2-6-2007

A SURVEY OF THE LIKELIHOOD APPROACH TO BIOEQUIVALENCE TRIALS

Leena Choi
Department of Biostatistics, School of Medicine, Vanderbilt University

Brian S. Caffo
The Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, bcaffo@jhsph.edu

Charles Rohde
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

A Survey of the Likelihood Approach to Bioequivalence Trials

Leena Choi[*], Brian Caffo and Charles Rohde[†]

February 6, 2007

Abstract

Bioequivalence trials are abbreviated clinical trials whereby a generic drug or new formulation is evaluated to determine if it is "equivalent" to a corresponding previously approved brand-name drug or formulation. In this manuscript, we survey the process of testing bioequivalence and advocate the likelihood paradigm for representing the resulting data as evidence. We emphasize the unique conflicts between hypothesis testing and confidence intervals in this area - which we believe are indicative of the existence of the systemic defects in the frequentist approach - that the likelihood paradigm avoids. We suggest the direct use of profile likelihoods for evaluating bioequivalence and examine the main properties of profile likelihoods and estimated likelihoods under simulation. This simulation study shows that profile likelihoods are a reasonable alternative to the (unknown) true likelihood for a range of parameters commensurate with bioequivalence research. Our study also shows that the standard methods in the current practice of bioequivalence trials offers only weak evidence from the evidential point of view.

keywords: likelihood principle, bioequivalence, profile likelihood, misleading evidence

[*] Department of Biostatistics, School of Medicine, Vanderbilt University, leena.choi@vanderbilt.edu

[†] Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University

1 Introduction

When pharmaceutical companies would like to market a generic drug after the patent of a brand-name drug expires or when they would like to market a new formulation of an approved drug, regulatory authorities do not require the performance of costly full scale clinical trials to demonstrate the efficacy and safety. Instead, pharmaceutical companies conduct bioequivalence (BE) trials to establish that the generic drug or new formulation ("the test") is bioequivalent to the brand-name drug or originally approved drug ("the reference").

It might seem strange, for those who are not familiar with BE trials, that a drug formulation containing the same active ingredient can show different effects or toxicities. Two formulations having different excipients, or the same excipients formulated differently, can result in different effects. Stated more succinctly, chemical equivalence of the active agent does not guarantee biological equivalence. Such problems often occur when the drugs have a narrow therapeutic index, as with digoxin (a heart medication), warfarin (a blood thinner), sustained-release theophylline formulations (an asthma medication) and phenytoin (an anticonvulsant or antiepileptic drug). For example, digoxin intoxication in 1977 received a great deal of public attention (Schulz and Steinijans, 1991) due to inadvertent toxicity attributed to generic drugs that were not bioequivalent to the brand name drug.

Balancing the need to protect patients from the failure of treatment or toxicity via rigorous evaluation methods, is the desire for safe and effective generic drugs, which are typically less expensive. As such, bioequivalence trials are of interest to many groups: pharmaceutical companies, insurance companies, prescribing doctors, pharmacists, patient-consumer groups, regulatory authorities, etcetera. Moreover, because their interests do not always coincide, discussions regarding bioequivalence statistical methodology are complex and even sometimes politically charged (see Metzler, 1974). In this manuscript, we propose the use of the likelihood as a useful first step in summarizing the data as evidence, regardless of the researcher's perspective. This likelihood paradigm represents a unified framework for quantifying evidence regarding average and population bioequivalence. Notably, the two major measures, the area under the curve (AUC)and the maximum plasma concentration reached (C_{max}) can be evaluated in the same unified framework - without concern for adjusting for multiplicity.

This manuscript outline is as follows. In Section 2.1-2.3, we review the basic concepts of BE while in Section 2.4 we examine problems in the current statistical practice in BE trials. Section 3 describes the likelihood paradigm while Section 4 illustrates how this paradigm can be applied to BE trials. Moreover, we examine important properties of profile likelihoods using simulation. A summary and discussion follows in Section 5.

2 A Review of Bioequivalence Testing

2.1 Definition and metrics of bioequivalence

The bioequivalence of a test and reference formulation depends on the closeness of characteristics of the extent and rate of absorption, generally referred to as the bioavailability of the drug. To measure bioavailability, pharmacokinetic (PK) studies are carried out. In PK studies, drug concentrations measured from blood samples obtained at pre-specified sampling times for each subject are summarized as AUC, C_{max}, and the time to reach the maximum concentration (t_{max}), all of which represent bioavailability. Comparisons between these measures are used to determine bioequivalence. Thus BE relies on the fundamental assumption that two formulations are therapeutically equivalent if their bioavailabilities are the same.

The metric AUC holds a special place amongst these summaries, being the required primary metric of the extent of absorption for most countries. The C_{max} is also an important metric, being a measure of the rate of absorption, although many researchers criticized its usage arguing that it is confounded by the amount of absorption. A number of alternative metrics have been suggested. For example, C_{max}/AUC (Endrenyi et al., 1991) or partial AUC (Chen, 1992), as a better measure of the rate of absorption, but none have been proven satisfactory (Bois et al., 1994). Sometimes t_{max} is employed as a measure of rate of absorption, although its poor temporal resolution, due to the discrete nature of the sampling times, limits its use. Despite this ongoing interest and research in other metrics, AUC and C_{max} remain the most important summaries for BE trials, and hence remain our primary focus.

2.2 Distributional assumptions for metrics in BE trials

Before performing a statistical analysis in BE trials, AUC and C_{max} are generally log transformed. The three most commonly cited reasons for using the log transformed AUC are that: $i)$ AUC is non-negative, $ii)$ the distribution of AUC is highly skewed, $iii)$ PK models are multiplicative, both theoretically and conceptually. Further discussion of the third reason is as follows.

As a conceptual rationale for the log-normal model, we note that many biological effects act multiplicatively, as well described in Limpert et al. (2001). If an outcome is the result of many random causes, each of which produces a small proportional effect, then the resulting distribution is often log-normal (Kenney and Keeping, 1951). Since the drug concentration at each time is a function of many random processes (absorption, distribution, metabolism and elimination) that reasonably would act proportionally to the amount of drug present in the body, this suggests that the resulting distribution is log-normal (Midha et al., 1993).

More theoretically, the FDA guidance (2001) provides a pharmacokinetic rationale based on Westlake (1988) which states that PK models are comprised of multiplicative components. Assuming that the elimination of the drug is first-order and only occurs from the central compartment, AUC can be expressed as follows:

$$AUC = \frac{FD}{CL} = \frac{FD}{Vk_e},$$

where F is the fraction of drug absorbed, D is the administered dose, CL is the clearance,

V is the apparent volume of distribution, and k_e is the elimination rate constant. Notice that AUC involves multiplicative terms of the PK parameters (F, V, and k_e). A log transformation of AUC results in the PK parameters entering as additively. Furthermore, if we are willing to assume that the distributions of PK parameters are log-normal, then the distribution of AUC is also log-normal.

There has been a small amount of research considering the distribution of PK parameters. A study with 54 healthy young subjects showed that some of the PK parameters of triazolam (which has relatively short half-life) were more consistent with the log-normal distribution than the normal distribution (Friedman et al., 1986). Lacey et al. (1997) investigated the distribution of PK parameters using four different compounds with 60, 69, 57 and 36 subjects, respectively. Using tests for normality (Shapiro and Wilk, 1965), they found that the majority (51%) of the distributions of PK parameters differed markedly from normality, whereas all were consistent with the log-normal distribution. In addition, they observed that log-normality is more apparent for highly variable drugs with high coefficients of variation which agrees with the discussion in Limpert et al. (2001). Finally Mizuta and Tsubotani (1985) looked at the distribution of PK parameters in samples from several drug administration routes.

Based on these results and rationale, our discussion assumes that the metric (AUC or C_{max}) is log transformed.

2.3 Design and analysis of BE trials

In a typical BE trial, the test (T) and the reference (R) formulations are administered to (12 to 30) healthy volunteers and the drug concentrations are measured over time. Frequently cross-over designs are employed, although parallel group designs are used as well. Cross-over designs are generally preferred, because of their ability to compare the test and reference formulations within a subject. As such, our discussions focus on BE trials using a 2×2 cross-over design.

Throughout we assume the critical assumption that there is no carry-over effect, or that the carry-over effect is negligible. Such carry-over effects can be due to left-over active drug in the previous period, due to psychological effects (Jones and Kenward, 2003) or other pharmacologic effects, such as induction of metabolism or elimination by the previously administered drug. However, the carry-over effect is often negligible in most BE trials (Zariffa et al., 2000; D'Angelo et al., 2001).

Design issues aside, analyses of BE trials often considers **average bioequivalence** (ABE) as a primary goal. The purpose of average bioequivalence studies is to show that the population means of the test and the reference are sufficiently close. Establishing ABE has been the only required criteria in BE trials for more than 20 years in many countries.

The current USA FDA guidelines (2001) declare the test and the reference as average bioequivalent if the difference in their population means is within the regulatory limit, say θ_A. That is

$$|\mu_T - \mu_R| \leq \theta_A,$$

where μ_T and μ_R are the population means of the log-transformed measure for the test and the reference, respectively, and (usually) $\theta_A = \log 1.25 = -\log 0.80 = 0.223$. This value is originated from the notion that the ratio of the population means in the original scale of $0.80 - 1.25$ (the mean of the test is $80 - 125\%$ of that of the reference) is considered as sufficiently close for drugs having an average therapeutic window.

Since Anderson and Hauck (1990) raised the issue of "switchability" between the old formulation and the new formulation, **individual bioequivalence** (IBE) and **population bioequivalence** (PBE) garnered more attention.

When a physician wants to switch a drug from an old formulation to a new one for her patient who has been titrated for the old formulation, she requires evidence that the new formulation is as safe and effective as the old. This concept is called switchability. Establishing IBE is intended to ensure switchability between two formulations within individuals. Anderson and Hauck (1990) defined two formulations as individually bioequivalent if they are sufficiently close for most subjects and proposed a method to evaluate IBE, based on the binomial distribution.

On the other hand, if physicians prescribe the new formulation for new patients, then there is a need to ensure that the two formulations are sufficiently close in the population. This concept is referred to as “prescribability”; population bioequivalence (PBE) is intended to ensure prescribability. The two formulations are declared population bioequivalent if the distributions (usually just the means and variances) of two formulations are sufficiently close. Thus, PBE conceptually includes ABE.

The USA FDA recommended replacing ABE with PBE and IBE. However, PBE and IBE are not required for approval of BE, perhaps because the suggested approach is not completely satisfactory from both practical and statistical viewpoints. Depending on the variability of the drug, they adopted the mixed-scaling approach for both PBE and IBE. A brief description of the current USA FDA guidance (2001) for PBE follows.

The test and the reference are population bioequivalent if the squared difference of their population means plus the difference in the total variances of the two formulations relative to a bounded version of the total variance of the reference is within the regulatory limit θ_P. That is:

$$\frac{(\mu_T - \mu_R)^2 + (\sigma^2_{TT} - \sigma^2_{RR})}{\max(\sigma^2_{RR}, \sigma^2_{T0})} \leq \theta_P,$$

where $\sigma^2_{TT} = \sigma^2_{WT} + \sigma^2_{BT}$ and $\sigma^2_{RR} = \sigma^2_{WR} + \sigma^2_{BR}$ are the total variances of the test and the reference. Here the subscripts W and B refer to “within” and “between” subjects. The constants, σ^2_{T0} and θ_P, are fixed regulatory standards.

As seen above, the USA guidance currently adopts an aggregate approach, using an aggregated test statistic for evaluating both means and variance components at the same time. In contrast, several disaggregate approaches have been suggested where tests for each component are performed separately. For example, Liu and Chow (1996) proposed to use a disaggregate approach for evaluating IBE where three components (intrasubject variability, subject-by-formulation interaction, and average) are separately tested applying multiple times of intersection-union tests. However, as the dimension (p) of tests increases, the power of the $(1-2\alpha)$ confidence set based approach, could decrease sharply for $p > 1$ as shown in Hwang (1996). Although the aggregate approach currently recommended by the FDA could avoid multiplicity issue, it is difficult to evaluate which component contributes bioinequivalence as well as the statistical properties for the suggested statistics are unknown.

We adopt a disaggregate approach, which can highlight a source of inequivalence more clearly.

2.4 Testing methodology

A review of the main articles in the development of BE tests reveals the (at a first glance) odd fact that $100(1-2\alpha)\%$ confidence intervals are often used when the level of type I error for the consumer's risk is to be controlled at most $\alpha\%$. In fact, there has been much debate among pharmaceutical scientists about which confidence interval level should be used, $100(1-2\alpha)\%$ or $100(1-\alpha)\%$. Table 1 illustrates several examples of BE tests with different operational confidence levels despite a constant desired nominal level of $\alpha = 0.05$. Currently, the USA FDA guidance adopts the two one-sided tests (TOST) as the standard method of ABE; hence, recommending the $100(1-2\alpha)\%$ confidence interval which (discussed below) is an operational equivalent of TOST.

Consider the problem where interest lies in estimating the difference in the population means of the two formulations, $\theta = \mu_T - \mu_R$. If BE holds, one would expect the estimate of θ to be within regulatory boundaries of 0, say between δ_L and δ_U. In this setting, the statement "the two formulations are bioequivalent if the $100(1-\boldsymbol{\alpha})\%$ confidence interval is contained within δ_L and δ_U" seems reasonable. On the other hand, consider casting the problem as two one-sided hypothesis tests consisting of the hypotheses $H_{01} : \theta \leq \delta_L$ vs. $H_{a1} : \theta \geq \delta_L$ and $H_{02} : \theta \geq \delta_U$ vs. $H_{a2} : \theta \leq \delta_U$. Then, the statement "the two formulations are bioequivalent if both null hypotheses are rejected at the level $\boldsymbol{\alpha}$" seems equally reasonable. However, if one accounts for the multiplicity of the two tests, then a difference in the methods and the confusion occur.

With regard this distinction Berger and Hsu (1996) commented in Section 2.3

> ... our conclusion is that the practice of defining bioequivalence tests in terms of $100(1-\boldsymbol{2\alpha})\%$ confidence intervals should be abandoned. If both a confidence interval and a test are required, a $100(1-\boldsymbol{\alpha})\%$ confidence intervals that corresponds to the given size-α test should be used.

They proved that the suggested $100(1-\alpha)\%$ confidence interval has the correct size. However, the suggested interval is exactly same as the classical $100(1-2\alpha)\%$ confidence interval when the interval includes zero, which is typical for most BE trials unless the variances of two formulations are very small or the two formulations are obviously bioinequivalent. Liu and Chow (1996) showed that the conclusion for bioequivalence/bioinequivalence would be the same from the two procedures in a variety of scenario.

These results beg the question of why these mathematically correct results defy experimental intuition. We believe that this conflict between intuition and mathematics, indicates a defect in the logical framework. This is one of the motivations of this research. We explore an alternative framework developed by Royall (1997), which does not suffer from some of the fundamental flaws in the current statistical practices in this area.

3 The Likelihood Paradigm

The source of the confusion amongst the frequentist approaches in BE trials arises from viewing the data as a decision making tool, rather than representing the data as evidence. Such practice skips the fundamental step of evaluating what the data say.

Given a statistical model for the observed data, the Law of likelihood plays the fundamental role in interpreting data as evidence. It is stated in Royall (1997):

Law of the Likelihood: If hypothesis A implies that the probability that a random variable X takes the value x is $p_A(x)$, while hypothesis B implies that the probability is $p_B(x)$, then the observation $X=x$ is evidence supporting A over B if and only if $p_A(x) > p_B(x)$, and the likelihood ratio, $p_A(x)/p_B(x)$, measures the strength of that evidence (Hacking, 1965).

This law has the so-called Likelihood Principle as its foundation. The Likelihood Principle, formally stated by Birnbaum (1962), suggests that under the assumption of a parametric statistical model, experimental results are fully characterized by the likelihood function. Therefore, two experiments resulting identical likelihood functions have the same evidential meaning.

The Likelihood Principle has far-reaching consequences for statistical practice. For example, it implies that other potential values of the data have no bearing on its evidential interpretation. Hence, frequency-style interpretations leading to P-values and confidence intervals, which depend on potential other values of the data or fictitious repetitions of the experiment, do not lead to evidential interpretations.

Royall, and other proponents of this likelihood paradigm, operationalize the Likelihood Principle using the Law of the Likelihood and suggest representing the data as evidence using a standardized likelihood plot. Reference lines drawn to indicate $1/k$ likelihood intervals can be used to summarize likelihood ratios. In particular, the values of $k = 8$ and 32 correspond to the likelihood ratios obtained when observing 3 and 5 successive heads when flipping a coin and comparing the (numerator) hypothesis that the coin is two headed versus the (denominator) hypothesis that the coin is fair. We refer to these values of k as “moderate” and “strong” evidence of the hypothesis in the numerator of the likelihood ratio over that in the denominator, respectively. (Of course, by symmetry, a likelihood ratio of $1/8$, for example, represent moderate evidence of the hypothesis in the denominator of the likelihood ratio to that in the numerator.) These values are akin to the $.05$ and $.01$ benchmarks commonly used to interpret P-values. We promote the use of this likelihood paradigm as an important first step in the analysis of bioequivalence trials.

An experiment needs not always produce moderate or strong evidence. For example, it may produce “weak evidence” in the form of a likelihood ratio between $1/k$ and k, or “misleading evidence” (Royall, 1997, 2000), where a likelihood of k (or $1/k$ respectively) is obtained when in fact the denominator (numerator) hypothesis is correct.

Strong misleading evidence cannot occur very often. A straightforward application of

Markov's inequality suggests that the probability of misleading evidence cannot exceed $1/k$, referred to as the universal bound by Royall (1997).

After a value for k is chosen, and after an experiment is completed, whether the data represent weak evidence or not is known. In contrast, it is impossible to know whether or not the evidence is misleading when the data produced strong evidence. Hence the misleading evidence is more important concept in this context. Later we evaluate the probability of such undesirable results, and adherence to the universal bound in the presence of nuisance parameters in the context of BE trials.

3.1 Likelihoods in the presence of nuisance parameters

When the likelihood function for a model is indexed by a single parameter, the likelihood provides the evidence for the parameter in the data, as stated in the Likelihood Principle. However, in the bioequivalence setting, the likelihood function typically has several parameters of interest, and nuisance parameters. As such, it is challenging to present the likelihood as a function of the parameter of interest alone.

Although there is no single universally adopted solution for eliminating nuisance parameters, there are several *ad-hoc* methods to circumvent this difficulty. Some of these methods include orthogonal parameterization, marginal likelihoods, conditional likelihoods, estimated likelihoods, and profile likelihoods (see Royall, 1997). The definitions for the estimated and profile likelihoods can be found in Pawitan (2001). Since marginal likelihoods, conditional likelihoods, and orthogonal likelihoods are all genuine likelihoods, they share the properties of likelihood, such as general results for the probability of misleading evidence. When these approaches are not available, we contend that the profile likelihood is the most promising alternative. Even though the universal bound on the probability of misleading evidence does not technically apply to profile likelihoods, it approximately applies to profile likelihoods for a large class of useful models. In the sequel we demonstrate, that in this setting, the profile likelihood is a good alternative and that the universal bound on the probability of misleading evidence appears to be applicable.

In Appendix A, we define our model without covariates and find the analytical solution for the profile likelihoods for the ratio of means and the ratio of variances of two formulations. In the presence of covariates, the profile likelihoods cannot be solved for analytically, but can be obtained numerically.

4 The Likelihood Paradigm: Application to BE Using Profile Likelihoods

Appropriate null and alternative hypotheses can be specified as follows:

$$H_1 : \theta \leq \delta_L \text{ or } \theta \geq \delta_U \quad \text{versus} \tag{1}$$

$$H_2 : \delta_L < \theta < \delta_U$$

where θ, is either the ratio of means or the ratio of variances, and the outcomes of interest are log transformed AUC and C_{max}. We begin by evaluating them separately. We note that a benefit of adopting the likelihood paradigm is that separate analyses do not require adjustment for multiplicity.

Examples of profile likelihood plots are shown in Figures 1-4. An accurate portrait of the evidence can be shown by a profile likelihood plot along with $1/k$ likelihood intervals and the predefined limit. The $1/k$ likelihood interval can be interpreted as follows: the best-supported value of the parameter θ (MLE) is at least k times better supported than all of the values outside the interval. The $1/k$ likelihood interval can be used as a measure of strength of evidence for BE versus BIE. If the $1/8$ likelihood interval lies completely within the limit, the data fairly strongly support BE. If the $1/32$ one lies within the limit, the data represent strong evidence for BE.

We illustrate how to evaluate ABE and PBE within the likelihood paradigm using data from Chow and Liu (2000) and several modified versions of their data where variances and period effects are modified.

4.1 Evidence for equivalence of the ratio of means (ABE)

The left panel in Figure 1 shows the profile likelihood for the ratio of means where the $1/32$ likelihood interval completely lies within the BE limit. Thus, the data provide strong evidence that the two formulations are average bioequivalent. With the 90% and 95% confidence intervals, two formulations are also to be concluded as BE. The 95% confidence intervals in the figure are almost the same as the $1/8$ likelihood intervals. There is a straightforward reason for this agreement. Specifically, Royall (1997) showed that if the measurements follow a normal distribution, the $1/8$ and $1/32$ likelihood intervals are approximately the same as the 95% and 99% confidence intervals, respectively.

To examine the effect of the variance in evaluating ABE, we artificially modified the data. First, the empirical standard deviation of the test formulation was increased by 70% compared to the reference. In the middle panel of Figure 1, the $1/8$ likelihood interval does not completely fall within the limit, but the $1/5$ one does. Thus, there is only weak evidence in favor of BE over BIE, even though the TOST (equivalently the 90% confidence interval) concludes BE.

Secondly, we modified the data so that both the standard deviations of the test and reference formulations are inflated by 50%. The profile likelihood is shown in the right panel of Figure 1. Notice that the interval is wide enough so that neither the $1/k$ likelihood intervals nor the 90% confidence interval lie within the regulatory limits; the profile likelihood plot suggests that the data does not provide enough evidence to support either BE or BIE. In contrast, TOST concludes BIE. The width of the profile likelihoods increases as the variability increases. The figures clearly show that the variability is large relative to the scale of interest. These results suggest a comparison of the variances of the two formulations.

4.2 Evidence for equivalence of the ratio of variances (PBE)

The profile likelihoods for the ratio of variances are shown in Figure 2 for the original (left panel) and two modified data sets (middle and right panels), respectively. The left panel of Figure 2 shows that the data support the equivalence of the variances. Notice that the profile likelihood for the ratio of variances is much wider than that for the ratio of the means. Therefore, more subjects are required to estimate the ratio of variances precisely.

On the other hand, the middle panel of Figure 2 suggests that there is clear evidence that the variance of the test formulation is larger than that of the reference. Thus, the two formulations do not appear to be population bioequivalent, even though they do appear to be average bioequivalent.

The right panel of Figure 2 appears to support the equivalence of the variances, though the $1/8$ interval is wide, ranging from $0.7-1.7$. This suggests, along with Figure 1, that we do not have enough information to clearly see whether the data supports bioequivalence or not. It is worth noting that after a second stage of data collection, it is straightforward to combine the information from the two stages within the likelihood paradigm. Specifically, there is no need for adjustment P-value as is required for frequentist sequential trials. Instead, one simply combines the two data sets and plots the profile likelihood for the parameter of interest.

4.3 Evaluating AUC and C_{max} jointly

In the current practice of BE trials in the USA, bioequivalence is determined using both AUC and C_{max}. Typically, these metrics are evaluated separately. Usually AUC and C_{max} are highly correlated, as they are calculated based on the drug concentrations measured from the same subject. Thus, it is natural to treat them as a vector of four measurements within each subject: AUC and C_{max} for the test and reference formulations, respectively. Let $\left(Y_{Ri}^{(A)}, Y_{Ti}^{(A)}, Y_{Ri}^{(C)}, Y_{Ti}^{(C)}\right)$ be the log-transformed AUC and C_{max} for the reference and the test on subject i. Assume that the distribution of these measures follows a multivariate normal (MVN) as:

$$\begin{pmatrix} Y_{Ri}^{(A)} \\ Y_{Ti}^{(A)} \\ Y_{Ri}^{(C)} \\ Y_{Ti}^{(C)} \end{pmatrix} \sim \text{MVN} \left\{ \begin{pmatrix} \mu_R^{(A)} \\ \mu_T^{(A)} \\ \mu_R^{(C)} \\ \mu_T^{(C)} \end{pmatrix}, \ \Omega \right\}, \tag{2}$$

where Ω is a covariance matrix. We reparametrize such that $\mu_T^{(A)} = \mu_R^{(A)} + \theta^{(A)}$ and $\mu_T^{(C)} = \mu_R^{(C)} + \theta^{(C)}$. Hence, $\theta^{(A)}$ and $\theta^{(C)}$ are the mean differences between two formulations for each outcome. Using the joint likelihood, we can find the profile likelihood with respect to $\theta^{(A)}$ and $\theta^{(C)}$ one at a time. That is, we treat one of them as the parameter of interest and consider the other as nuisance parameter along with other nuisance parameters.

Figure 3 shows the profile likelihood for $\theta^{(A)}$ and $\theta^{(C)}$ using a data set obtained from a recent BE trial with 48 subjects performed at a pharmaceutical company. The data were perturbed prior to analysis, to maintain confidentiality. The profile likelihood for $\theta^{(A)}$ represents strong evidence supporting BE for AUC, whereas the profile likelihood for $\theta^{(C)}$ does not support BE for C_{max}. The C_{max} of the test formulation is smaller than C_{max} of the reference. Thus, the test formulation appears to be absorbed more slowly than the reference even though overall amounts of drug absorbed are similar.

4.4 Evaluating potential confounding effects

The profile likelihoods with and without adjusting for covariates (sequence and period) are shown in Figure 4 using the data from Chow and Liu (2000) (left panel) and another modified version of their data (right panel), where the values for the second period were about 10% increased from the mean; hence in this modified data set, a period effect is present.

The $1/5$, $1/8$ and $1/32$ likelihood intervals along with 90% and 95% confidence intervals are shown for comparison. When there are no period and sequence effects, the profile likelihoods with and without adjustment are almost same (left panel). In contrast, the likelihood without adjustment is much flatter than the one with adjustment (right panel) when a period effect really exists. This illustrates the point that when period or sequence effects exist, the unadjusted profile likelihood will represent weaker evidence than the adjusted one, because the variation explained by the period effects gets absorbed into the error. As confounding effects, such as carry-over effects, which are indistinguishable from treatment-period interaction or sequence effects, could be present in cross-over designs, it is advisable to always look at the profile likelihoods, with and without adjustment. A large discrepancy between the two suggests potential carry-over effects, treatment-period interaction or sequence effects.

4.5 Probabilities of weak and misleading evidence

We examine the probabilities of weak and misleading evidence produced by the profile likelihoods in the BE setting. As there is no closed form solution for calculating the probabilities of weak and misleading evidence for the interval hypotheses (1) used in evaluating BE, a simulation study using Model (3) was performed assuming that the two formulations are marginally bioinequivalent with common error variances. We focused on the degree of similarity to the true likelihood under parameter values that are reasonable for BE studies. As an analogue of the probability of misleading evidence, we estimated the probability of incorrectly presenting the data as BE using likelihood intervals (given k) obtained from the true, profile and estimated likelihoods.

The probability of incorrectly presenting the data as BE ("misleading probability") is calculated as the number of times the entire $1/k$ likelihood interval is contained within the regulatory limits divided by the number of simulations. Figure 6 shows the estimated probabilities of misleading evidence as functions of k and the sample size n for $\rho = (0.5, 0.7)$ and $\sigma = (0.1, 0.2, 0.3)$. The type I error probability for TOST and a reference line 0.05 are shown for comparison. Notice that the probabilities of misleading evidence from the true and profile likelihoods are almost the same, regardless of the sample size, parameter values and choice of k. This small difference diminishes as the sample size increases.

In contrast, the probability of misleading evidence from the estimated likelihood is much larger than those from the true and profile likelihoods.

Interestingly, the probabilities of misleading evidence from the true and profile likelihoods always achieve the maximum possible values for a given k, for a wide range of sample sizes and model parameters. It is interesting to contrast this observation from what is known for point hypotheses, whose probability of misleading evidence goes to zero with the sample size. A reason for this phenomenon can be explained as follows. For point hypotheses, where Δ (the difference in the two hypothesized means) is fixed, the maximum probability of misleading evidence is reached at $n = (2 \log k)(\sigma/\Delta)^2$ (see pages 90-93 in Royall (1997)). In contrast, for interval hypotheses in this BE setting, Δ is varying and hence there are many sample sizes where the maximum probability of misleading evidence can be reached. Thus, the probability of misleading evidence persists for a wide range of sample sizes. Notice, however, that the probabilities of misleading evidence for the profile likelihood does not go beyond that maximum value, suggesting that the universal bound is applicable. Also of note is that it does not appear to be applicable to the estimated likelihood.

Another simulation was performed using the same model, but the two formulations were assumed to be bioequivalent. We examine the probability of failure to present evidence for BE, an analogue of the probability of weak evidence. This is akin to the Type II error probability. However, we present this property in terms of the probability of presenting evidence for BE when the two formulations are truly bioequivalent (the analogue of power), as shown in Figure 7. The profile likelihood represents the data as BE less often than it should, but eventually becomes close to that of the true likelihood as either σ decreases or the sample size increases. Because the discrepancy between the pseudo-likelihoods and the true likelihood tend to zero (with probability one) as $n \to \infty$, eventually the probabilities of presenting evidence for BE based on all three likelihoods also tend to one as the sample size increases (regardless of k).

5 Summary and Discussion

In this manuscript we explored an alternative method for presenting and interpreting bioequivalence data as evidence, using likelihood methods. Motivated by simulations studies and prior theoretical development, we recommend the use of the profile likelihood as the relevant measure of evidence in the presence of nuisance parameters. In particular, the simulations results suggest that the profile likelihood behaves similarly to the true likelihood, as long as the sample size is moderate. For example, with 14 subjects in each treatment sequence and $\sigma = 0.2$ (a moderate error variance in BE trials) the probability of presenting fairly strong evidence ($k = 8$) using the profile likelihood is more than 0.95, which is similar to that of the true likelihood. In addition, regardless of the parameter values and the sample size, the probability of misleading evidence is very small, about 0.02, which is very similar to that of the true likelihood, suggesting that the universal bound for the probability of misleading evidence is applicable.

We also presented a straightforward extension of likelihood analysis to evaluate population bioequivalence and similarly considered ABE via likelihood analysis jointly for AUC and C_{max}. Our likelihood paradigm also can be extended easily to evaluate IBE if the three components (intrasubject variability, subject-by-formulation interaction, and average) would be sufficient. However, it requires higher order cross-over design which is unfavorable design from both statistical and study subjects' viewpoints. Finding a good metric for evaluating IBE without relying on higher order cross-over design would be promising research area.

The standard method in the current practice of BE trials, TOST, is based on the Neyman-Pearson testing theory. Likelihood theory and Neyman-Pearson testing theory have much in common, in that they both explicitly compare two hypotheses and depend on likelihood ratios. The simulation studies suggested that the overall properties of TOST are similar to those of the profile likelihood with $k = 4.5$, which only represents weak evidence. However, with TOST, it is difficult to see how much the data support BE or BIE, because of the emphasis on decision making rather than evidential interpretation. On the other hand, the likelihood plot gives the most direct and complete representation of the data as evidence.

Finally, we would also note that the decision for declaring BE or BIE is ultimately in the hands of the regulatory authorities and clinical pharmacologists. After examining what the data say, the regulatory authorities can decide BE or BIE depending on the characteristics of drug. For example, if the therapeutic index of a drug is narrow, they might want to use a more strict criteria. In contrast, if the therapeutic index of a drug is wide and the variability of a drug is large, then a less stringent criteria might be applied, or additional data required. In this manuscript, we clarified the distinction between evidence and decision making in the BE setting and hence proposed a new paradigm to represent bioequivalence data as evidence.

References

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

Anderson, S. and Hauck, W. W. (1990). Consideration of individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics*, 18:259–273.

Berger, R. L. and Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets (Disc: P303-319). *Statistical Science*, 11:283–302.

Birnbaum, A. (1962). On the foundations of statistical inference (Com: P307-326). *Journal of the American Statistical Association*, 57:269–306.

Bois, F., Tozer, T., W.W., H., M.L., C., R., P., and R.L., W. (1994). Bioequivalence: performance of several measures of rate of absorption. *Pharmaceutical Research*, 11:966–974.

Chen, M. (1992). An alternative approach for assessment of rate of absorption in bioequivalence studies. *Pharmaceutical Research*, 9:1380–1385.

Chow, S.-C. and Liu, J.-P. (2000). *Design and analysis of bioavailability and bioequivalence studies*. Marcel Dekker, 2nd edition.

D'Angelo, G., Potvin, D., and Turgeon, J. (2001). Carry-over effects in bioequivalence studies. *Journal of Biopharmaceutical Statistics*, 11:35–43.

Endrenyi, L., Fritsch, S., and Yan, W. (1991). Cmax/auc is a clearer measure than cmax for absorption rates in investigations of bioequivalence. *International Journal of Clinical Pharmacology, Therapy, and Toxicology*, 29:394–399.

FDA guidance (2001). Statistical approaches to establishing bioequivalence.

Friedman, H., Greenblatt, D., Burstein, E., Harmatz, J., and Shader, R. (1986). Population study of triazolam pharmacokinetics. *British Journal of Clinical Pharmacology*, 22:639–642.

Hacking, I. (1965). *Logic of statistical inference*. New York: Cambridge University Press.

Hwang, J. T. G. (1996). Comment on "Bioequivalence trials, intersection-union tests and equivalence confidence sets". *Statistical Science*, 11:313–315.

Jones, B. and Kenward, M. G. (2003). *Design and analysis of cross-over trials*. Monographs on statistics and applied probability. Chapman & Hall/CRC, second edition.

Kenney, J. and Keeping, E. (1951). *Mathematics of Statistics*. D. Van Nostrand, New York.

Kirkwood, T. (1981). Bioequivalence testing: a need to rethink (reader reaction). *Biometrics*, 37:589–591.

Lacey, L., Keene, O., Pritchard, J., and Bye, A. (1997). Common noncompartmental pharmacokinetic variables: Are they normally or log-normally distributed? *Journal of Biopharmaceutical Statistics*, 7:171–178.

Limpert, E., Stahel, W., and Abbt, M. (2001). Log-normal distributions across the sciences: Keys and clues. *BioScience*, 51:341–352.

Liu, J.-p. and Chow, S.-C. (1996). Comment on "Bioequivalence trials, intersection-union tests and equivalence confidence sets". *Statistical Science*, 11:306–312.

Locke, C. (1984). An exact confidence interval for untransformed data for the ratio of two formulation means. *Journal of Pharmacokinetics and Biopharmaceutics*, 12:649–655.

Metzler, C. M. (1974). Bioavailability - a problem in equivalence. *Biometrics*, 30:309–317.

Midha, K., Ormsby, E., Hubbard, J., McKay, G., Hawes, E., and Gavalas, L. (1993). Logarithmic transformation in bioequivalence: application with two formulations of perphenazine. *Journal of Pharmaceutical Sciences*, 82:138–144.

Mizuta, E. and Tsubotani, A. (1985). Preparation of mean drug concentration-time curves in plasma: A study on the frequency distribution of pharmacokinetic parameters. *Chemical Pharmaceutical Bulletin*, 33:1620–1632.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press.

Royall, R. (2000). On the probability of observing misleading statistical evidence (C/R: P768-780). *Journal of the American Statistical Association*, 95(451):760–768.

Royall, R. M. (1997). *Statistical evidence: a likelihood paradiam*. Chapman & Hall/CRC.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680.

Schulz, H.-U. and Steinijans, V. W. (1991). Striving for standards in bioequivalence assessment: a review. *International Journal of Clinical Pharmacology, Therapy and Toxicology*, 29:293–298.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52:591–611.

Westlake, W. J. (1976). Symmetric confidence intervals for bioequivalence trials. *Biometrics*, 32:741–744.

Westlake, W. J. (1988). *Biophamaceutical statistics for drug development*, chapter Bioavailability and bioequivalence of pharmaceutical formulations, pages 329–352. Marcel Dekker.

Zariffa, N. M.-D., Patterson, S. D., Boyle, D., and Hyneck, M. (2000). Case studies, practical issues and observations on population and individual bioequivalence. *Statistics in Medicine*, 19(20):2811–2820.

A Profile Likelihoods

A.1 Profile likelihood for the ratio of means in evaluating ABE

When there is no sequence or period effects, the measures for the test and the reference formulations from a 2×2 cross-over BE trial can be assumed to be bivariate log-normal. We assume that the distribution of log transformed test and reference measures on the ith subject, $Y_{Ri} = \log Y^*_{Ri}$ and $Y_{Ti} = \log Y^*_{Ti}$, follows a bivariate normal as:

$$\left(\begin{array}{c} Y_{Ri} \\ Y_{Ti} \end{array} \right) \sim \text{BVN} \left(\left(\begin{array}{c} \mu_R \\ \mu_T \end{array} \right), \left(\begin{array}{cc} \sigma_R^2 & \rho\sigma_R\sigma_T \\ & \sigma_T^2 \end{array} \right) \right). \tag{3}$$

Let y_{Ri} and y_{Ti} be log transformed observations for the reference and the test formulations on the ith subject, $i = 1, \ldots, n$, and $\boldsymbol{y_R}$ and $\boldsymbol{y_T}$ be the associated vectors. The likelihood function for $\mu_R, \mu_T, \sigma_R, \sigma_T, \rho$ can be written as:.

$$
\begin{aligned}
&L(\mu_R, \mu_T, \sigma_R, \sigma_T, \rho \mid \boldsymbol{y_R}, \boldsymbol{y_T}) \\
=&\prod_{i=1}^{n} \frac{1}{2\pi\sigma_R\sigma_T\sqrt{1-\rho^2}} \\
&\times \exp\Big\{-\frac{1}{2(1-\rho^2)}\Big[\frac{(y_{Ri}-\mu_R)^2}{\sigma_R^2} - 2\rho\frac{(y_{Ri}-\mu_R)(y_{Ti}-\mu_T)}{\sigma_R\sigma_T} + \frac{(y_{Ti}-\mu_T)^2}{\sigma_T^2}\Big]\Big\},
\end{aligned} \tag{4}
$$

where $\sigma_R > 0$, $\sigma_T > 0$ and $-1 < \rho < 1$.

After exponentiating, the difference of means in the log transformed scale is the ratio of the means in the original scale. Note that this equivalence relationship is only true in the instance of equal variances for the two formulations. More precisely, the ratio of medians in the original scale is equivalent to the exponentiated difference of medians in the log transformed scale. Regardless, we focus entirely on the difference of means in the log scale even though we allow non-constant variance across the two arms. This is because we are interested in whether or not the central tendency of the two formulations are sufficiently close.

We reparametrize $\theta = \mu_T - \mu_R$ and $\gamma = \mu_R$, and reexpress the likelihood function for $\theta, \gamma, \sigma_R, \sigma_T, \rho$ as:

$$
\begin{aligned}
&L(\theta, \gamma, \sigma_R, \sigma_T, \rho \mid \boldsymbol{y_R}, \boldsymbol{y_T}) \\
\propto&\Big(\frac{1}{\sigma_R^2\sigma_T^2(1-\rho^2)}\Big)^{n/2} \exp\Big\{-\frac{1}{2(1-\rho^2)} \\
&\times \Big[\frac{\sum_{i=1}^n (y_{Ri}-\gamma)^2}{\sigma_R^2} - 2\rho\frac{\sum_{i=1}^n (y_{Ri}-\gamma)(y_{Ti}-\theta-\gamma)}{\sigma_R\sigma_T} + \frac{\sum_{i=1}^n (y_{Ti}-\theta-\gamma)^2}{\sigma_T^2}\Big]\Big\}.
\end{aligned} \tag{5}
$$

The profile likelihood of θ and γ for the likelihood function (5) can be written as:

$$
\begin{aligned}
&L_P(\theta, \gamma \mid \boldsymbol{y_R}, \boldsymbol{y_T}) = \max_{\sigma_R, \sigma_T, \rho} L(\theta, \gamma, \sigma_R, \sigma_T, \rho \mid \boldsymbol{y_R}, \boldsymbol{y_T}) = L(\theta, \gamma, \tilde{\sigma}_R, \tilde{\sigma}_T, \tilde{\rho} \mid \boldsymbol{y_R}, \boldsymbol{y_T}) \\
&\propto \Big\{\sum_{i=1}^n (y_{Ri}-\gamma)^2 \sum_{i=1}^n (y_{Ti}-\theta-\gamma)^2 - \Big[\sum_{i=1}^n (y_{Ri}-\gamma)(y_{Ti}-\theta-\gamma)\Big]^2\Big\}^{-n/2},
\end{aligned} \tag{6}
$$

where

$$\begin{aligned}\tilde{\sigma}_R^2 &= \frac{\sum_{i=1}^n (y_{Ri}-\gamma)^2}{n},\\ \tilde{\sigma}_T^2 &= \frac{\sum_{i=1}^n (y_{Ti}-\theta-\gamma)^2}{n} \quad \text{and}\\ \tilde{\rho} &= \frac{\sum_{i=1}^n (y_{Ri}-\gamma)(y_{Ti}-\theta-\gamma)}{\sqrt{\sum_{i=1}^n (y_{Ri}-\gamma)^2 \sum_{i=1}^n (y_{Ti}-\theta-\gamma)^2}}.\end{aligned}$$

Then the profile likelihood of θ is:

$$\begin{aligned}L_P(\theta) = L_P(\theta \mid \boldsymbol{y_R}, \boldsymbol{y_T}) &= \max_{\gamma}\ L_P(\theta, \gamma \mid \boldsymbol{y_R}, \boldsymbol{y_T}) = L_P(\theta, \tilde{\gamma} \mid \boldsymbol{y_R}, \boldsymbol{y_T})\\ &\propto \Big\{\sum_{i=1}^n (y_{Ri}-\tilde{\gamma})^2 \sum_{i=1}^n (y_{Ti}-\theta-\tilde{\gamma})^2 - \Big[\sum_{i=1}^n (y_{Ri}-\tilde{\gamma})(y_{Ti}-\theta-\tilde{\gamma})\Big]^2\Big\}^{-n/2},\end{aligned}$$

where

$$\tilde{\gamma} = \frac{\sum y_{Ri}\Big[\sum(y_{Ti})^2-\sum(y_{Ri}y_{Ti})\Big]+\sum y_{Ti}\Big[\sum(y_{Ri})^2-\sum(y_{Ri}y_{Ti})\Big]-\theta\sum y_{Ri}\Big(\sum y_{Ti}-\sum y_{Ri}\Big)-n\theta\Big[\sum(y_{Ri})^2-\sum(y_{Ri}y_{Ti})\Big]}{n\sum(y_{Ti}-y_{Ri})^2-\Big(\sum y_{Ti}-\sum y_{Ri}\Big)^2}.$$

A.2 Profile likelihood for the ratio of variances in evaluating PBE

The parameter of interest is the ratio of variances σ_T/σ_R while the means μ_R and μ_T and ρ are the nuisance parameters. Using the reparameterization $\theta = \sigma_T/\sigma_R$ and $\gamma = \sigma_R$, the likelihood function (4) for $\mu_R, \mu_T, \sigma_R, \sigma_T, \rho$ can be reexpressed as:

$$\begin{aligned}&L(\mu_R, \mu_T, \theta, \gamma, \rho \mid \boldsymbol{y_R}, \boldsymbol{y_T}) \qquad (7)\\ &= \prod_{i=1}^n \frac{1}{2\pi\gamma^2\theta\sqrt{1-\rho^2}}\ \exp\Big\{-\frac{1}{2(1-\rho^2)}\\ &\quad \times \big[\frac{(y_{Ri}-\mu_R)^2}{\gamma^2} - 2\rho\frac{(y_{Ri}-\mu_R)(y_{Ti}-\mu_T)}{\gamma^2\theta} + \frac{(y_{Ti}-\mu_T)^2}{\gamma^2\theta^2}\big]\Big\},\end{aligned}$$

where $-1 < \rho < 1$.

The profile likelihood of θ for the likelihood function (7) can be written as:

$$\begin{aligned} L_P(\theta) = L_P(\theta \mid \boldsymbol{y_R}, \boldsymbol{y_T}) &= \max_{\mu_R,\mu_T,\gamma,\rho} \; L(\mu_R, \mu_T, \theta, \gamma, \rho \mid \boldsymbol{y_R}, \boldsymbol{y_T}) \\ &= L(\tilde{\mu}_R, \tilde{\mu}_T, \theta, \tilde{\gamma}, \tilde{\rho} \mid \boldsymbol{y_R}, \boldsymbol{y_T}) \propto \left(\frac{1}{\tilde{\gamma}^2\theta\sqrt{1-\tilde{\rho}^2}}\right)^n \text{exp}\Big\{-\frac{1}{2(1-\tilde{\rho}^2)} \\ &\times \big[\frac{\sum_{i=1}^n (y_{Ri}-\tilde{\mu}_R)^2}{\tilde{\gamma}^2} - 2\tilde{\rho}\frac{\sum_{i=1}^n (y_{Ri}-\tilde{\mu}_R)(y_{Ti}-\tilde{\mu}_T)}{\tilde{\gamma}^2\theta} + \frac{\sum_{i=1}^n (y_{Ti}-\tilde{\mu}_T)^2}{\tilde{\gamma}^2\theta^2}\big]\Big\}, \end{aligned} \tag{8}$$

where

$$\begin{aligned} \tilde{\mu}_R &= \frac{\sum_{i=1}^n y_{Ri}}{n} = \bar{y}_R, \\ \tilde{\mu}_T &= \frac{\sum_{i=1}^n y_{Ti}}{n} = \bar{y}_T, \\ \tilde{\rho} &= \frac{\sum_{i=1}^n (y_{Ri}-\bar{y}_R)(y_{Ti}-\bar{y}_T)}{\sqrt{\sum_{i=1}^n (y_{Ri}-\bar{y}_R)^2 \sum_{i=1}^n (y_{Ti}-\bar{y}_T)^2}} \quad \text{and} \\ \tilde{\gamma}^2 &= \frac{1}{2n(1-\tilde{\rho}^2)}\Big[\sum_{i=1}^n (y_{Ri}-\tilde{\mu}_R)^2 - 2\tilde{\rho}\frac{\sum_{i=1}^n (y_{Ri}-\tilde{\mu}_R)(y_{Ti}-\tilde{\mu}_T)}{\theta} + \frac{\sum_{i=1}^n (y_{Ti}-\tilde{\mu}_T)^2}{\theta^2}\Big]. \end{aligned}$$

Notice that only $\tilde{\gamma}$ depends on the parameter of interest.

Paper	Operational Method
Metzler (1974); Kirkwood (1981)	$100(1-\alpha)\%$ confidence interval
FDA guidance (2001)	$100(1-2\alpha)\%$ confidence interval
Westlake (1976)	$100(1-\alpha)\%$ symmetric confidence interval
Anderson and Hauck (1983)	P-value (level α)
Locke (1984)	$100(1-\alpha)\%$ confidence interval
Schuirmann (1987)	two one-sided tests (level α for each test)

Table 1: The comparison of operational methods and the nominal level of α among several proposed BE tests.

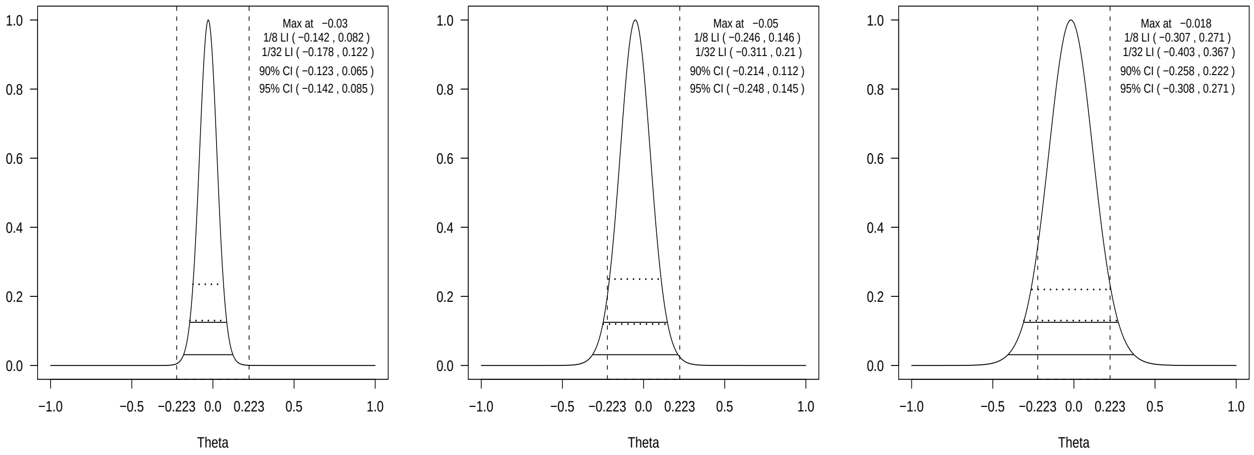

Figure 1: The profile likelihood, 1/8 (upper solid line) and 1/32 (lower solid line) likelihood intervals for the difference of means of $\log AUC$ using the data in Chow and Liu (2000) (left panel), the modified version of Chow and Liu's data (2000) where the standard deviation of the test drug is 1.7 times greater than the standard deviation of the reference drug (middle panel), and the modified version of Chow and Liu's data (2000) where the standard deviations of the test drug and the reference drug are both inflated by 50% (right panel). The horizontal dotted lines represent the 90% (upper) and 95% (lower) confidence intervals estimated by a random effects model without covariates and the vertical lines represent the regulatory lower (δ_L) and upper (δ_U) limits.

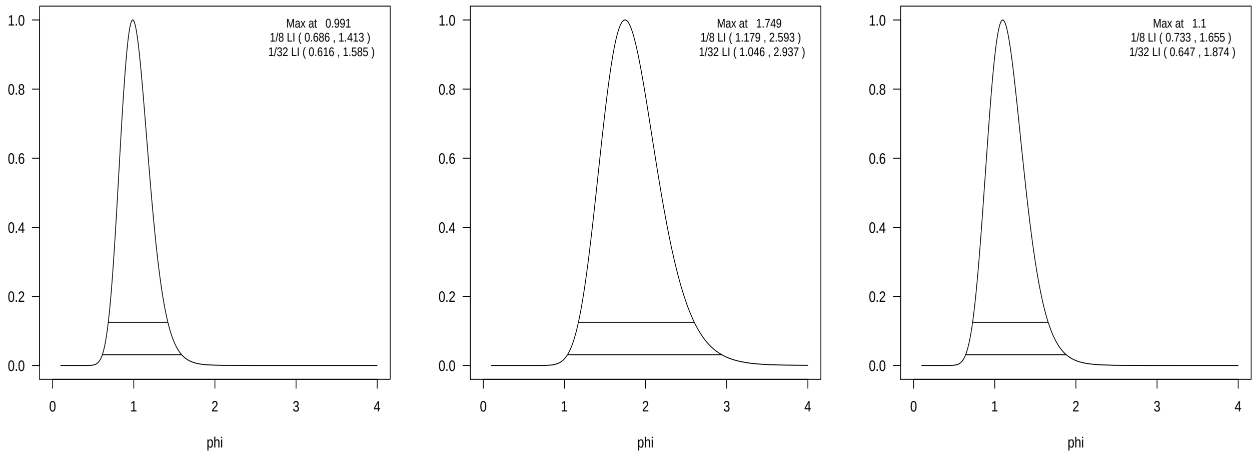

Figure 2: The profile likelihood, 1/8 (upper solid line) and 1/32 (lower solid line) likelihood intervals of the ratio of variances of the test drug and the reference drug using the data in Chow and Liu (2000) (left panel), the modified version of Chow and Liu's data (2000) where the standard deviation of the test drug is 1.7 times greater than the standard deviation of the reference drug (middle panel), and the modified version of Chow and Liu's data (2000) where the standard deviations of the test drug and the reference drug are both inflated by 50% (right panel). Notice that there are no regulatory limits available for the ratio of variances.

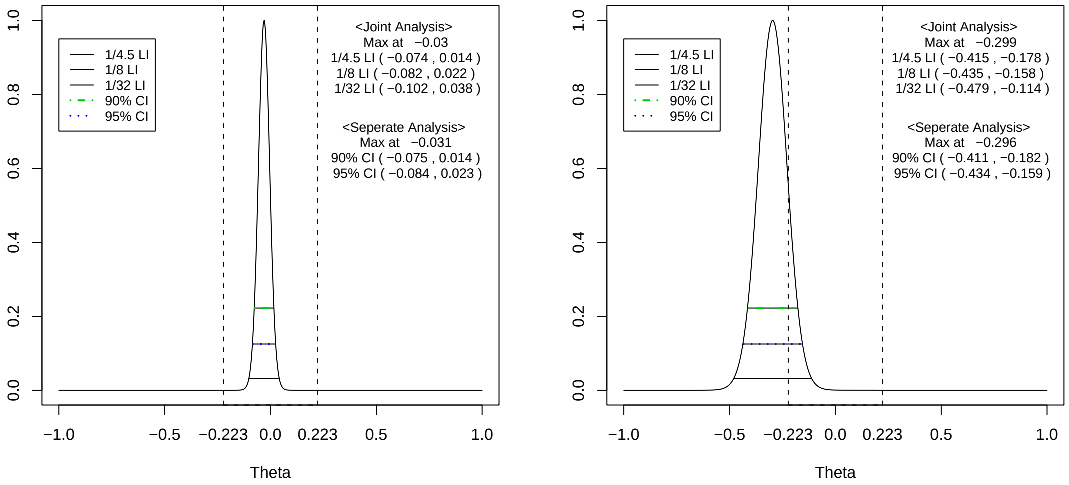

Figure 3: The profile likelihood, 1/4.5 (upper solid line), 1/8 (middle solid line) and 1/32 (lower solid line) likelihood intervals of $\theta^{(A)}$ (left panel) and $\theta^{(C)}$ (right panel). The horizontal dotted lines represent the 90% (upper) and 95% (lower) confidence intervals estimated by a random effects model without covariates and the vertical lines represent the regulatory lower (δ_L) and upper (δ_U) limits. The joint likelihood for AUC and C_{max} are profiled one at a time.

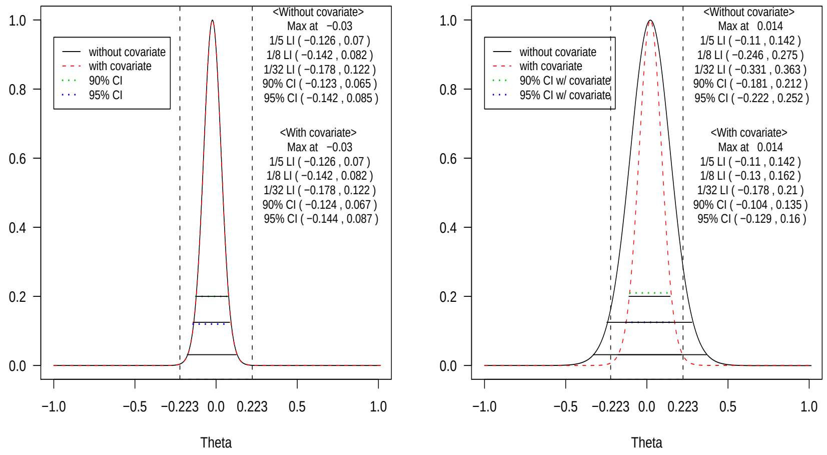

Figure 4: The profile likelihoods with and without covariates, 1/5 (upper solid line), 1/8 (middle solid line) and 1/32 (lower solid line) likelihood intervals for the difference of means of $\log AUC$ using the data in Chow and Liu (2000) (left panel) and the modified version of Chow and Liu's data (2000) where the values for the second period are inflated so that period effect exists (right panel). The horizontal dotted lines represent the 90% (upper) and 95% (lower) confidence interval estimated by a random effects models with and without covariates and the vertical lines represent the regulatory lower (δ_L) and upper (δ_U) limits.

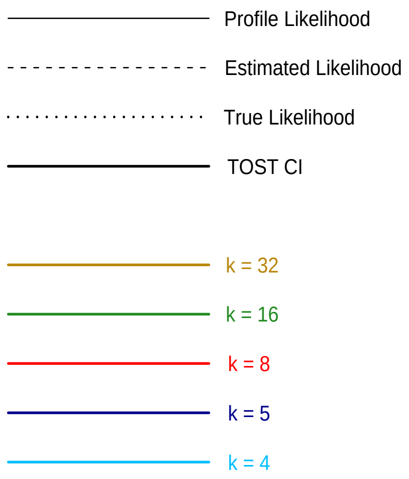

Figure 5: The legend used for Figures 6 and 7.

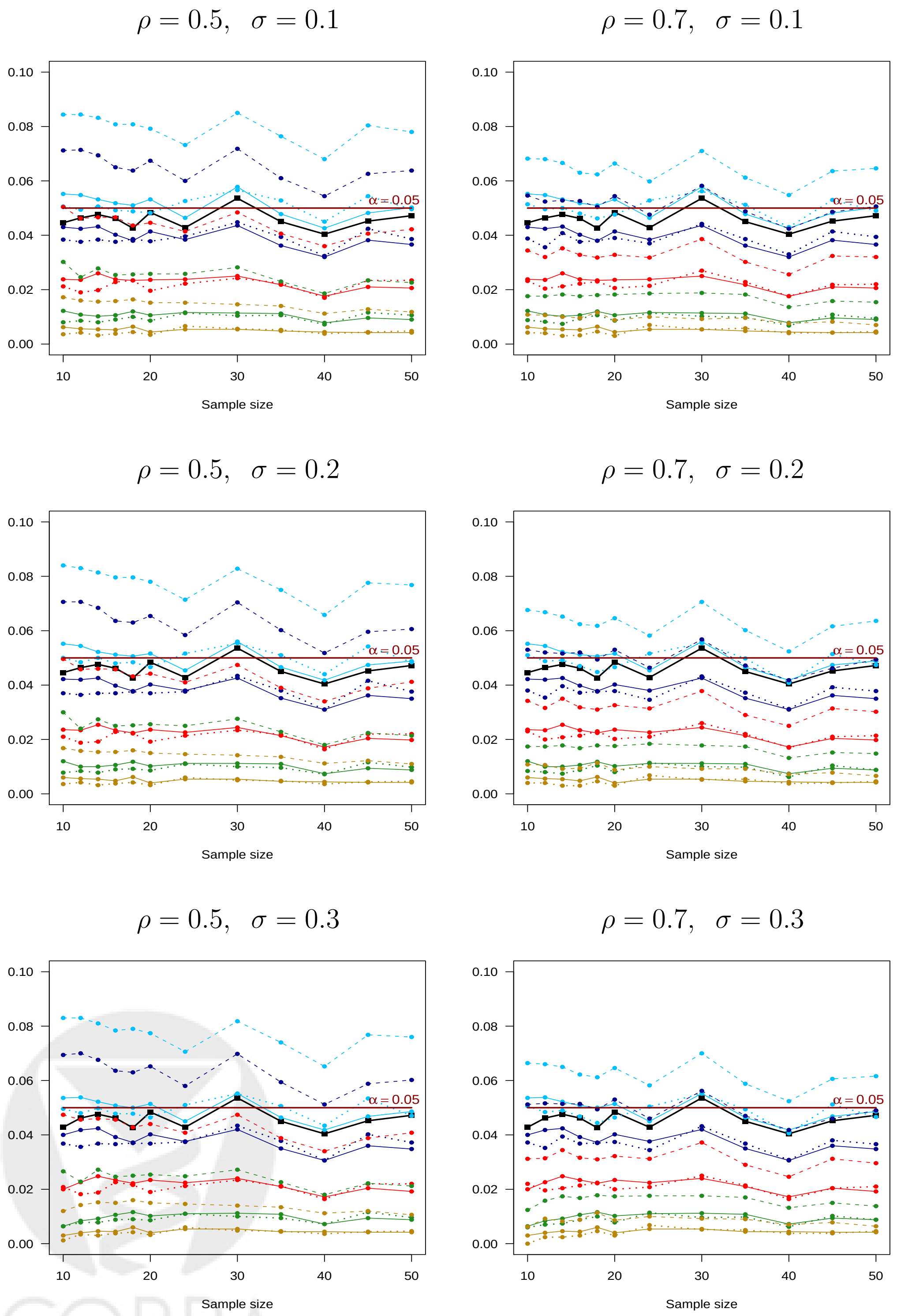

Figure 6: The probability of producing evidence for BE using the true, profile and estimated likelihood intervals when the two formulations are marginally bioinequivalent ($\theta = \theta_L$) as a function of $k = 4, 5, 8, 16, 32$ for $\rho = 0.5$ (left panel), $\rho = 0.7$ (right panel), $\sigma = 0.1$ (top panel), $\sigma = 0.2$ (middle panel) and $\sigma = 0.3$ (bottom panel). The type I error for TOST and the line for ($\alpha = 0.05$) is shown for comparison.

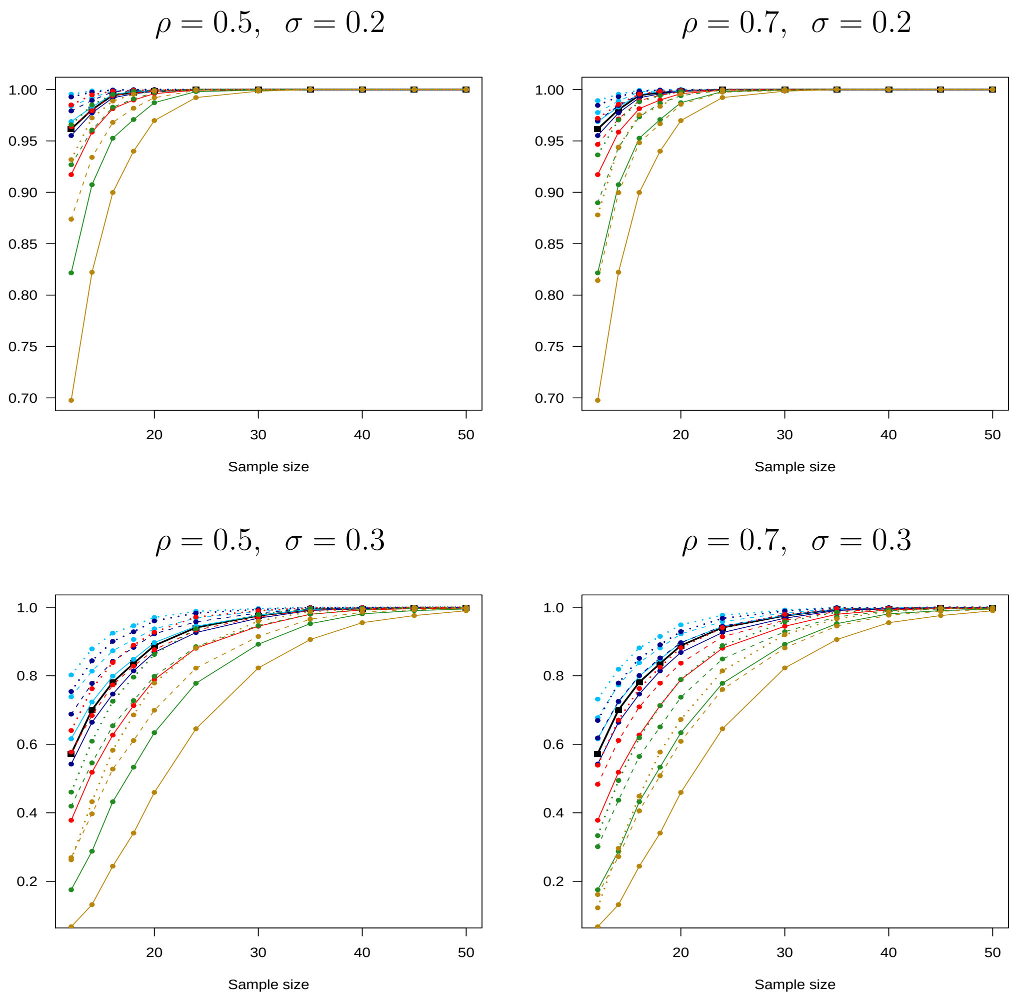

Figure 7: The probability of correctly concluding BE using the true, profile and estimated likelihood intervals when the two formulations are truly bioequivalent as a function of $k = 4, 5, 8, 16, 32$ for $\rho = 0.5$ (left panel), $\rho = 0.7$ (right panel), $\sigma = 0.2$ (top panel) and $\sigma = 0.3$ (bottom panel). The power for TOST is shown for comparison.