, do not really answer the question: Can the two dissolution profiles be considered equivalent? The document Guidance for Industry, SUPAC-MR: Modified Release Solid Oral Dosage Forms (1997) states the following for equivalence of dissolution profiles:

> "...the average difference at any dissolution sampling time point should not be greater than 15 percent between the changed drug product and the biobatch or marketed batch (unchanged drug product) dissolution profiles."

In other words, two dissolution profiles are considered equivalent if, at every given time point, the difference between the two profiles is not greater than 15%. Obviously, if the two profiles coincide at any time point, the difference between them is less than 15%. However, by testing the hypotheses (2), the null hypothesis can be rejected even for profiles that differ by less than 15% at every time point, thus concluding that the profiles are not equivalent when they actually are. Schuirmann (1987) presents a very good discussion on this issue when the difference between the means is fixed and not a percentage as in our case. Although his problem is not exactly the same as ours, the rationale is similar for our case.

The f_2 method was developed to assess the difference between the two profiles and decide if at every time point they differ by less than 15% or not, whereas Hotelling's T^2 and mixed effects models are methods designed to conclude if the profiles coincide or not.

We present here a different way of setting up the hypotheses so that the question of profile equivalence is directly addressed, and give an easy to implement size α intersection-union test (IUT) for the new set of hypotheses. A comparison of power is made for the IUT and the f_2 measure by means of simulation in section 4. Sample size considerations are given in section 5 and some final comments are given in section 6.

2 Notation and Assumptions

For the same notation presented in the previous section, let $\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i = (\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_p)'$ be the average of the observed values for the vectors $\boldsymbol{X}_i$ and $\boldsymbol{S}_X = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})'$ be the observed variance-covariance matrix. $\overline{\boldsymbol{Y}}$ and $\boldsymbol{S}_Y$ are defined similarly. We assume that $\boldsymbol{X}_i$ for $i = 1, \ldots, n$ constitute a random sample from a $N_p(\boldsymbol{\mu}_X, \boldsymbol{\Sigma})$ population and that $\boldsymbol{Y}_j$ for $j = 1, \ldots, m$ is a random sample from a $N_p(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma})$ population, independent from the $\boldsymbol{X}_i$'s. Note that we are assuming the same variance-covariance structure for both populations, however, this model allows unequal variances for different time points as well as considering the values measured on a single tablet over time to be correlated. Let $\boldsymbol{A}$ denote the pooled variance-covariance matrix of the samples, that is

$$\boldsymbol{A} = \frac{(n-1)\boldsymbol{S}_X + (m-1)\boldsymbol{S}_Y}{n+m-2}.$$

Denote by a_{ij} the $(i,j)^{th}$ element of $\boldsymbol{A}$. It is important to note that a_{kk} corresponds to the pooled variance of the samples at time point t_k.

3 The Hypotheses and an IUT

As was mentioned before, two dissolution profiles are considered equivalent if the percentage difference between them, at any given time point, is at most 15%. In other words, they are considered equivalent if $\frac{1}{1.15} < \frac{\mu_{Xk}}{\mu_{Yk}} < 1.15$ for all $k = 1, \ldots, p$. In a bioequivalence setting, it is well understood that the proper alternative hypothesis is that the ratio is within the specified bounds; see, for example, Anderson and Hauck (1983, 1984). The hypotheses for the dissolution profile testing problem should be stated in a similar way as those for bioequivalence problems to properly guard against declaring the two profiles equivalent when they are not. Therefore, we consider the problem of testing

$$
\begin{array}{ll}
& \mathrm{H}_0 : \frac{\mu_{Xk}}{\mu_{Yk}} \leq \frac{1}{\delta} \quad \text{or} \quad \frac{\mu_{Xk}}{\mu_{Yk}} \geq \delta \quad \text{for some } k = 1, \ldots, p, \\
\text{versus} & \\
& \mathrm{H}_1 : \frac{1}{\delta} < \frac{\mu_{Xk}}{\mu_{Yk}} < \delta, \qquad\qquad \text{for all } k = 1, \ldots, p,
\end{array}
\tag{3}
$$

where $\delta > 1$ is a fixed constant. For the problem of dissolution profiles testing, $\delta = 1.15$. The hypotheses (3) can be expressed as

$$
\begin{array}{ll}
& \mathrm{H}_0 : \cup_{k=1}^{p} \left[\left(\frac{\mu_{Xk}}{\mu_{Yk}} \leq \frac{1}{\delta} \right) \cup \left(\frac{\mu_{Xk}}{\mu_{Yk}} \geq \delta \right) \right] \\
\text{versus} & \\
& \mathrm{H}_1 : \cap_{k=1}^{p} \left[\left(\frac{\mu_{Xk}}{\mu_{Yk}} > \frac{1}{\delta} \right) \cap \left(\frac{\mu_{Xk}}{\mu_{Yk}} < \delta \right) \right].
\end{array}
\tag{4}
$$

The parameter μ_{Yk} is positive for all $k = 1, \ldots, p$ because it represents the average proportion of tablet dissolved at time t_k. Therefore the hypotheses (4) can be restated as

$$
\begin{array}{ll}
& \mathrm{H}_0 : \cup_{k=1}^{p} \left[(\delta\mu_{Xk} - \mu_{Yk} \leq 0) \cup (\delta\mu_{Yk} - \mu_{Xk} \leq 0) \right] \\
\text{versus} & \\
& \mathrm{H}_1 : \cap_{k=1}^{p} \left[(\delta\mu_{Xk} - \mu_{Yk} > 0) \cap (\delta\mu_{Yk} - \mu_{Xk} > 0) \right].
\end{array}
\tag{5}
$$

A test statistic for testing hypotheses like (5) can be developed by using IUT's. Let us consider the following hypotheses that refer to a single time point,

$$
\begin{array}{ll}
& \mathrm{H}_{0k} : \delta\mu_{Xk} - \mu_{Yk} \leq 0 \quad \text{or} \quad \delta\mu_{Yk} - \mu_{Xk} \leq 0 \\
\text{versus} & \\
& \mathrm{H}_{1k} : \delta\mu_{Xk} - \mu_{Yk} > 0 \quad \text{and} \quad \delta\mu_{Yk} - \mu_{Xk} > 0.
\end{array}
\tag{6}
$$

The likelihood ratio test for (6) was developed by Sasabuchi (1980, 1988a, 1988b). He proved that the size α LRT for (6) rejects H_{0k} if and only if both T_{1k} and T_{2k} are greater than or equal to $t_{\alpha,r}$ where

$$
T_{1k} = \frac{\delta\overline{X}_k - \overline{Y}_k}{\sqrt{a_{kk}\left(\frac{\delta^2}{n} + \frac{1}{m}\right)}} \quad , \qquad T_{2k} = \frac{\delta\overline{Y}_k - \overline{X}_k}{\sqrt{a_{kk}\left(\frac{1}{n} + \frac{\delta^2}{m}\right)}}
$$

and $r = n + m - 2$. We call this test the T_{1k}/T_{2k} test. Berger and Hsu (1996) recognized this test also as a size α IUT for (6). Now, by combining the T_{1k}/T_{2k} tests for all k we obtain a level α IUT for (5) or, equivalently, for (3). Therefore the test that rejects H_0 of

(3) if and only if T_{1k} and T_{2k} are greater than or equal to $t_{\alpha,r}$ for all $k = 1, \ldots, p$ is a level α test for (3). To prove that this is actually a size α test, consider the parameter point $\boldsymbol{\mu}_X^* = (\mu_{X1}, \mu_{X2}, \ldots, \mu_{Xp})'$, $\boldsymbol{\mu}_Y^* = (\delta\mu_{X1}, \mu_{X2}, \ldots, \mu_{Xp})'$, that is, a parameter point on the boundary of H_{01} but inside H_{1k} for $k = 2, \ldots, p$ and let $\sigma_{kk} \to 0$ for all $k = 1, 2, \ldots, p$. Now, applying Theorem 2 in Berger and Hsu (1996) or in Berger (1982), we obtain that this is a size α test for (5) or equivalently for (3). A simpler way to write this test procedure is the following: reject H_0 if and only if

$$\min_{1 \le k \le p} \{T_{1k}, T_{2k}\} \ge t_{\alpha,r}. \tag{7}$$

This is in the form of the "min test" of Laska and Meisner (1989). In the next section we compare the power of this test against the f_2 measure of similarity recommended by the FDA.

4 Power Comparison: IUT versus f_2

In this section we present a simulation study on the power of the IUT and the f_2 measure of similarity. For this study, we use basically the same set up as the one presented by O'Hara et al. (1997), with some variations mainly in the variance structure of the data. The time points considered are 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 10, 12 and 24 hours and a value of $\boldsymbol{\mu}_X = (0.087, 0.171, 0.332, 0.494, 0.609, 0.706, 0.783, 0.835, 0.875, 0.925, 0.952, 0.999)'$ is used for the reference drug. The values of $\boldsymbol{\mu}_Y$ are computed by using the formula

$$\mu_{Yk} = 1 - (1 - \mu_{Xk})^r$$

for each of 57 values of r equally spaced on a logarithmic scale in the range between 0.497 and 2.014. By generating the means in this manner, the ratio $\frac{\mu_{Xk}}{\mu_{Yk}}$ lies between $\frac{1}{\delta}$ and δ, for all $k = 1, 2, \ldots, p$ with $\delta = 1.15$, only for values of $\log(r)$ in the range from -0.145851 to 0.146842. For values of $\log(r)$ outside this range, at least one $\frac{\mu_{Xk}}{\mu_{Yk}}$ is not within the limits given. Therefore, we would expect that a good method of dissolution profile testing should not reject, or reject with a small probability, the null hypothesis of (3) for values of $\log(r)$ outside this interval, and reject with a high probability for values inside this range.

O'Hara et al. (1997) used, for generating the reference and test data, a random 'between tablet' term which was common to all observations on the same tablet and a 'within tablet' error term which was unique to each observation. Both terms had constant standard deviation for their simulation. By doing this, they are giving the data the same variance at each time point. We believe that a more reasonable variance structure would be to assume smaller variance for early and late time points and bigger variance for time points in the middle. Moreover, the document Guidance for Industry: Dissolution Testing for Immediate Release Solid Oral Dosage Forms states that to allow the use of procedures that use mean data for profile comparison, the percent coefficient of variation at early time points should not be more than 20%, and at other time points should not be more than 10% (the coefficient of variation is defined as $\frac{st.dev.}{mean}$).

O'Hara et al. (1997) did two simulations. In the first simulation they used 0.02 for the standard deviation of the between tablet term and 0.015 for the standard deviation of the within tablet term. In the second simulation they used 0.01 and 0.0075 for the corresponding standard deviations.

For the mean values used in their simulation, only the second set of standard deviations satisfy that the coefficient of variation for the first time point is less than 20% and less than 10% for the remaining.

Therefore, we use different standard deviations for generating the values for the 'within' tablet term. More specifically, for generating this term we use the values 0.0075, .0085, .0095, .0125, .0140, .0150, .0150, .0140, .0125, .0095, .0085 and .0075 for the standard deviations at the corresponding time points. We keep the standard error of the 'between tablet' term fixed at the value 0.01. For computing the f_2 measure, we use $\omega_k = 1$ for all k. The test is performed with a value of $\alpha = 0.05$.

As was mentioned before, the FDA recommends the use of the f_2 measure and requires also that the sample means differ by no more than 15% ate each time point (FDA 1997b). We computed the proportion of times that the f_2 measure was between 50 and 100 out of 1000 repetitions. We also computed the proportion of times that the condition $\frac{1}{\delta} \le \frac{\overline{X}_k}{\overline{Y}_k} \le \delta$ for all $k = 1, 2, \ldots, p$ was satisfied with $\delta = 1.15$, we call this the Conditioning test (CT). The power of the IUT was also computed as a proportion of times that the null hypothesis of (3) was rejected. The sample size for the reference and test drugs was taken to be 12 for the first simulation. The results are presented in Figure 1.

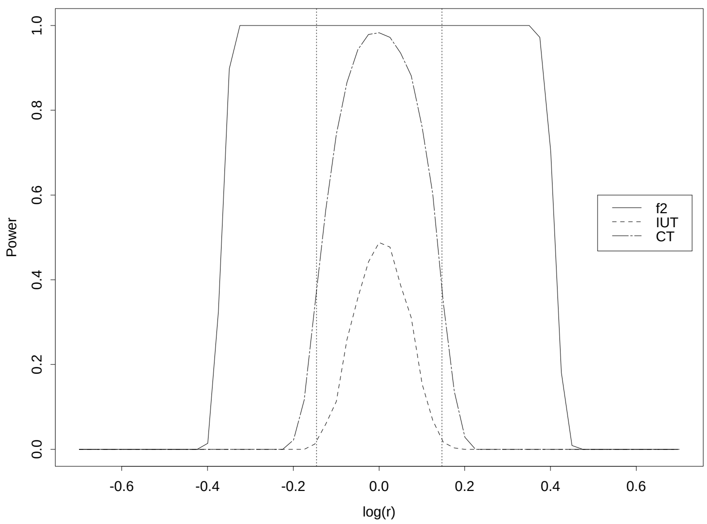

Figure 1: Power comparison between the f_2, CT and IUT, $n = m = 12$.

The two vertical lines in the graphic mark the border between the null and alternate hypotheses. At the boundary of H_0, the value of the power of the IUT is very close to the nominal $\alpha = 0.05$. We observe that the power for the IUT is not very high for $\log(r) = 0$, the value at which the profiles coincide. This is due to the sample size used as well as the value for the coefficient of variation at each time point. This will be discussed in the next section. Note that both, f_2 and CT are more powerful than the IUT inside the range

-.145851, .146842 but they also reject the null hypothesis with high probability for values outside this range, and as a result of this, their levels as statistical hypothesis tests are much bigger than α. Actually f_2 has level 1 and CT has an approximate level of .4 in this case. This means that, by using f_2 or CT as a testing procedure for dissolution profile testing, the probability of declaring two profiles equivalent when they are not is high.

Two more studies were performed with sample sizes 20 and 45, the results are presented in Figures 2 and 3 respectively. We observe that the power of the IUT increases as the sample size increases. We also note that in these cases, both f_2 and CT are too liberal whereas IUT maintains the level of $\alpha = 0.05$. In fact, it can be shown that, for large samples, CT will have a rejection probability of .5 on the border between H_0 and H_1. This means that CT will be liberal for any sample size.

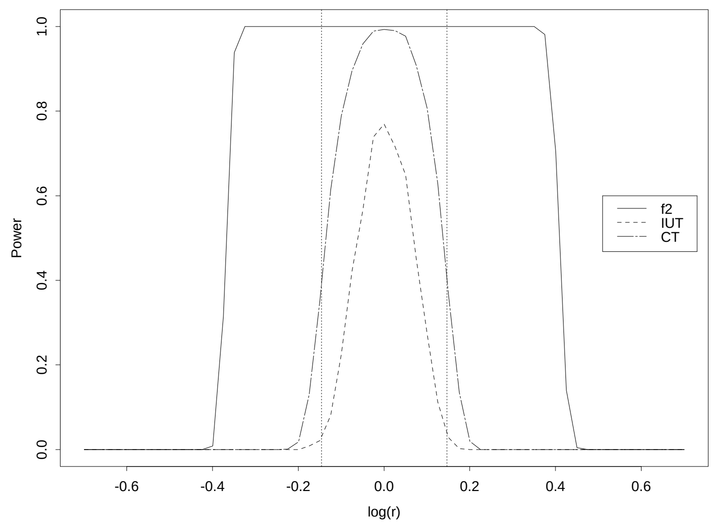

Figure 2: Power comparison between the f_2, CT and IUT, $n = m = 20$.

5 Sample Size Considerations

The standard sample size for dissolution profile testing required by the FDA is 12. However, and depending on the variance structure of the data, this sample size may not be large enough to be able to detect profile equivalence, even when the profiles coincide at each time point. In this section we obtain an approximate sample size that ensures detecting profile equivalence, when the profiles are actually the same at the sampled time points, with a high probability, say $1-\beta$. In this section we will show that a good approximation of the sample size is given by

$$n = \frac{(z_\alpha + z_{\beta/2})^2 (CV_k^*)^2 (1+\delta^2)}{(\delta-1)^2}, \tag{8}$$

where CV_k^* is the largest coefficient of variation among all time points.

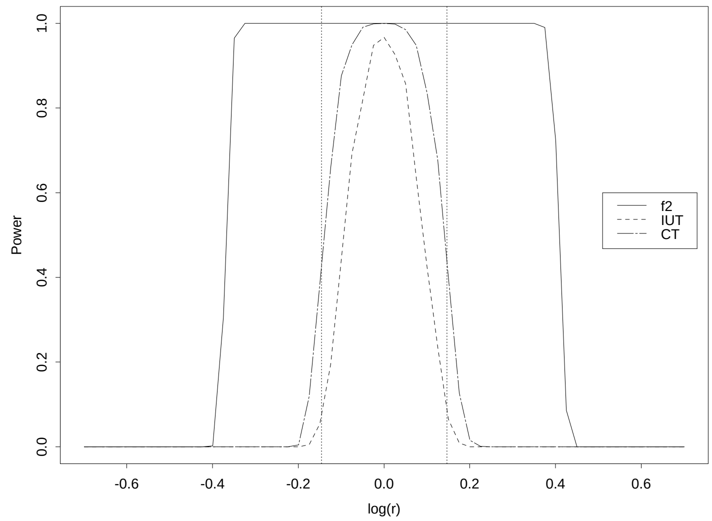

Figure 3: Power comparison between the f_2, CT and IUT, $n = m = 45$.

We would like to be able to reject H_0 with a high probability whenever $\boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$, i.e., whenever the profiles coincide. This can be written as

$$P(\text{ reject H}_0 \quad | \quad \mu_{Xk} = \mu_{Yk} \quad \text{for all} \quad k = 1, 2, \ldots, p) \geq 1 - \beta \tag{9}$$

for a small $0 < \beta < 1$.

Now, according to our previous discussion, our IUT will reject H_0 if and only if

$$\frac{\delta \overline{X}_k - \overline{Y}_k}{\sqrt{a_{kk}\left(\dfrac{\delta^2}{n} + \dfrac{1}{m}\right)}} \geq t_{\alpha,r} \quad \text{and} \qquad \frac{\delta \overline{Y}_k - \overline{X}_k}{\sqrt{a_{kk}\left(\dfrac{1}{n} + \dfrac{\delta^2}{m}\right)}} \geq t_{\alpha,r} \quad \text{for all} \quad k = 1, \ldots, p.$$

Assuming $\boldsymbol{\Sigma}$ known, the rejection region would be still determined by the same inequalities but replacing a_{kk} by σ_{kk} and $t_{\alpha,r}$ by z_α. In fact, Sasabuchi (1980) showed that in this case, this is actually the size α likelihood ratio test for testing (3). Now, the probability of rejecting H_0 when $\boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ is

$$P_{\boldsymbol{\mu}_X = \boldsymbol{\mu}_Y}(\text{ reject H}_0) = P_{\boldsymbol{\mu}_X = \boldsymbol{\mu}_Y}\left(\frac{\delta \overline{X}_k - \overline{Y}_k}{\sqrt{\sigma_{kk}\left(\dfrac{\delta^2}{n} + \dfrac{1}{m}\right)}} \geq z_\alpha, \frac{\delta \overline{Y}_k - \overline{X}_k}{\sqrt{\sigma_{kk}\left(\dfrac{1}{n} + \dfrac{\delta^2}{m}\right)}} \geq z_\alpha \quad \text{for all} \quad k\right). \tag{10}$$

The inequalities in the right hand side of (10) can be written as

$$\frac{\delta \overline{X}_k - \overline{Y}_k - (\delta \mu_{Xk} - \mu_{Yk})}{\sqrt{\sigma_{kk}\left(\frac{\delta^2}{n} + \frac{1}{m}\right)}} \geq z_\alpha - \frac{\delta \mu_{Xk} - \mu_{Yk}}{\sqrt{\sigma_{kk}\left(\frac{\delta^2}{n} + \frac{1}{m}\right)}}, \qquad \frac{\delta \overline{Y}_k - \overline{X}_k - (\delta \mu_{Yk} - \mu_{Xk})}{\sqrt{\sigma_{kk}\left(\frac{1}{n} + \frac{\delta^2}{m}\right)}} \geq z_\alpha - \frac{\delta \mu_{Yk} - \mu_{Xk}}{\sqrt{\sigma_{kk}\left(\frac{1}{n} + \frac{\delta^2}{m}\right)}},$$

and when $\boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ and the sample sizes are the same for both, the reference and the test drugs, i.e., $n = m$, we have

$$\frac{\delta\overline{X}_k - \overline{Y}_k - \mu_k(\delta-1)}{\sqrt{\frac{\sigma_{kk}(\delta^2+1)}{n}}} \geq z_\alpha - \frac{\mu_k(\delta-1)}{\sqrt{\frac{\sigma_{kk}(\delta^2+1)}{n}}}, \qquad \frac{\delta\overline{Y}_k - \overline{X}_k - \mu_k(\delta-1)}{\sqrt{\frac{\sigma_{kk}(1+\delta^2)}{n}}} \geq z_\alpha - \frac{\mu_k(\delta-1)}{\sqrt{\frac{\sigma_{kk}(1+\delta^2)}{n}}}.$$

Let $Z_{1k} = \frac{\delta\overline{X}_k - \overline{Y}_k - \mu_k(\delta-1)}{\sqrt{\frac{\sigma_{kk}(1+\delta^2)}{n}}}$ and $Z_{2k} = \frac{\delta\overline{Y}_k - \overline{X}_k - \mu_k(\delta-1)}{\sqrt{\frac{\sigma_{kk}(1+\delta^2)}{n}}}$, then each Z_{1k} and each Z_{2k} are standard normals, although not independent. Actually, the vector $\boldsymbol{Z} = (Z_{11}, Z_{12}, \ldots, Z_{1p}, Z_{21}, Z_{22}, \ldots, Z_{2p})'$ has a $N_{2p}(\mathbf{0}, \boldsymbol{P})$ distribution, where

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{\rho} & \frac{-2\delta}{1+\delta^2}\boldsymbol{\rho} \\ \frac{-2\delta}{1+\delta^2}\boldsymbol{\rho} & \boldsymbol{\rho} \end{pmatrix}$$

and $\boldsymbol{\rho}$ is the common correlation matrix of the $\boldsymbol{X}_i$'s and the $\boldsymbol{Y}_j$'s. Therefore the probability of rejecting when the means are equal can be expressed as

$$P\left(Z_{1k} \geq z_\alpha - \frac{\mu_k(\delta-1)}{\sqrt{\frac{\sigma_{kk}(\delta^2+1)}{n}}}, \quad Z_{2k} \geq z_\alpha - \frac{\mu_k(\delta-1)}{\sqrt{\frac{\sigma_{kk}(\delta^2+1)}{n}}} \quad \text{for all } k \right).$$

We note that this probability depends only on δ, n and the coefficient of variation at each time point ($CV_k = \frac{\sqrt{\sigma_{kk}}}{\mu_k}$). We also observe that this probability is decreasing with respect to each CV_k.

We restrict our attention to only one pair of variables, Z_{1k} and Z_{2k} for a fixed k, to find a lower bound on the sample size in order to get a probability of at least $1-\beta$ of rejection when the profiles coincide. The random vector $(Z_{1k}, Z_{2k})'$ has a bivariate normal distribution with mean $\mathbf{0}$ and variance covariance matrix

$$\begin{pmatrix} 1 & \frac{-2\delta}{1+\delta^2} \\ \frac{-2\delta}{1+\delta^2} & 1 \end{pmatrix}.$$

For $\delta = 1.15$ the quantity $\frac{-2\delta}{1+\delta^2}$ is less than -0.99 and hence, for $b < 0$, $P(Z_{1k} \geq b, Z_{2k} \geq b) \approx 1 - 2P(Z < b)$ for $Z \sim N(0,1)$. Therefore we could use $-z_{\beta/2}$ instead of b to get $P(Z_{1k} \geq b, Z_{2k} \geq b) \approx 1 - \beta$.

Table 1 contains the computed values of $P(Z_{1k} \geq b, Z_{2k} \geq b)$ for different values of b as well as $1 - 2P(Z < b)$ for $\delta = 1.15$. Note that the approximation is correct to four decimal places if $b \leq -0.3$. The value of b will typically be chosen to make the probability (power) .8 or .9. So the approximation is certainly adequate in this case. Actually, from Bonferroni's inequality,

$$P(Z_{1k} \geq b, Z_{2k} \geq b) \geq 1 - 2P(Z < b)$$

so, using $b = -z_{\beta/2}$ is slightly conservative.

Therefore, we have

$$z_\alpha - \frac{(\delta-1)\sqrt{n}}{CV_k\sqrt{1+\delta^2}} = -z_{\beta/2}$$

Table 1: Probability values for different cutting points b, and $\delta = 1.15$.

b	$P(Z_{1k} \geq b, Z_{2k} \geq b)$	$1 - 2P(Z < b)$
0.0	.0222	.0000
-0.1	.0815	.0797
-0.2	.1586	.1585
-0.3	.2358	.2358
-0.4	.3108	.3108
-0.5	.3829	.3829
-0.6	.4515	.4515
-0.7	.5161	.5161
-0.8	.5763	.5763
-0.9	.6319	.6319
-1.0	.6827	.6827
-1.1	.7287	.7287
-1.2	.7699	.7699
-1.3	.8064	.8064
-1.4	.8385	.8385
-1.5	.8664	.8664
-1.6	.8904	.8904
-1.7	.9109	.9109
-1.8	.9281	.9281
-1.9	.9426	.9426
-2.0	.9545	.9545

and solving for n we obtain

$$n = \frac{(z_\alpha + z_{\beta/2})^2 (CV_k)^2 (1 + \delta^2)}{(\delta - 1)^2}.$$

Note that the value of n is an increasing function of CV_k, and therefore by using the largest value of CV_k we obtain the value of n that will make the probabilities $P(Z_{1k} \geq b, Z_{2k} \geq b) \geq 1 - \beta$ for all $k = 1, 2, \ldots, p$, this leads to formula (8). The value of n that we obtain by formula (8) is, however, only a lower bound for the value required to achieve the probability given in (9).

Now, for $\delta = 1.15, z_\alpha = 1.645$, $z_{\beta/2} = 1.645$ (corresponding to a power of .9) and $CV_k^* = .2$ in (8), we obtain $n \approx 45$.

For $n = 12$, $\delta = 1.15$, $z_\alpha = 1.645$ and $CV_k^* = .1436$ which is the largest coefficient of variation in our simulation, we obtain $z_{\beta/2} = .728$ and then the probability of rejection when the profiles are the same is approximately .5346. This value is close to what we got in our simulation (Figure 1).

With the same values given and $n = 20$ we have that the power should be close to .8444 and for $n = 45$ the power should be approximately .9968. These values are close to the values for the power that we observed in Figures 2 and 3 for $\log(r) = 0$ in our simulation studies.

6 Conclusions

The f_2 and CT methods are procedures recommended by the FDA to test the equivalence of two dissolution profiles, however, the probability of declaring that two profiles are equivalent when they are not with the CT method is high and this problem is even worse with the f_2 method.

We could still use Hotelling's T^2 or mixed effects models, but these methods will declare that the profiles are not equivalent with high probability even when the profiles differ by no more than 15% at each time point.

The IUT introduced here controls both aspects. By setting a level for the test, the probability of declaring two profiles equivalent when they are not is at most α, and by establishing the hypotheses as (3) and using an appropriate sample size, the probability of declaring profile equivalence correctly can be as big as desired ($n = 45$ will give a probability of about .90 when $CV_k^* \leq .2$).

References

Anderson, S. and W. W. Hauck (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications is Statistics: Theory and Methods 12*, 2663–2692.

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics 24*, 295–300.

Berger, R. L. (1989). Uniformly more powerful tests for hypotheses concerning linear inequalities and normal means. *Journal of the American Statistical Association 84*, 192–199.

Berger, R. L. and D. D. Boos (1999). Confidence limits for the onset and duration of treatment effect. *Biometrical Journal 41*, 517–531.

Berger, R. L. and J. C. Hsu (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science 11*, 283–319.

Casella, G. and R. L. Berger (1990). *Statistical Inference*. Pacific Grove, California: Brooks/Cole Publishing Company.

FDA (1995). *Guidance for Industry, Immediate Release Solid Oral Dosage Forms.* U.S. Department of Health and Human Services, FDA, CDER, November 1995.

FDA (1997a). *Guidance for Industry: Dissolution Testing of Immediate Release Solid Oral Dossage Forms.* U.S. Department of Health and Human Services, FDA, CDER, August 1997.

FDA (1997b). *Guidance for Industry, SUPAC-MR: Modified Release Solid Oral Dosage Forms.* U.S. Department of Health and Human Services, FDA, CDER, September 1997.

Hauck, W. W. and S. Anderson (1984). A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *Journal of Pharmacokinetics and Biopharmaceutics 12*, 83–91.

Laska, E. M. and M. J. Meisner (1989). Testing whether an identified treatment is best. *Biometrics 45*, 1139–1151.

Moore, J. W. and H. H. Flanner (1996). Mathematical comparison of dissolution profiles. *Pharmaceutical Technology*, June 1996, pp 64–74.

O'Hara, T., A. Dunne, A. Kinahan, S. Cunningham, P. Strak, and J. Devane (1997). Review of methodologies for the comparison of dissolution profile data. In *In Vitro-In Vivo Correlations*, pp. 167–171. Plenum Press, New York. Edited by Young et al.

Rescigno, A. (1992). Bioequivalence. *Pharmaceutical Research 9*, 925–928.

Sasabuchi, S. (1980). A test for a multivariate normal mean with compsite hypotheses determined by linear inequalities. *Biometrika 67*, 429–439.

Sasabuchi, S. (1988a). A multivariate one-sided test with composite hypotheses when the covariance matrix is completely unknown. *Memoirs of the Faculty of Science, Kyushu University, Ser. A 42*, 37–46.

Sasabuchi, S. (1988b). A multivariate test with composite hypotheses determined by linear inequalities when the covariance matrix has an unknown scale factor. *Memoirs of the Faculty of Science, Kyushu University, Ser. A 42*, 9–19.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailaty. *Journal of Pharmacokinetics and Biopharmaceutics 15*, 657–680.

Tsong, Y. and T. Hammerstrom (1994). Statistical issues in drug quality control based on dissolution testing. In *Proceedings of the Biopharmaceutical Section of the American Statistical Association*, pp. 295–300.