2013

Statistical considerations of noninferiority, bioequivalence and equivalence testing in biosimilars studies

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

STATISTICAL CONSIDERATIONS OF NONINFERIORITY, BIOEQUIVALENCE AND EQUIVALENCE TESTING IN BIOSIMILARS STUDIES

by

SIYAN XU

B.A., Wuhan University, 2006
M.S., Marquette University, 2009

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2013

Approved by

First Reader [redacted]

Ralph B. D'Agostino, Sr., Ph.D.
Professor of Mathematics and Statistics and Professor of Epidemiology and Biostatistics; Director, Statistics and Consulting Unit; Executive Director, Biostatistics Program

Second Reader [redacted]

Kerry B. Barker, Ph.D.
Senior Director, Biotechnology Clinical Development of Pfizer Inc.

Third Reader [redacted]

Sandeep M. Menon, Ph.D.
Adjunct Professor of Biostatistics; Executive Director, Biotherapeutics Clinical Research of Pfizer Inc.

Acknowledgements

I would like to thank God Almighty for all His love and blessings.

I am sincerely grateful to my thesis advisor, Professor Ralph B. D'Agostino for guiding and supporting me over the years, and Mrs. LeiLanie D'Agostino for her kindness and consideration. I am also sincerely grateful to my intern advisor, Dr. Kerry Barker for giving me opportunities and guidance over the year. I whole heartedly thank my committee members, Dr. Sandeep Menon, Professor Gheorghe Doros and Professor Joseph Massaro for all their help and guidance.

I would like to acknowledge Dr. Steven Hua, Dr. Ronald Menton and for their inspiration and patience. I am grateful to all professors/instructors/staffs at Boston University and all previous schools I attended. I want to thank my fellow classmates and friends. Special thanks to Mo Pei and his family.

I am deeply indebted to my parents for their love, support and understanding on the way of my growth and pursuing dream. They teach me to be a good person.

Thank God for giving me life, and thank all my family members, my advisors, the full committee members, colleagues, friends to be part of my life.

STATISTICAL CONSIDERATIONS OF NONINFERIORITY, BIOEQUIVALENCE AND EQUIVALENCE TESTING IN BIOSIMILARS STUDIES

SIYAN XU

Boston University Graduate School of Arts and Sciences, 2013

Major Professor: Ralph B. D'Agostino Sr., Professor of Mathematics and Statistics and Professor of Epidemiology and Biostatistics

ABSTRACT

In recent years, the development of follow-on biological products (biosimilars) has received increasing attention. The dissertation covers statistical methods related to three topics of Non-inferiority (NI), Bioequivalence (BE) and Equivalence in demonstrating biosimilarity. For NI, one of the key requirements is constancy assumption, that is, the effect of reference treatment is the same in current NI trials as in historical superiority trials. However if a covariate interacts with the treatment arms, then changes in distribution of this covariate will result in violation of constancy assumption. We propose a modified covariate-adjustment fixed margin method, and recommend it based on its performance characteristics in comparison with other methods. Topic two is related to BE inference for log-normal distributed data. Two drugs are bioequivalent if the difference of a pharmacokinetics (PK) parameter of two products falls within prespecified margins. In the presence of unspecified variances, existing methods like two one-sided tests and Bayesian analysis in BE setting limit our knowledge on the extent that inference of BE is affected by the variability of the PK parameter. We

propose a likelihood approach that retains the unspecified variances in the model and partitions the entire likelihood function into two components: F-statistic function for variances and t-statistic function for difference of PK parameter. The advantage of the proposed method over existing methods is it helps identify range of variances where BE is more likely to be achieved. In the third topic, we extend the proposed likelihood method for Equivalence inference, where data is often normal distributed. In this part, we demonstrate an additional advantage of the proposed method over current analysis methods such as likelihood ratio test and Bayesian analysis in Equivalence setting. The proposed likelihood method produces results that are same or comparable to current analysis methods in general case when model parameters are independent. However it yields better results in special cases when model parameters are dependent, for example the ratio of variances is directly proportional to the ratio of means. Our research results suggest the proposed likelihood method serves a better alternative than the current analysis methods to address BE/Equivalence inference.

Table of Contents

List of Tables

List of Figures

1. Introduction

A small molecule drug is defined as a drug manufactured through a chemical process and can be structurally replicated. As such, the process to approve a generic version of a small molecule is well defined. On the other hand, biologics are often hundreds of times the size of a small molecule and a quite complex. These products, often proteins, are produced via living cells. The manufacturing of these products is never completely pure and there is natural variation between production lots. Therefore, regulatory authorization on biosimilars cannot follow the precedent set by small molecule generics.

According the FDA guidance document (2012) they define biosimilar “the biological product is highly similar to the reference product notwithstanding minor differences in clinically inactive components”, and that “there are no clinically meaningful differences between the biological product and the reference product in terms of the safety, purity, and potency of the product”. FDA recommends “stepwise approach” to develop biosimilar products. A stepwise approach includes “a comparison of the proposed product and the reference product with respect to structure, function, animal toxicity, human pharmacokinetics (PK) and pharmacodynamics (PD), clinical immunogenicity, and clinical safety and effectiveness.” Biopharmaceutical companies are heading towards

demonstrating a proposed therapeutic protein product is biosimilar to a licensed reference product after the Patient Protection and Affordable Care Act signed into law by President Barack Obama in 2010.

In biosimilar problems, the three most important terminologies are Bioequivalence (BE), Non-Inferiority (NI) and Equivalence. For small molecules, establishment of BE is a regulatory requirement prior to claiming that the proposed generic version is equivalent to the branded-name drug. For example, according to regulations applicable in the European Economic Area, two medicinal products are bioequivalent if they are pharmaceutically equivalent and if their bioavailabilities after administration in the same molar doses are similar to such a degree that their effects, with respect to both efficacy and safety, will be essentially the same. A similar wording of the BE definition can found in the US FDA guidelines (FDA guidance 2001). Thus the definitions of BE place the focus of statistical inference on whether the parameter of interest falls within the equivalence margins more rigorously than estimating the parameter itself. It is very likely to have similar regulatory understanding of BE for biologics, although has not been well established yet. For NI and equivalence trials we are looking at the endpoint of interest (often efficacy), so we are looking at the actual measure of interest (unlike BE where we are looking at a ‘biomarker’). NI is to show we are not worse than the reference product. Equivalence is not only looking at the lower bound (non-inferior) but also an upper bound (non-superior). The NI margins and equivalence margins can be derived in the same way.

My dissertation research arose from clinical trials in demonstrating biosimilarity. The thesis will cover three broad inter related ideas on testing for BE (used in Phase I, II for PK variables), NI and Equivalence (used in Phase III for efficacy endpoint).

1. Current NI methods re-evaluated
2. A likelihood method of evaluating BE
3. Extension of BE evaluation method to Equivalence

One of the key assumptions on the NI test is constancy assumption, that is, the effect of reference treatment is the same in current NI trials as in historical superiority trials. However, if a covariate interacts with the treatment arms, then changes in distribution of this covariate will likely result in violation of constancy assumption. In the first part of the dissertation, we propose four new NI methods and compare them with two existing methods to evaluate the change of background constancy assumption on the performance of these six methods. To achieve this goal, we study the impact of three elements: 1) Strength of covariate; 2) Degree of interaction between covariate and treatment and; 3) Differences in distribution of the covariate between historical and current trials have on both the type I error rate and power using three different measures of association: difference, log relative risk and log odds ratio.

Two drugs are bioequivalent if the ratio of a pharmacokinetic (PK) parameter of two products falls within equivalence margins. The distribution of PK parameters

is often assumed to be log-normal, therefore bioequivalence (BE) is usually assessed on the difference of logarithmically transformed PK parameters (δ). In the presence of unspecified variances, test procedures like two one-sided tests (TOST) use sample estimates for those variances; profile likelihood replaces them with restricted maximum likelihood estimators, while Bayesian models integrate them out in the posterior distribution. These methods limit our knowledge on the extent that inference about BE is affected by the variability of PK parameters. In the second part of the dissertation, we propose a likelihood approach that retains the unspecified variances in the model and partitions the entire likelihood function into two components: *F*-statistic function for variances and *t*-statistic function for δ. Demonstrated with published real life data, the proposed method not only produces results that are same as TOST and comparable to Bayesian method, but also helps identify ranges of variances which could make the determination of BE more achievable. Our findings manifest the advantages of the proposed method in making inference about the extent that BE is affected by the unspecified variances, which cannot be accomplished either by TOST or Bayesian method.

Equivalence trials aim to demonstrate that new and standard treatments are equivalent within pre-defined clinically relevant limits. We focus when inference of equivalence is made in terms of the ratio of two normal means. In the presence of unspecified variances, methods such as likelihood-ratio test use sample estimates for those variances; Bayesian models integrate them out in the

posterior distribution. These methods limit our knowledge on the extent that inference about equivalence is affected by variability of the parameter of interest. To account the presence of unspecified variances, in the third part, we extend the likelihood approach from log-normal data to normal data. The proposed method retains the unspecified variances in the model and partitions the full-likelihood function into two components: *F*-statistic function for variances and *t*-statistic function for the ratio of means. We show that, the proposed method can help identify ranges of variances where equivalence is more likely to be achieved. In addition, the proposed method can produce results that are same as the likelihood-ratio test and comparable to Bayesian analysis in general case when model parameters are independent. In a special case of dependent parameters, for example the ratio of two variances is directly proportional to the ratio of two means, the proposed method yields better results in inference about equivalence than the likelihood-ratio test which relies solely on the *t*-statistic or Bayesian analysis which integrates out the variances over some non-informative prior in the posterior distribution.

The aim of this dissertation is to develop advanced methodology/procedures in BE, NI and equivalence testing. We have written these three topics to methodology manuscripts and submitted to peer-review statistical journals, so each topic in the dissertation is self contained. The proposed work is the first in this specific area, and will help in developing effective ways to demonstrate whether new products are biosimilar to reference products.

2. Covariate effect on constancy assumption in noninferiority clinical trials (First Topic)

2.1. Introduction

Superiority test is the standard statistical test employed in the randomized clinical trial (RCT) [16, 36] with the aim to show that one treatment is superior to the other. However, given changes in patient populations, quality of life, and treatment options, it is becoming increasingly difficult to develop a clinically more efficacious new treatment. It is also often unethical to give patients placebo when an active treatment is available on the market. Therefore, pharmaceutical companies are now more interested in developing new treatments by showing that they are equivalent or non-inferior to an existing treatment. In these cases the focus has shifted from statistical significance to clinical meaningful difference and comparability. The Patient and Affordable Care Act of 2010 has also led to further influx of non-inferiority and equivalence clinical trials designs.

To avoid misuse of superiority test, non-inferior tests are more appropriate, in which significant results lead to conclusion of non-inferiority that is not due to sampling error [34, 50]. An overall improvement in patient care is taken into consideration in addition to establishment of non-inferiority for efficacy endpoint.

For example, if the new treatment also shows other benefits, such as cheaper price for the new drug, easier mode of administration (subcutaneous or oral instead of an intravenous administration), less toxicity and so on.

Unlike superiority test which directly demonstrates the efficacy of a test treatment, non-inferiority test is an indirect approach to demonstrate the efficacy of a test treatment. NI test combines current NI trial, which compares test treatment with reference treatment, with one or more historical superiority trials, which compared reference treatment with placebo [25, 28]. In order to do this, one of the key assumptions of the NI test is the constancy assumption, i.e., the effect of reference treatment is the same in current NI trial as in historical superiority trials [12, 13, 18, 23]. However, this assumption may not hold due to differences in the inclusion/exclusion criteria between current trial and historical trials, mode of administration, length of follow-up, design quality etc. When the constancy assumption is violated, standard statistical methods—fixed margin method and synthesis method may become invalid, as shown by Wang et al. [49].

If a covariate is associated with the effect size of the reference product, changes in distribution of this covariate will result in violation of the constancy assumption. Nie and Soon [35] proposed a covariate-adjustment regression model approach to assess the test treatment effect when population difference between the historical trial and the current NI trial causes constancy assumption violation.

However, they did not compare performance characteristics of proposed methods with standard methods. Moreover, adjustment of treatment effect was made only when the constancy assumption was violated. They did not take advantage of the fact that the reference treatment effect was estimable using the covariate mean of the new population based on the model-based regression approach. Reference treatment effect estimated using the current population is more accurate than using historical trial population for NI test, regardless of violation of the constancy assumption. Furthermore, the covariate effect was evaluated and estimated using historical data, but was not confirmed using the NI data.

In this topic, we propose modified covariate-adjustment fixed margin/synthesis methods and two-stage covariate-adjustment fixed margin/synthesis methods. For the modified covariate-adjustment fixed margin/synthesis methods, we estimate the effect size of reference product using covariate mean of current population. For the two-stage covariate-adjustment fixed margin/synthesis methods, the first stage is to assess whether the effect of covariate is important enough to be included in the model using current NI data, and the second stage is to use either the fixed margin/synthesis methods (if covariate not included) or modified covariate-adjustment fixed margin/synthesis methods (if covariate included) based on the results of the first stage. We will also study the impact of three elements: 1) Strength of covariate; 2) Degree of interaction between covariate and treatment; and 3) Differences in distribution of the covariate

between historical and current trials have on both the type I error rate and power using three different measures of association: difference, log relative risk and log odds ratio.

Note that the covariate-adjustment fixed margin/synthesis approach requires individual patient data (IPD) rather than aggregate data (AD) in historical trial Nie and Soon [35]. IPD are complete data information including each patient's assigned treatment arm, outcome and covariate measurements. IPD is the gold standard for meta-analysis to obtain historical estimates [8, 44]. If the investigators of both NI and the historical trials are the same, there should be no problem to access the raw data. However, if the investigators are different, then sharing data among sponsors/regulatory agencies are encouraged to design future NI trials.

For simplicity, we assume 1:1 balanced randomized clinical trials, with patient-level data available, and focus on factors whose impact on treatment difference can be quantified, for example population differences. We will address the question of how the changes of background distribution (i.e. violation of constancy assumption) will affect the performance of various NI methods, using the three different measures of association (difference, log relative risk and log odds ratio). Section 2.2 will describe the methods. Section 2.3 will give the details of the simulation set up and the results of various simulations. Section 2.4 will

summarize our conclusions on the performance evaluation based on type I error rate, power and some relevant sensitively analysis.

2.2. NI Test Problem and Methods

Suppose the outcome is a dichotomous efficacy variable. For convenience we will assume that a higher response rate is desirable (if the opposite is true one can always just multiple by negative one to get the desired direction). Let T, C, and P be response rates of test drug, active control drug and putative placebo in NI trial, respectively. Let C_0 and P_0 be response rates of active control drug and placebo in historical trial, respectively. Note that active control is the reference product. In NI trials there is no placebo arm, but sponsors and regulatory agencies are still usually interested in estimating what the effect size of test product over placebo would have been if the placebo had been included in the current NI trial. Since no direct comparison is available, we impute a placebo, called a putative placebo, from historical data to assist in this indirect comparison.

Figure 2.1. Plot for response rates of test drug (T), active control drug (C) and putative placebo (P) in NI trial.

Figure 2.1 shows the relationship among response rates of test drug (T), active control drug (C) and putative placebo (P) in NI trial. “0” is included to indicate direction.

Formulation of the hypothesis depends on the primary objectives. According to Blackwelder [6], the general objective of NI is to show the test treatment is not worse than the active control by a pre-specified margin, that is

$$H_0: \mathrm{C} - \mathrm{T} \geq \delta \quad \text{vs.} \quad H_1: \mathrm{C} - \mathrm{T} < \delta \tag{1}$$

C-T stands for “difference” between active control and test product, it could be difference, relative risk, odds ratio, log of relative risk or log of odds ratio (mathematic expression can be modified accordingly). The margin δ is the largest acceptable difference between the two products that would be considered not clinically meaningful and must be carefully defined in advance. Selection of the margin δ is extremely challenging, involving both careful clinical and statistical judgments [12, 13, 18, 23, 24, 27, 48].

The minimum standard for regulatory agencies’ approval of a test treatment is to show “superiority over placebo”. However, with missing placebo arm in active

controlled clinical trials, direct demonstration of such superiority is impossible. Furthermore, with reference products have already been marketed for years with positive risk-benefit profiles, “superiority over placebo” is a necessary but not sufficient condition to approve a test product. The goal is to choose a margin so that the test product retains some fraction of reference product efficacy (FDA, 1999). If all uncertainties are properly and comprehensively considered, ruling out the margin δ within an acceptable statistical error from the non inferiority trial can be helpful in also demonstrating superiority over placebo with great confidence.

Given an objective of retaining a portion of the active control treatment efficacy, equation (1) can be rewritten as equation (2):

$$H_0: \mathrm{C} - \mathrm{T} \geq (1-\lambda)(\mathrm{C} - \mathrm{P}) \text{ vs. } \quad H_1: \mathrm{C} - \mathrm{T} < (1-\lambda)(C-P) \tag{2}$$

where λ is the level of percent effect retention from previous active trial. A key question is what estimate to use for C-P? Does one use the point estimate or should one use a more conservative estimate? Regulatory guidance document suggests using a lower bound of a confidence interval and/or taking “discount” of any related uncertainties. Note that such estimate of effect size of reference product can be either point estimate or lower bound of some confidence interval, or taking “discount” of any related uncertainties [45]. Our discussions below are based on the retention test hypothesis (2).

2.2.1. Fixed margin method

This method rejects the null hypothesis H_0 in (2) and concludes that the test treatment is not inferior to the reference treatment if

$$\hat{C} - \hat{T} + z_{\alpha}\sigma_{TC} < (1-\lambda)\{\tilde{C}_0 - \tilde{P}_0 - z_{\alpha}\sigma_{PC0}\} \quad (3)$$

where z_{α} is the (1- α)th percentile of the standard normal distribution. $\tilde{C}_0$ and $\tilde{P}_0$ are estimators of C and P from historical trial, σ_{PC0} estimates the standard error of (C-P). Similarly, $\hat{C}$ and $\hat{T}$ are estimators of C and T from current NI trial, σ_{TC} estimates the standard error of (C-T).

The non-inferiority margin δ is a function (often taking the smaller value) of a statistical margin and a clinical margin, and must be predetermined before the NI trial starts. However, when conducting this NI analysis, the margin is treated as a fixed constant. The clinical margin comes from a comprehensive review of literature combined with clinical judgment. While a statistical margin often comes from an analysis of one or more historical superiority trials. In this paper we will assume that the statistical margin is smaller than the clinical margin, but we should bear in mind that choosing a fixed margin is always combining both the statistical and the clinical information.

If we use $\alpha = 0.025$, the fixed margin method is known as 95-95 method, where the first 95 means the 95% confidence interval obtained from the historical trial,

and the second 95 means the 95% confidence interval obtained from the non inferiority trial. The fixed margin method is also known as confidence interval method. This method is conservative due to the fact that the margin is determined by discounting the lower bound of $\tilde{C}_0 - \tilde{P}_0$. For further details of this method, please refer Hung et al., [23]; Hung et al., [24]; Hung et al., [25]; Wang et al., [47]; Wang and Hung, [48];Tsong et al., [46]; Wiens, [52]; Laster et al., [30] and references therein.

2.2.2. Synthesis method

Synthesis method rejects the null hypothesis H_0 in (2) if

$$Z_{pv} = \frac{\hat{C} - \hat{T} - (1-\lambda)\{\tilde{C}_0 - \tilde{P}_0\}}{\sqrt{\sigma_{TC}^2 + (1-\lambda)^2 \sigma_{PC0}^2}} < -z_\alpha \quad (4)$$

where z_α is the (1-α)th percentile of the standard normal distribution. This method is also known as preservation test method.

This method synthesizes two sources of data—current NI study (estimates the effect of test treatment over reference treatment $\hat{C} - \hat{T}$) and historical studies (estimated the effect of reference treatment over placebo $\tilde{C}_0 - \tilde{P}_0$). Here there is no need to pre-specify a non-inferiority margin. $\tilde{C}_0 - \tilde{P}_0$ is not a fixed constant but an additional random variable similar to the random variable $\hat{C} - \hat{T}$. Because the

test statistic is constructed by dividing a relevant combination of the estimate of test treatment effect and active control effect by the combined standard error, this method purports to control type I error rate under the constancy assumption. Since the variance of fixed margin method is larger than the variance of synthesis method, the fixed margin method is more conservative and hence the nominal type I error rate is often much less than the desired target type I error rate (even when the constancy assumption holds). The synthesis method has been shown to be very sensitive to the constancy assumption. Violation of the constancy assumption will result in the synthesis method being liberal. However, since the fixed margin method uses the worst limit of the confidence interval of the active control effect to define the margin, it may be less sensitive to the violations of the constancy condition. The degree of which depends on the level of preservation specified in the defining the margin. For more details of this method, please refer Hung et al., [23]; Wang et al., [49]; Wang and Hung, [48]; Snapinn, [45]; Hasselblad and Kong, [19]; ICH, [26]; Holmgren, [21]; Wang and Hung, [47]; Rothmann et al., [37] and references therein.

2.2.3. Modified covariate-adjustment fixed margin/synthesis methods

Nie and Soon [35] proposed a covariate-adjustment regression model approach to assess the test treatment effect when population difference between the

historical trial and the NI trial causes constancy assumption violation. Applying this approach, they were not only able to measure the impact of population difference on the degree of constancy assumption violation, but were also able to re-estimate the effect size of the active control when constancy assumption did not hold. For historical superiority trial, the following logistic model is used to describe association between dichotomous outcome, treatment arm and covariate:

$$logit\big(P(Y=1)\big) = \alpha + \beta * Trt + \beta_1 * X + \gamma_1 * X * Trt \quad (5)$$

, where Y is the primary efficacy endpoint (outcome); Trt is the treatment arm, 1 for placebo and 0 for reference product; X is a binary covariate, coded as 1 or 0, and $X * Trt$ is the interaction term for treatment and covariate. This basic model can easily extend to a more complicated one, such as including two or more covariates, and some of them may be independent from treatment arm. Individual patient data is needed for this modeling.

Based on the above model, the effect of active control is $logit(C_0) - logit(P_0) = -\beta - \gamma_1 \bar{X}_{.H}$, where $\bar{X}_{.H}$ is the mean of covariate X for historical trial population. NI trial population can also be used to estimate the mean of covariate X. Constancy assumption holds if the difference between such effects using two populations is small. To estimate the effect size of active control, current NI trial population is preferable if it is available; otherwise, the historical population can be used to estimate this. We modified Nie and Soon's method by always using covariate

mean of current NI trial population, because it will give us a more accurate estimator of active control efficacy in NI test. We further apply this active control effect to 2.1 and 2.2 for modified covariate-adjustment fixed margin method and modified covariate-adjustment synthesis method.

a) Modified covariate-adjustment fixed margin method (CovFM)

Replace the estimator of $\tilde{C}_0 - \tilde{P}_0$ and its standard error σ_{PC0} in (3) using parameters estimated in (5) and covariate mean of current NI trial population.

b) Modified covariate-adjustment synthesis method (CovSyn)

Replace the estimator of $\tilde{C}_0 - \tilde{P}_0$ and its standard error σ_{PC0} in (4) using parameters estimated in (5) and covariate mean of current NI trial population.

2.2.4. Two-stage covariate-adjustment fixed margin/synthesis methods

Often one does not know the true effect of a covariate. In section 2.3 modified covariate-adjustment method, the covariate effect was confirmed via model selection using historical data, and did not use current NI data. In the proposed two-stage method, the first stage is to confirm that the effect of covariate is sufficiently present in the current study to such an extent that it should be incorporated in the current analysis; the second stage is to use either the covariate method or unadjusted method based on the results of stage 1.

Stage 1: Test of covariate

For a given NI trial, suppose RR_1 and RR_2 are response rates of subgroups defined by a binomial covariate. To confirm the effect of covariate in the first stage, we propose the following hypothesis test using threshold 0.15 instead of commonly used 0.05:

$$H_0: RR_1 = RR_2 \text{ vs. } H_a: RR_1 \neq RR_2 \quad (6)$$

Stage 2: Test of NI

In stage 2, use fixed margin method or synthesis method if the null hypothesis in stage 1 is not rejected, i.e., the effect of covariate is not worth putting into the model; use modified covariate-adjustment fixed margin method or modified covariate-adjustment synthesis method if null hypothesis can be rejected. We will denote the above by two-stage covariate-adjustment fixed margin method if the fixed margin method is used in stage 2 and two-stage covariate-adjustment synthesis method if the synthesis method is used in stage 2.

2.2.5. Paradigms of the two NI Methods: Fixed margin vs. Synthesis

All methods used here involve using either the fixed margin approach or the synthesis approach. Note that these two approaches operate under different sets

of paradigms. The fixed margin method uses conventional clinical trial inference, which is based on a comparison of two randomized groups within a study. However, synthesis method uses across-study inference, which is based on synthesizing analysis from multiple studies.

2.3. Simulation

2.3.1. Motivating example

HER2 stands for Human Epidermal Growth Factor Receptor 2, is a member of the HER family. This family has four distinct receptors, the Epidermal Growth Factor Receptor (EGFR), HER2, HER3 and HER4. HER3/HER2 status is considered a potential strong covariate for Her2+ cancer as it has been shown in some studies that HER3+/HER2+ subjects tend to respond better than HER3-/HER2+ patients.

Consider a hypothetical but realistic scenario. Suppose in a historical trial, 50% of enrolled subjects are HER3+/HER2+, and 50% are HER3-/HER2+. In active controlled trial percent of enrolled subjects with HER3+/HER2+ may vary from

10% to 70%. We will consider four scenarios. The response rates of these four scenarios are as follows:

1) Response rates of taking placebo and active control are 0.18, 0.50 for HER3+/HER2+ patient; response rates of taking placebo and active control are 0.18, 0.50 for HER3-/HER2+ patient.

2) Response rates of taking placebo and active control are 0.28, 0.60 for HER3+/HER2+ patient; response rates of taking placebo and active control are 0.13, 0.45 for HER3-.HER2+ patient.

3) Response rates of taking placebo and active control are 0.40, 0.60 for HER3+/HER2+ patient; response rates of taking placebo and active control are 0.13, 0.45 for HER3-/HER2+ patient.

4) Response rates of taking placebo and active control are 0.18, 0.78 for HER3+/HER2+ patient; response rates of taking placebo and active control are 0.31, 0.38 for HER3-/HER2+ patient.

Figure 2.2. Plot for response rates, HER3/HER2 status and treatment arms.

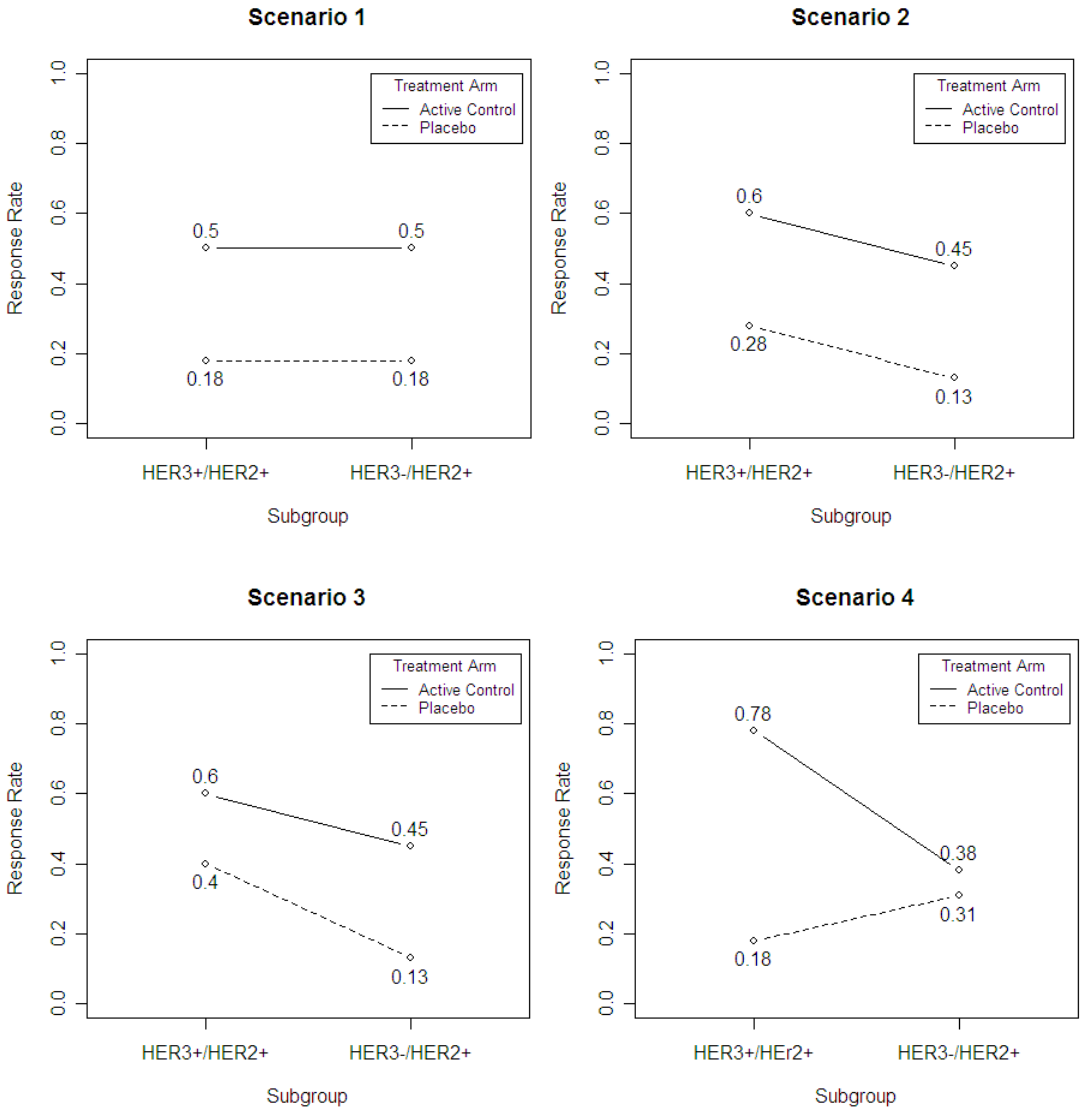


Figure 2.2 describes the relationship among HER3/HER2 status, response rates and treatment arms. To determine whether HER3/HER2 status is a covariate, and to what extent it interacts with treatment arm, Table 2.1 gives numerical information.

In scenario 1, HER3/HER2 status is not associated with response rate, so it is not a covariate. While in scenario 2 to 4, HER3/HER2 status is associated with response rate, and thus it is a covariate. Moreover, in scenario 1, HER3/HER2 status does not interact with treatment arm. In scenario 2, 3 and 4, covariate has mild, moderate and high interaction with treatment arm.

If covariate is associated with the effect size of reference product, which means covariate interacts with treatment arm, then changes in percent of subjects with HER3/HER2 status will result in the constancy assumption being violated. Otherwise, the change in percent of subjects with HER3/HER2 status will not result in constancy assumption violation.

Table 2.1. Table for association between HER3/HER2 status and outcome (column 4), and association between HER3/HER2 status and treatment arm (column 7).

Scenario	Response Rate (1:1 randomization)		Fold change of response rate	Active Control vs. Placebo Odds Ratio		Fold change of Odds Ratio
	HER3+/HER2+	HER3-/HER2+	(HER3+/HER2+ vs. HER3-/ER2+)	HER3+ /HER2+	HER3-/HER2+	(HER3+/HER2+ vs. HER3-/HER2+)
1	0.34	0.34	1	4.556	4.556	1
2	0.44	0.29	1.51	3.857	5.476	0.704
3	0.50	0.29	1.724	2.25	5.476	0.411
4	0.48	0.345	1.39	16.152	1.364	11.842

2.3.2. Simulation details

Simulations were performed in order to evaluate and compare the operating characteristics of the six methods described in the previous section. The six methods are:

a) Fixed Margin method (FM);

b) Synthesis Method (Syn);

c) Modified covariate-adjustment fixed margin method (CovFM);

d) Modified covariate-adjustment synthesis method (CovSyn);

e) Two-stage covariate-adjustment fixed margin method (2sFM) and;

f) Two-stage covariate-adjustment synthesis method (2sSyn).

In addition we will examine three different measures of association for each of the six methods. These three measures are: difference, log relative risk and log odds ratio. The log scale of relative risk and odds ratio were used because of the following reasons: 1) log(A/B)=log(A)-log(B)=-log(B/A). In this case, one can interpret the difference between A and B as the inverse of the difference between B and A; 2) log scale data can be approximated with Gaussian distribution theorem.

For each method and measure of association combination, 10,000 simulations were run to:

1) Compare empirical type I error using target alpha level of 2.5%;

2) Compare power using target power level of 80% at the approximate the same empirical alpha level;

3) Perform extensive sensitivity analyses to assess the performance of each method in lieu of violation of the constancy assumption. In particular to compare the degree of departure of the constancy assumption needed to double or triple the empirical type I error rate. Hence, this evaluation will evaluate the robustness of each method to deviations from the assumption used in the analysis.

Since the null hypothesis is composite, we will measure empirical type I error rate when equality in null hypothesis of NI testing holds. Power is to be compared when the constancy assumption holds, i.e. 50% HER3+/HER2+ in both historical trial and current NI trial. If the constancy assumption is seriously violated, the FM and Syn methods will appear to have good power. However, that is misleading since the type I error rate are inflated. To compare power, assign the response rate of test product 10 possible values, with maximum equals to the response rate of reference product, and minimum equals to the response rate of test product used in computing type I error rate. When examining the power plots, the type I error rates when percent of HER3+/HER2+ is 50% can be also found.

Simulation steps:

(i) Use underlying true parameters to generate 1000 per arm (2 arms) in historical trial, 300 per arm (3 arms) in NI trial.

(ii) Apply each of the six methods to reject/accept H_0.

(iii) Repeat (i) and (ii) 10,000 times, to get proportion of rejection for all six methods and three measures of association.

2.3.3. Simulation results

For each scenario and measure of association combination, results are presented in a panel of four plots in one figure. Row one presents the type I error rate while row two presents the power. Column one gives results over broad range of all values while column two zooms in on values of interest. The upper left plot has 7 groups of points for each of the six methods (corresponding to percent of HER3+/HER2+ in NI ranging from 10% to 70% by 10% increments). Note that 50% is the same percent of HER3+/HER2+ in historical trial, so points around it (40%, 50% and 60% of HER3+/HER2+ in NI trial) are zoomed, which are shown on the upper right. The bottom left plot is 10 groups of points for each of the six methods (corresponding to 10 points between ranges of possible test product effect). Note that the 10^{th} group of points is the power when test product

effect equals to that computed in type I error rate. So they are actually type I error rate when the percentage of HER3+/HER2+ is 50% in NI trial. Closer look at power around the type I error rate are shown on the bottom right plot.

This section will only show the results for scenario 2 (mild interaction) only since the conclusions drawn from scenario 2 also hold for the other three scenarios. To see simulation results for scenario 1, 3 and 4, please refer to Appendix A.

Scenario 2:

Response rates of placebo and active control treatment groups are 0.28, 0.60 for HER3+/HER2+ patients; response rates of placebo and active control treatment groups are 0.13, 0.45 for HER3-.HER2+ patients.

In this scenario, HER3/HER2 status is a covariate. This covariate does not interact with treatment arm using metric difference, but mildly interacts with treatment arm using metrics log relative risk (LRR) and log odds ratio (LOR). Thus change of percent of HER3+/HER2+ in NI will not cause constancy assumption violation for metric difference, but will result in constancy assumption violation for LRR and LOR.

Plots for type I error rate and power, using three measures of association (difference, log relative risk and log odds ratio) are shown in Figure 2.3, Figure 2.4 and Figure 2.5.

Figure 2.3. Performance Characteristics Comparisons--Measure of Association: Difference.

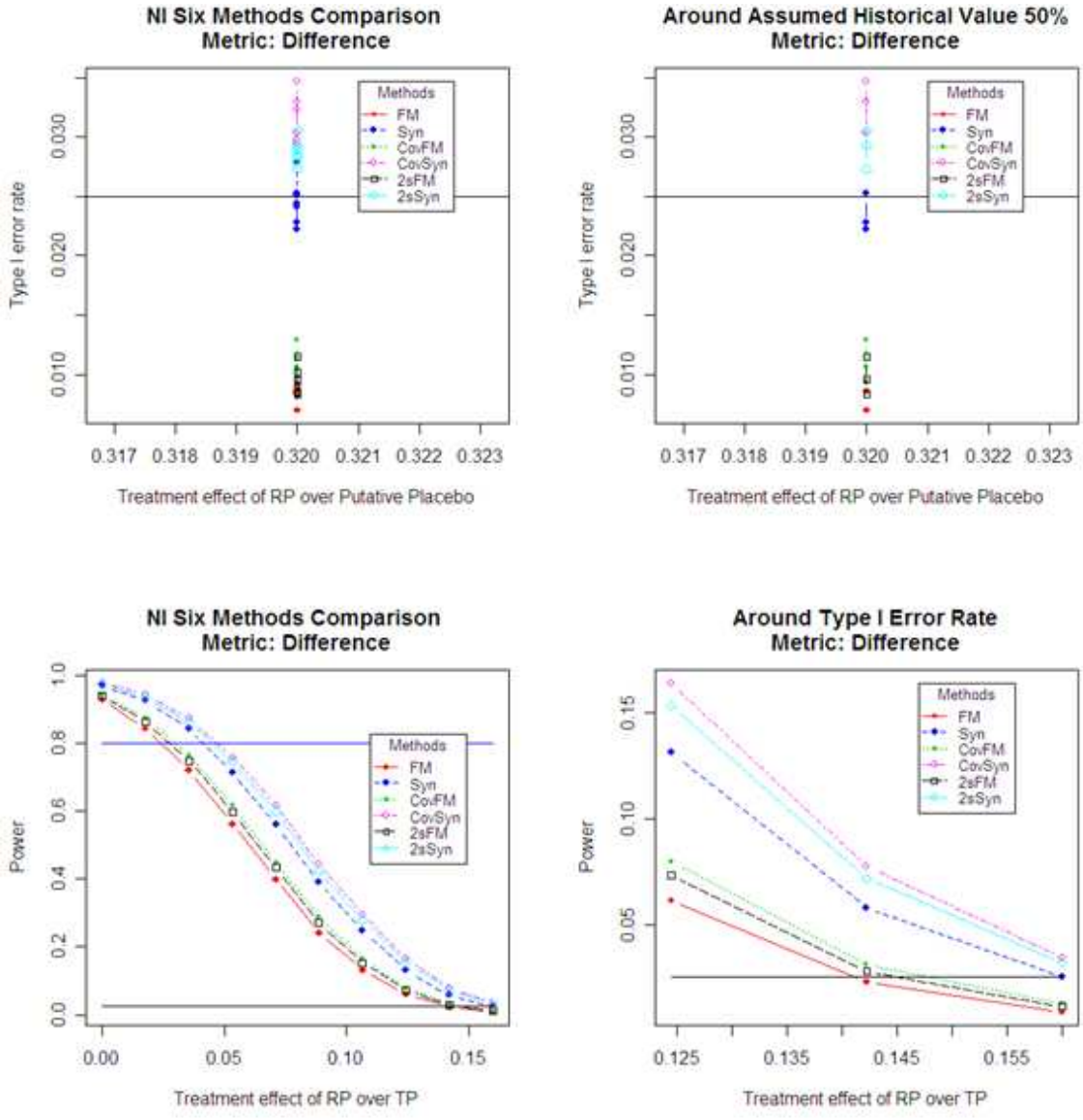


Figure 2.4. Performance Characteristics Comparisons--Measure of Association: Log Relative Risk.

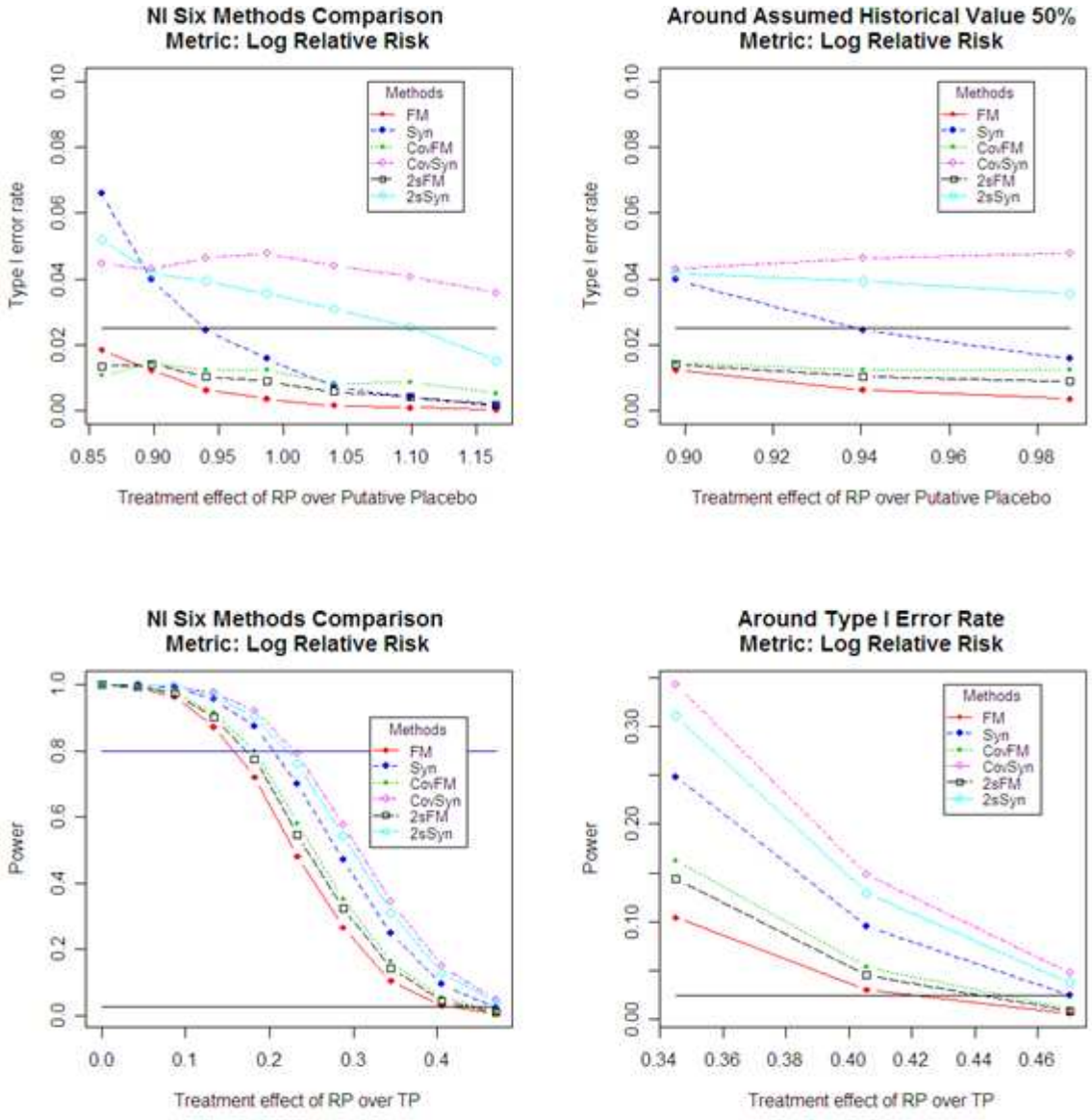
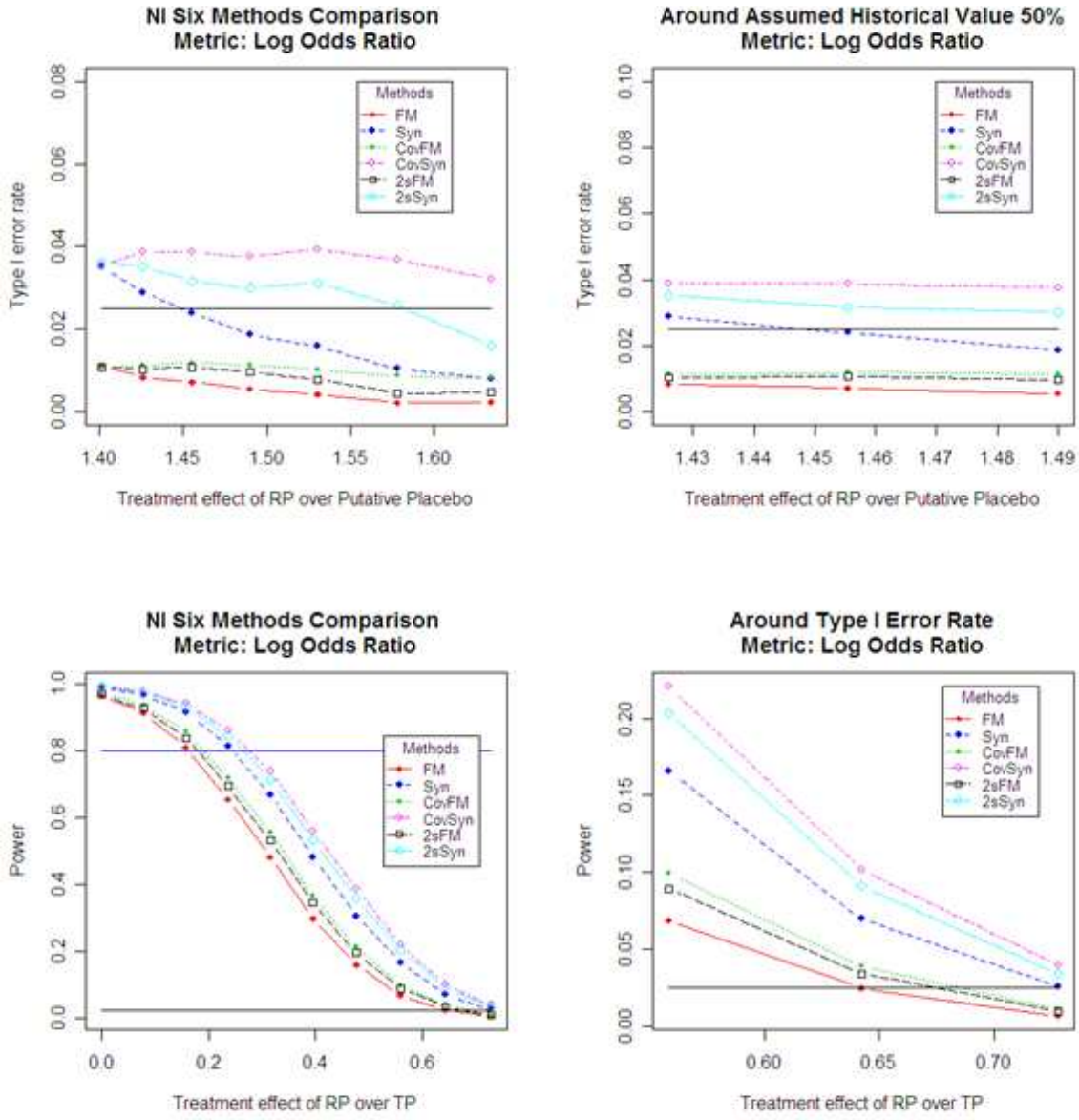


Figure 2.5. Performance Characteristics Comparisons--Measure of Association: Log Odds Ratio.



When using LRR and LOR, increasing the percent of HER3+/HER2+ will decrease the effect of reference product over putative placebo, and also

decrease the effect of reference product over test product. Thus the 10% point corresponding to the biggest “Treatment effect of RP over Putative Placebo” in type I error plots in Figure 2.4 and 2.5.

Type I error rates of FM, CovFM and s2FM are all under 0.025, while Syn, CovSyn and 2sSyn have higher type I error rates and can be greater than 0.025 at some points. As can be seen in Figure 2.4 and 2.5, even slight departure from the constancy assumption (50%) causes the type I error rate inflated. Looking at upper left plots in Figures 2.3, 2.4 and 2.5: type I error rates of CovSym and 2sSyn are larger than 0.025 when constancy assumption holds (note this is the case where we adjusted the covariate when it is unnecessary). As can be seen in upper right plots of Figures 2.4 and 2.5, the synthesis method is extremely sensitive to constancy assumption such that any small departure from constancy assumption, and/or any small estimation variation inflates the type I error rate. The fixed margin method is more stable in controlling the type I error rate. CovFM and 2sFM improve the conservativeness of FM, but can still control type I error rate at target value 0.025.

Note the reason that we only see type I error rate inflated in one direction is because the bigger the relative effect of reference product over placebo, the bigger is the relative effect of reference product over test product. Thus it will make it harder for us to reject the null hypothesis of NI.

Figures 2.3, 2.4 and 2.5 Power plots show that CovFM has higher power than both 2sFM and FM. CovFM has similar operating characteristics across all scenarios, even when there is no interaction effect (Figure 2.3). As expected 2sFM has larger power than FM since 2sFM rejects the null hypothesis in (6) when there is a strong observed covariate effect and therefore will use CovFM, which is a model-based regression approach. However 2sFM has less power than CovFM because of potential type II error associated with testing hypothesis (6), i.e. fail to reject the null hypothesis in (6) even though the covariate effect is warranted.

Table 2.2 summarizes the type I error rates of the six NI methods, for different percentage of HER3+/HER2+ (10% to 70% by 10% increment) using three different measures of associations (difference, LRR, and LOR). Bold rows are percents of HER3+/HER2+ 40%, 50% and 60%. Table 2.3 summarizes the power of the six NI methods, when 50% of HER3+/HER2+ population in NI trial.

Table 2.2. Scenario 2-- Type I Error Rate

Metric	Perc_HER3+/HER2+	P	T	C	FM	Syn	CovFM	CovSyn	2sFM	2sSyn
Difference	0.1	0.145	0.305	0.465	0.0085	0.0279	0.0082	0.0295	0.0085	0.0283
	0.2	0.16	0.32	0.48	0.0084	0.0251	0.0087	0.0297	0.0084	0.0278
	0.3	0.175	0.335	0.495	0.0083	0.0241	0.0105	0.0303	0.0095	0.0285
	0.4	0.19	0.35	0.51	**0.007**	**0.0228**	**0.0106**	**0.033**	**0.0096**	**0.0293**
	0.5	0.205	0.365	0.525	**0.0086**	**0.0253**	**0.0129**	**0.0347**	**0.0115**	**0.0306**
	0.6	0.22	0.38	0.54	**0.007**	**0.0222**	**0.0093**	**0.0303**	**0.0083**	**0.0273**
	0.7	0.235	0.395	0.555	0.009	0.0244	0.0114	0.0323	0.0101	0.029
LRR	0.1	0.145	0.259663	0.465	3.00E-04	0.0015	0.0053	0.0357	0.0019	0.0152
	0.2	0.16	0.277128	0.48	0.001	0.0042	0.0087	0.0407	0.0042	0.0255
	0.3	0.175	0.294321	0.495	0.0016	0.0075	0.0083	0.044	0.0057	0.031
	0.4	0.19	0.311288	0.51	**0.0035**	**0.0158**	**0.0125**	**0.0478**	**0.0089**	**0.0356**
	0.5	0.205	0.328062	0.525	**0.0063**	**0.0245**	**0.0124**	**0.0464**	**0.0104**	**0.0394**
	0.6	0.22	0.344674	0.54	**0.0123**	**0.0397**	**0.0145**	**0.043**	**0.014**	**0.0418**
	0.7	0.235	0.361144	0.555	0.0185	0.0659	0.0107	0.0446	0.0135	0.0519
LOR	0.1	0.145	0.277419	0.465	0.0023	0.008	0.0083	0.0322	0.0048	0.016
	0.2	0.16	0.295434	0.48	0.0021	0.0104	0.0087	0.0369	0.0045	0.0258
	0.3	0.175	0.313179	0.495	0.0042	0.0159	0.0102	0.0392	0.0078	0.0311
	0.4	0.19	0.330704	0.51	**0.0055**	**0.0187**	**0.0113**	**0.0376**	**0.0097**	**0.0301**
	0.5	0.205	0.348049	0.525	**0.0071**	**0.0239**	**0.0119**	**0.0387**	**0.0107**	**0.0316**
	0.6	0.22	0.365247	0.54	**0.0083**	**0.0289**	**0.0112**	**0.0388**	**0.0103**	**0.0352**
	0.7	0.235	0.382323	0.555	0.0112	0.0351	0.0106	0.0356	0.0107	0.0362

Table 2.3. Scenario 2--Power

Metric	P	T	C	FM	Syn	CovFM	CovSyn	2sFM	2sSyn
Difference	0.205	0.507222	0.525	0.8424	0.9244	0.8742	0.9433	0.8628	0.9353
LRR	0.205	0.503118	0.525	0.9898	0.9978	0.9951	0.9989	0.9938	0.9986
LOR	0.205	0.505339	0.525	0.9132	0.968	0.937	0.9801	0.928	0.9751

2.4. Conclusions and discussion

Both FM and Syn have inflated type I error rate when constancy assumption is seriously violated (scenario 4 as shown in Appendix A), Syn is the more liberal of the two. Syn is also very sensitive to violations of the constancy assumption. Small departure from constancy assumption, and/or small estimation variation could inflate the type I error rate. While not as severe as Syn, both the CovSyn and 2sSyn methods also have inflated type I error rates due to sensitivity of Syn.

FM has empirical type I error rate over the desired value when constancy assumption is seriously violated, although it is not as sensitive as the Syn method. This can be explained mathematically: from the formula $\hat{C} - \hat{T} + z_{0.025}\sigma_{TC} < (1-\eta)\{\tilde{C}_0 - \tilde{P}_0 - z_{0.025}\sigma_{PC0}\}$, if constancy is seriously violated, $\tilde{C}_0 - \tilde{P}_0$ will become smaller, and so it is easier to reject the null hypothesis, resulting in inflated type I error rate. When constancy assumption does not hold due to covariate effect, adjusting for this effect will result in significantly less type I error. This is why the CovFM and 2sFM methods perform better than FM method in controlling the type I error rate. 2sFM may falsely accept the null hypothesis in (6). Method CovFM performs consistently well across all scenarios: CovFM method has almost the same power as 2sFM in scenario 1 (no covariate no interaction) and scenario 4 (covariate and high interaction), while CovFM has

larger power than 2sFM in scenario 2 (mild interaction) and scenario 3 (moderate interaction).

Based on these results we recommend using CovFM (modified covariate-adjustment fixed margin method) since it controls the type I error rate and has satisfactory power for any of the three measures of association.

The most logical use for this methodology is when a company wants to show NI for a new formulation, for example, a new Sub Q dose for the company's existing IV drug. Thus the company has individual patient data. It is our hope that with the emphasis on comparative effectiveness initiatively found in the Affordable Care Act passed by the US Congress on March 23, 2010 that sharing individual patient data among sponsors/agencies will be possible in the future.

3. Inference of Bioequivalence for log-normal distributed data with unspecified variances (Second Topic)

3.1. Introduction

Establishment of bioequivalence (BE) is a regulatory requirement needed prior to claiming that a proposed generic version (test product) of a drug is equivalent to the branded-name drug (reference product). For example, U.S. FDA guidance document [17] recommends that 90% confidence interval (CI) values for the ratio of the relative means for AUC0-t, AUC0-∞, and Cmax of the new and reference drugs should fall between 0.80 and 1.25. These limits, set by regulations applicable in both U.S. and the European Economic Area, are acceptance criteria to assure that the difference between the test and reference products is no more than clinical relevance. AUC and Cmax are two pharmacokinetic (PK) parameters that represent area under the concentration-time curve and maximum concentration of a drug. The usual BE definition places the focus of statistical inference on whether the parameter of interest falls within the equivalence margins even more rigorously than estimating the parameter itself. Since the distribution of PK data is often assumed to be log-normal, BE is assessed on the difference in a logarithm-transformed PK parameter.

Statistical methods in establishing BE are readily available in literature including testing procedures developed by Metzler [34], Kirkwood [29], Westlake [3], Anderson and Hauck [1], Locke [32], Schuirmann [43], Berger and Hsu [53], among others. A comparison of operating characteristics in these testing procedures can be found in a review paper by Choi *et al.* [11]. The above methods are based on frequentist theory. One short fall of the frequentist theory is it attempts to use the same probabilities to measure both the chance of errors and the strength of evidence (Royall, [38, 39]). In [11], the authors proposed the use of a likelihood method by visually analyzing a standardized likelihood plot [38, 39]. This has generated interest in whether statistical evidence using a likelihood method can be used as an alternative or supplement to the usual practice of decision making based on testing procedures in BE setting.

When inference about BE based on the difference (δ) is of interest and the true variances (σ^2) is unknown, the traditional solutions involve replacing or removing σ^2 out of the analysis model. For instance, the hypothesis testing by the two one-sided tests (TOST) uses sample estimates for the σ^2, Bayesian method handles this by integrating out σ^2 over some non-informative prior. Since σ^2 is either replaced or removed, it raises a concern that these current analysis methods conceal the extent to which inference about δ is affected by σ^2 [7].

In this topic, we propose a likelihood approach that retains the unspecified variances in the model via three parameters: the ratio of two variances (γ^2), the

variance of the reference drug σ_R^2, and correlation of the two drugs (ρ) for cross-over studies only. “Unspecified” means those parameters are included in the model and allowed to vary over a range of values, to show the robustness of BE inference conditional on different values. We expand the likelihood approach for the ratio of two normal means by Diaz-Frances and Sprott [14] to address average bioequivalence of log-normal data by expressing the entire likelihood function in the product of two components: a function of F-statistic for variances and the a function of the t-statistic for δ. The proposed method helps identify ranges of the unspecified variances where BE is more likely to be achieved, which cannot be accomplished by methods that replace or remove these parameters. The proposed method addresses both cross-over and parallel designs.

The likelihood approach is specifically chosen for its focus on showing strength of evidence using data, which is consistent with our aim on generating evidence and measuring the extent that BE is affected by the unspecified variances. On the other hand, testing procedures such as TOST are only interested in whether the data support the nonequivalence or the equivalence hypothesis. Bayesian analysis presents a posterior probability distribution of δ, which is determined by the prior probability distribution of δ and likelihood function (probability model for the observed data). It focuses on inference in δ changing from prior, through observed data (via likelihood function) to posterior. We derived a Bayesian probability distribution using the method by Rubio and Perez-Elizalda [40] to

address our specific problem. It can be used to describe the probability that δ falls within the BE margins. The proposed method is compared with TOST and Bayesian method using published data.

The remainder of this whole part is organized as follows. Section 3.2 describes the proposed likelihood method. Section 3.3 presents the Bayesian model applicable to our specific question. The impact on BE by the unspecified variances is described in section 3.4. Comparison of different methods will be shown in section 3.5. Last section 3.6 is summary.

3.2. The Likelihood Method

3.2.1. Crossover Design

Suppose two variables $x_i = \log(x_i^*)$ and $y_i = \log(y_i^*), i = 1, \dots, n,$ that represent log-transformed individual observations of a PK parameter (AUC or Cmax) for the two drugs, follow bivariate normal distribution (BVN):

$$\begin{pmatrix} x_i = \log\left(x_i^*\right) \\ y_i = \log\left(y_i^*\right) \end{pmatrix} \sim BVN\left(\begin{pmatrix} \delta + \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \gamma^2\sigma^2 & \rho\gamma\sigma^2 \\ \rho\gamma\sigma^2 & \sigma^2 \end{pmatrix}\right) \quad (1)$$

For simplicity, we focus on log(AUC) (log(AUC_T), log(AUC_R), *T*=test, *R*=reference) throughout this section, the implication of log(Cmax) follows similarly. Parameters in (1) represent $\delta = E(\log(AUC_T)) - E(\log(AUC_R))$, $\mu = E(\log(AUC_R))$ and γ^2 the ratio of two variances ($\gamma^2 = \sigma_T^2/\sigma_R^2$) and $\sigma^2 = \sigma_R^2$ (Note log() means general logarithm, ln() means natural logarithm). The correlation coefficient ρ is to address cross-over designs where subjects start with one drug and then are switched to the other drug usually after a wash-out period, as such x_i and y_i are observations from the same subject, while in parallel designs ρ=0. Model (1) assumes there is no sequence or period effect.

The equivalence hypotheses with the pre-specified margins of θ_L=ln(0.8)≈-0.223 (natural logarithm) and θ_U=ln(1.25)≈0.223, can be stated as

$$H_{01}: \delta \le \theta_L \text{ or } H_{02}: \delta \ge \theta_U$$

$$H_{a1}: \delta > \theta_L \text{ and } H_{a2}: \delta < \theta_U \quad (2)$$

Schuirmann's TOST [43] can be used to test the hypothesis in (2). By rejecting the pair of null hypotheses H_{01} and H_{02}, equivalence can be declared if and only if:

$$T_L > t_{1-\alpha,n-1} \text{ and } T_U < -t_{1-\alpha,n-1}$$

where $T_L = \frac{(\bar{x}-\bar{y}-\theta_L)}{\hat{\sigma}_{(\bar{x}-\bar{y})}}$, $T_U = \frac{(\bar{x}-\bar{y}-\theta_U)}{\hat{\sigma}_{(\bar{x}-\bar{y})}}$, $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, and $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$, $t_{1-\alpha,n-1}$ is the (1 – α)-th percentile of the Student's *t*-distribution with n-1 degrees of freedom.

The likelihood approach focuses on showing the statistical evidence instead of making decision to choose between equivalence and nonequivalence. The likelihood function based on model (1) is in the form of

$$L(\delta, \mu, \rho, \gamma, \sigma | X, Y) \propto \frac{1}{\gamma^n \sigma^{2n} (1-\rho^2)^{n/2}} exp\left(-\frac{1}{2\sigma^2(1-\rho^2)} A\right) \quad (3)$$

where $A = \frac{S_x}{\gamma^2} + S_y + \frac{n(\bar{x}-\delta-\mu)^2}{\gamma^2} + n(\bar{y}-\mu)^2 - \frac{2\rho}{\gamma}\sum_{i=1}^{n}(x_i - \delta - \mu)(y_i - \mu)$

$S_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$ and $S_y = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

Without the knowledge of true μ and σ, a likelihood function of δ, γ and ρ can be constructed by replacing μ in (3) with its restricted maximum likelihood estimator (rMLE) $\hat{\mu}_{rM}(\delta, \gamma, \rho) = \frac{(1-\rho\gamma)\bar{x} + \gamma(\gamma-\rho)\bar{y} - (1-\rho\gamma)\delta}{(1-2\rho\gamma+\gamma^2)}$, followed by replacing σ^2 with its rMLE $\hat{\sigma}^2_{rM} = \frac{B}{2n(1-\rho^2)}$ where *B*=$\left[\frac{S_x}{\gamma^2} + S_y + \frac{n(\bar{x}-\bar{y}-\delta)^2(1-\rho^2)}{(1-2\rho\gamma+\gamma^2)} - \frac{2\rho}{\gamma}S_{xy}\right]$ and $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$. That is

$$L_{max}(\delta, \rho, \gamma | X, Y) \propto \frac{(1-\rho^2)^{\frac{n}{2}}}{\gamma^n}\left[\frac{S_x}{\gamma^2} + S_y + \frac{n(\bar{x}-\bar{y}-\delta)^2(1-\rho^2)}{(1-2\rho\gamma+\gamma^2)} - \frac{2\rho}{\gamma}S_{xy}\right]^{-n} \quad (4)$$

Following some algebra, the function in (4) can be subsequently partitioned into two parts:

$$L_{max}(\delta, \rho, \gamma | X, Y) = L_F(\rho, \gamma) \times L_t(\delta | \rho, \gamma) \quad (5)$$

where

$$L_F(\rho, \gamma) \propto \left(\frac{1}{FS_y\gamma\sqrt{1-\rho^2}}\right)^n \quad (6)$$

$$L_t(\delta | \rho, \gamma) \propto \left(1 + \frac{t^2}{(2n-2)}\right)^{-n} \quad (7)$$

$F = \frac{\frac{S_x}{\gamma^2} + S_y - \frac{2\rho}{\gamma} S_{xy}}{2S_y(1-\rho^2)}$, $t^2 = \left(\frac{\bar{x} - \bar{y} - \delta}{\hat{\sigma}_{(\bar{x}-\bar{y})}}\right)^2 = \frac{(\bar{x} - \bar{y} - \delta)^2}{\frac{(1 - 2\rho\gamma + \gamma^2)}{n}\hat{\sigma}^2}$, and $\hat{\sigma}^2 = \frac{\left(\frac{S_x}{\gamma^2} + S_y - \frac{2\rho}{\gamma} S_{xy}\right)}{2(n-1)(1-\rho^2)}$.

Derivation of (3)-(7) is outlined in Appendix A. Taking $\delta = \theta_L$ and $\delta = \theta_U$ the corresponding t in (7) equal to T_L and T_U respectively. This means $L_t(\delta | \rho, \gamma)$ in (7) can produce the results of TOST and more importantly the likelihood function in (5) which is the product of (6) and (7), will provide no less information than TOST alone on the inference about BE.

3.2.2. Parallel Design

In a parallel design, the distribution of $x_i, i = 1, \ldots, n$, and $y_j, j = 1, \ldots, m$ can be referred to (1) with $\rho = 0$. The corresponding maximum likelihood function can be

obtained by replacing μ and σ^2 by their rMLEs $\hat{\mu}_{rM}(\delta,\gamma) = \frac{n(\bar{x}-\delta)+m\gamma^2\bar{y}}{n+m\gamma^2}$ and $\hat{\sigma}^2_{rM} = \frac{B}{n+m}$ with B=$\left[\frac{S_x}{\gamma^2} + S_y + \frac{nm(\bar{x}-\bar{y}-\delta)^2}{n+m\gamma^2}\right]$,

$$L_{max}(\delta,\gamma|X,Y) = L_F(\gamma) \times L_t(\delta|\gamma) \tag{8}$$

where

$$L_F(\gamma) \propto F^{m/2}[(n-1)+(m-1)F]^{-(n+m)/2} \tag{9}$$

$$L_t(\delta|\gamma) \propto \left(1 + \frac{t^2}{n+m-2}\right)^{-\left(\frac{n+m}{2}\right)} \tag{10}$$

$F = \frac{S_y/[(m-1)\sigma^2]}{S_x/[(n-1)\gamma^2\sigma^2]} = \frac{(n-1)S_y\gamma^2}{(m-1)S_x}$, $t^2 = \frac{nm\cdot(\bar{x}-\bar{y}-\delta)^2}{(m\gamma^2+n)\hat{\sigma}^2}$, and $\hat{\sigma}^2 = \frac{\frac{S_x}{\gamma^2}+S_y}{m+n-2}$.

By taking $\delta = \theta_L$ and $\delta = \theta_U$, the corresponding t in (10) equal to T_L and T_U respectively.

3.3. Bayesian approach

3.3.1. Crossover Design

To help derive the Bayesian posterior distribution a change of parameters will be needed. Rather than using (1) we employ the following form of distribution for x_i and y_i (which bears no deviation from the distribution (1) and its likelihood function (3))

$$\begin{pmatrix} x_i = \log(x_i^*) \\ y_i = \log(y_i^*) \end{pmatrix} \sim BVN\left(\begin{pmatrix} \mu_1 \\ \mu \end{pmatrix}, \begin{pmatrix} \gamma^2\sigma^2 & \rho\gamma\sigma^2 \\ \rho\gamma\sigma^2 & \sigma^2 \end{pmatrix}\right) \tag{11}$$

The density function is in the form of

$$f(x_i, y_i) = \frac{1}{2\pi\gamma\sigma^2\sqrt{1-\rho^2}} exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_i-\mu_1)^2}{\gamma^2\sigma^2} + \frac{(y_i-\mu)^2}{\sigma^2} - \frac{2\rho(x_i-\mu_1)(y_i-\mu)}{\gamma\sigma^2}\right]\right\}$$

The likelihood function for $\mu_1, \mu, \gamma, \sigma, and\ \rho$ can be expressed as

$$L(\mu_1, \mu, \gamma, \sigma, \rho | X, Y) \propto \frac{1}{\gamma^n\sigma^{2n}(1-\rho^2)^{n/2}} exp\left(-\frac{1}{2\sigma^2(1-\rho^2)}C\right) \tag{12}$$

where $C = \frac{S_x}{\gamma^2} + S_y + \frac{n(\bar{x}-\mu_1)^2}{\gamma^2} + n(\bar{y}-\mu)^2 - \frac{2\rho}{\gamma}\sum_{i=1}^{n}(x_i-\mu_1)(y_i-\mu)$, S_x and S_y follow (3).

A suitable reference prior was suggested by Berger and Sun [44]

$$\pi_{RO}(\mu_1, \mu, \gamma, \sigma, \rho) \propto \frac{1}{\gamma^2\sigma^3(1-\rho^2)^{\frac{3}{2}}} = \sigma^{-3}\gamma^{-2}(1-\rho^2)^{-\frac{3}{2}}$$

Using this prior, the posterior density of the parameters $(\mu_1, \mu, \gamma, \sigma, \rho)$ is in the form of

$$\pi(\mu_1, \mu, \gamma, \sigma, \rho|X, Y) \propto \gamma^{-n-2}\sigma^{-2n-3}(1-\rho^2)^{-\frac{n}{2}-\frac{3}{2}} exp\left(-\frac{1}{2\sigma^2(1-\rho^2)}C\right) \qquad (13)$$

where *C* is defined in (12).

After some algebra and integrating out parameters other than δ, the marginal posterior distribution of δ originated from (13) can be presented in the following form

$$\pi(\delta, |X, Y) \propto \int_{-1}^{1}\int_{0}^{1} u^{\frac{n-1}{2}}(1-u)^{\frac{n-1}{2}}(1-\rho^2)^{\frac{n-1}{2}}\left[1-2\rho\sqrt{(1-u)u}\right]^{-1} \cdot \Big[S(u,\rho)^2 + 1-u\omega(x-\delta-y)2+4\rho21-uux-\delta y-\rho21-uu(x-\delta+y)2\,1-2\rho(1-u)u-n\cdot 1-F(t(\delta,u,\rho))2\,dud\rho \qquad (14)$$

where $F(t(\delta, u, \rho))$ is the student's *t*-distribution function with n degrees of freedom evaluated at $t(\delta, u, \rho)$. Derivation of (14) along with the expression of $t(\delta, u, \rho)$ is shown in Appendix C.

3.3.2. Parallel Design

In parallel designs, the distribution of $x_i, i = 1, \dots, n;\ y_j, j = 1, \dots, m$ can be referred

to (11) with ρ=0, the likelihood function of $\mu_1, \mu, \gamma, and\ \sigma$ follows

$$\mathrm{L}(\mu_1, \mu, \gamma, \sigma|\, X, Y) \propto \frac{1}{\gamma^n \sigma^{n+m}} exp\left(-\frac{1}{2\sigma^2} D\right) \tag{15}$$

where $D = \frac{\sum_{i=1}^{n}(x_i-\mu_1)^2}{\gamma^2} + \sum_{j=1}^{m}\left(y_j - \mu\right)^2 = \frac{S_x}{\gamma^2} + S_y + \frac{n(\bar{x}-\mu_1)^2}{\gamma^2} + m(\bar{y}-\mu)^2$

A suitable prior distribution as suggested in [44] is in the form of

$$\pi_{RO}(\mu_1, \mu, \gamma, \sigma) \propto \frac{1}{\gamma^2 \sigma^3} = \sigma^{-3}\gamma^{-2}$$

Then the posterior density function of parameters $(\mu_1, \mu, \gamma, \sigma)$ can be expressed as

$$\pi(\mu_1, \mu, \gamma, \sigma|X, Y) \propto\ \gamma^{-n-2}\sigma^{-(n+m)-3} exp\left(-\frac{1}{2\sigma^2} D\right) \tag{16}$$

After some algebra and integrating out parameters other than δ, the marginal posterior distribution of δ originated from (16) can be expressed as:

$$\pi(\delta|X, Y) \propto \int_0^1 u^{\frac{m-1}{2}}(1-u)^{\frac{n-1}{2}}[S(u)^2 + (1-u)u(\bar{\mathrm{x}} - \delta - \bar{\mathrm{y}})^2]^{-\frac{n+m+1}{2}} \cdot [1 -$$

$$F(t(\delta,u))\ du \tag{17}$$

Where $F(t(\delta, u))$ is the student's t-distribution function with $(n + m + 1)$ degrees of freedom evaluated at

$$t(\delta, u) = \sqrt{\frac{(n+m+1)}{S(u)^2 + (1-u)u(\bar{\mathrm{x}} - \delta - \bar{\mathrm{y}})^2}} \cdot \{-[(1-u)(\bar{x}-\delta) + u\bar{y}]\}$$

Derivation of (17) follows similarly to that of (14).

3.4. Inference on BE affected by unspecified variances

In this section, we illustrate the proposed likelihood method, using data from a published crossover study, to assess the impact of unspecified variances on the inference about BE. The unspecified variances are characterized by γ^2 (the ratio of two variances), σ_{R}^2 (variance for the reference drug) and ρ (correlation coefficient). To do so, we construct a likelihood function of δ by retaining these three parameters in the function (5).

3.4.1. A log-normal data set and TOST result

Data in Example 1 came from the published crossover BE study by Marzo *et al.* [33].

Test drug ln(AUC0-t): {7.89, 6.53, 6.48, 6.62, 7.93, 6.99, 6.14, 6.78, 6.89, 7.43, 6.52, 6.78, 5.56, 6.84, 7.04, 6.52, 6.70, 6.63, 6.29, 6.41, 5.88, 6.25, 6.92, 7.23}.

Reference drug ln(AUC0-t): {7.61, 6.80, 6.61, 6.16, 7.83, 6.94, 5.45, 6.45, 6.94, 7.08, 6.39, 6.71, 5.22, 6.68, 6.88, 6.32, 6.85, 6.75, 6.38, 6.86, 6.44, 5.96, 7.06, 6.97}.

The study compared a new formulation of ticlopidine hydrochloride (test) to the reference Tiklid (marketed drug). Ticlopidine hydrochloride is an inhibitor of

platelet aggregation used in the management and prevention of thromboembolic disorders. Using the notations in 3.2.1, the observed mean values of ln(AUC) are $\bar{x}$=6.72 and $\bar{y}$=6.64 for test and reference respectively; S_x=6.94, S_y=7.64 and S_{xy}=6.34, and correlation is $\hat{\rho}$ =0.870. Results of TOST using SAS© 9.2, PROC T-Test Procedure are summarized in Table 3.1.

Table 3.1. TOST results, confidence intervals and equivalence assessments for Examples 1 and 2 data

	Example 1	Example 2
N	24	12
Mean: $\bar{x} - \bar{y}$	0.0796	0.0417
95% CI Mean	(-0.0421, 0.2013)	(-0.2888, 0.3721)
90% CI Mean	(-0.0213, 0.1804)	(-0.2280, 0.3113)
SD	0.2883	0.5201
95% CI SD	(0.2240, 0.4044)	(0.3685, 0.8831)
TOST (α=0.05)		
p-value (upper)	<.0001	0.0528
p-value (lower)	0.0114	0.1261
p-value (overall)	0.0114	0.1261
Assessment	Equivalent	Not equivalent

The 90% CI of $\hat{\delta}=\bar{x}-\bar{y}$ is (-0.0213, 0.1804) which falls within the equivalence margins (-0.223, 0.223), and overall p-value, *p*=0.0114 is significant at 0.05 level. These results clearly suggest rejection of the null hypothesis in (2).

3.4.2. Standardized profile likelihood (SPL)

Figure 3.1 shows the plot of the standardized profile likelihood (SPL) for $L_{max}(\delta,\rho,\gamma|X,Y)$ of δ as shown in (5) with $\hat{\gamma}$=0.953 and $\hat{\rho}$=0.87 (sample estimates) and δ in the range of (-0.4, 0.4). The standardization was taken to have the maximum value of 1 for the vertical scale of the plot: SPL=$L_{max}(\delta|X,Y,\rho=.87,\gamma=.953)/L_{max}(\delta=\bar{x}-\bar{y}|X,Y,\rho=.87,\gamma=.953)$ and because both ρ and γ are specified, the component $L_F(\rho,\gamma)$ in (5) is cancelled out in SPL. According to [38], SPL=1/8 and 1/32 are two common reference lines, values that fall into the likelihood interval (LI) corresponding to SPL=1/8 and 1/32 (namely 1/8 LI and 1/32 LI) indicate “moderate strong” and “strong” evidence supported by the data, respectively. Two vertical solid lines represent BE margins -0.223 and 0.223. The horizontal dashed line, SPL=0.226 was obtained by inserting one-sided critical value of t distribution, *t*(0.95,23)=1.714 into SPL. The two points on the SPL curve intersecting with SPL=0.226 are -0.0213 and 0.1804, yielding the same two limits of the 90% CI using TOST shown in Table 3.1. To avoid the confusion with traditional CIs, we name these two limits on the

SPL curve as SPL CI, and SPL CI with *t*(0.95,*df*) as 90% SCI. Hence, with γ and ρ taking values given above, the (1-2α)% SCI matches the (1-2α)% CI using TOST. This can be proven by letting $t^2 = \left(\frac{\bar{x}-\bar{y}-\delta}{\hat{\sigma}_{(\bar{x}-\bar{y})}}\right)^2 = t^2$(1-$\alpha$,$\upsilon$) in (7), which results in $\delta_{1,2}$ (on SPL) = $\bar{x} - \bar{y} \pm t$(1-α,υ)$\cdot\ \hat{\sigma}_{(\bar{x}-\bar{y})}$.

In Figure 3.1, the 90% SCI, 95% SCI and 1/8 LI are all comprised within the BE margins, though 1/32 LI does not, suggesting the evidence is somewhere between moderate to strong in favor of BE. Note that 1/8 LI is very close to 95% SCI (and hence 95% CI), which is expected as Royall [38] showed for normally distributed random variables that1/8 and 1/32 LIs are approximately the same as the 95% and 99% CIs, though 1/32 LI appears to be excessive in the BE setting. Note that the width of (1-2α)% SCI equaling to w=2 t(1-α,υ)$\cdot\ \hat{\sigma}_{(\bar{x}-\bar{y})}$, is proportion to $\hat{\sigma}_{(\bar{x}-\bar{y})}$ and according to (7) contains the parameters γ and ρ. We use contour plots to characterize how widths of 90% and 95% SCIs are related to γ and ρ. Figure 3.2 shows the contour plots of widths of SCIs using Example 1 data, in terms of γ and ρ. The solid thick line represents the width of 90% SCI=0.2013 (on left panel) and 95% SCI =0.2435 (on right panel) which are both narrower than the width of the BE margins (≈0.446). With $\hat{\gamma}$=0.953, $\hat{\rho}$= 0.87 and $\hat{\sigma}_{\bar{x}-\bar{y}}$ = 0.0588 the data suggest adequate evidences of equivalence. Given that there are a range of pairs of γ and ρ associated with the same width of a SCI, it should help identify the source of the variability either due to heterogeneity, weak

correlation or large variances of the two samples that could reduce the probability in achieving BE.

Figure 3.1. Plot of SPL of Delta for Example 1 data

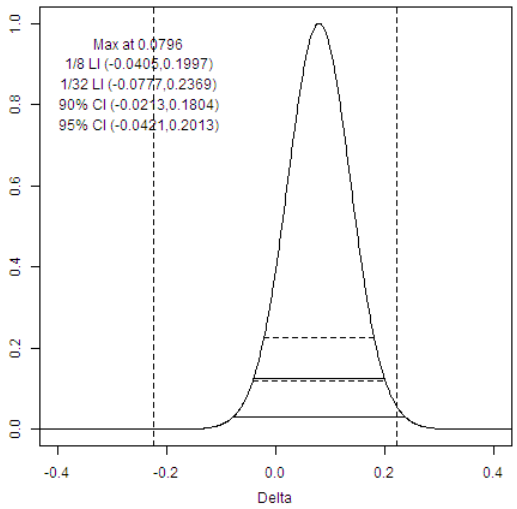
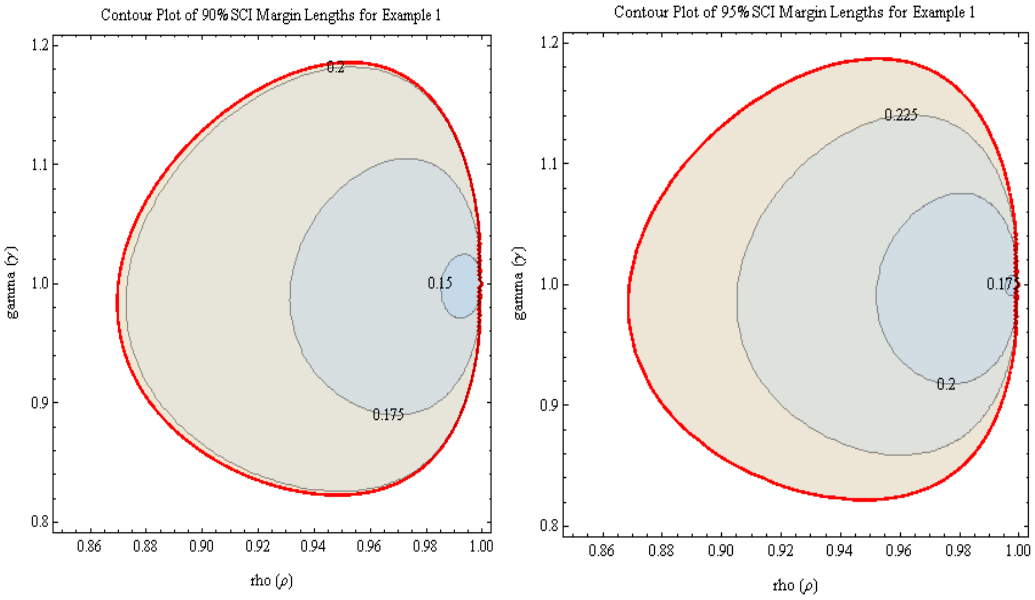


Figure 3.2. Contour plots for the widths of 90% and 95% SCIs for Example 1



3.4.3. Inference about BE affected by γ

In this section, we show the impact unspecified variances have on BE inference. To do so, we retain γ in the likelihood function (4) but replace ρ with its sample estimate $\hat{\rho}$ =0.87. The use of the sample estimate for ρ allows us to explore how inference about BE (based on δ) can be affected by γ. The distances between the upper and lower margins are considered in demonstrating the sensitivities of the inferences to γ. Sample estimator $\hat{\gamma} = \sqrt{S_x/S_y}$=0.953 is chosen as the base value. Alternative values of γ can be obtained by certain confidence intervals or other meaningful justification. Figure 3.3 presents ratios of γ at its alternative values against its base value (note that a reciprocal value is taken if ratio is less than 1). It also shows the ratios of lengths between the upper and lower margins observed at the alternative over the base value of γ. The ratios of lengths are presented at three levels of SPL, those corresponding to 90% and 95% confidence intervals (SCIs) and those at 1/8 LI.

Figure 3.3. Ratios of gamma and length of margin at the alternative values against base value

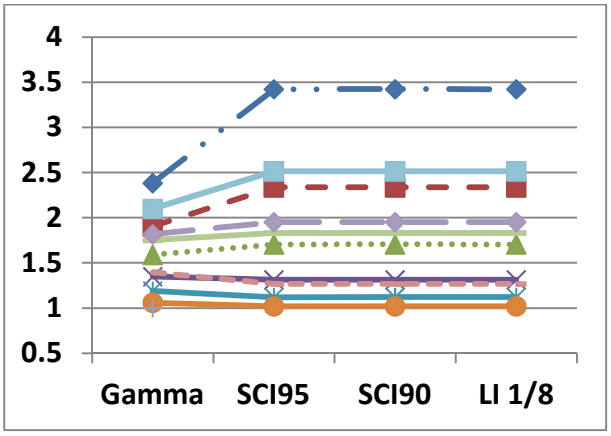


Figure 3.3 shows the ratios of the margin lengths are almost proportional to the ratios of γ's. Moreover, the ratios of margin lengths, and therefore the sensitivities, are independent of the level of SPL. An exact calculation based on (8) gives us a range of γ where BE is achieved using Example 1 data is $0.670 \leq \gamma \leq 1.439$ at 90% SCI level. The 90% SCI when γ=0.670 or γ=1.439 is (-0.064, 0.223). But in either case, the variances of the two drugs are no longer homogeneous raising concerns that one of the drugs carries more variability in rate of absorption than the other one, even though the 90% SCI is still within the BE margins.

Analysis of BE by TOST use sample estimate so that test statistics follows student t-distribution. It conceals the extent to which inference about BE is affected by γ, which was also noted in [7]. Moreover, when analyzing BE data using TOST, it's not uncommon that variances of two groups are pooled

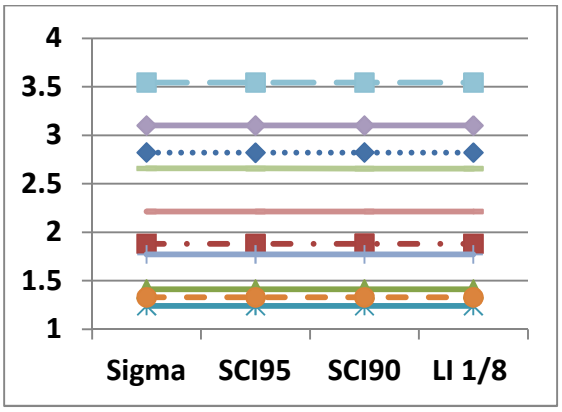
regardless whether they are homogeneous. As indicated by Hua *et al.* [22] the control of type I error rate may be inflated when pooling of two variances is unjustifiable. We believe that, in the presence of unspecified variances and the extent of heterogeneity is a concern, the proposed model $L_{max}(\delta, \rho, \gamma|X, Y)$ (5) could help determine a range of γ where BE is more likely to be achieved.

3.4.4. Inference about BE affected by σ

When γ^2 is a known constant (e.g. γ^2=c^2), unspecified variances can be represented by the reference drug σ^2 (σ_R^2=σ^2). Again the distances between the upper and lower margins are considered in demonstrating the sensitivities of the inferences to σ. Sample estimator $\hat{\sigma}$ =0.56 is chosen as the base value. Alternative values of σ can be obtained by confidence intervals or other meaningful justification. Figure 3.4 presents the sensitivities of the inferences to σ. The figure shows the ratios of σ at its alternative values against its base value (take reciprocal if ratio is less than 1), and also the ratios of the lengths between the upper and lower margins observed at the alternative over the base value. Again, the ratios of lengths are presented at three levels of SPL: those corresponding to 90% and 95% confidence intervals (SCI) and those at 1/8 LI. Since the width of length is proportional to $\hat{\sigma}$, the ratio of $\hat{\sigma}$ at alternative values

and the corresponding ratios at three levels of SPLs are constant, as shown below.

Figure 3.4. Ratios of sigma and length of margin at the alternative values against base value



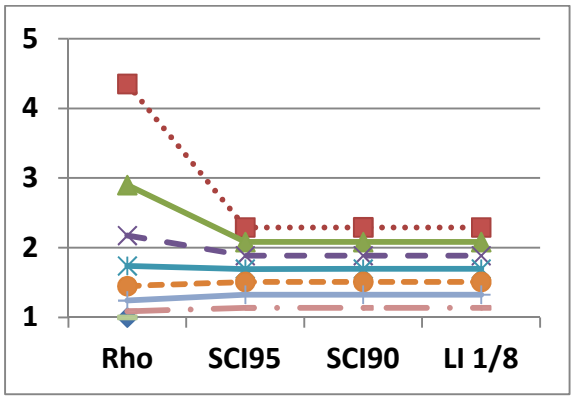
The range of σ where BE is more likely to be achieved for Example 1 data is $\sigma \leq$0.817, with 90% SCI (-0.064, 0.223).

Since coefficient of variation (CV) of the original scale and σ of the natural log-transformed data follow $CV^2 = \exp(\sigma^2) - 1$, the impact (on BE) by a large σ in log-transformed data is synonymous to a large CV in original scale. Thus in the presence of unspecified variances and the extent of CV is a concern, the proposed model $L_{max}(\delta, \rho, \gamma | X, Y)$ (5) could help determine a range of σ where BE is more likely to be achieved.

3.4.5 Inference about BE affected by ρ

Following the approach in the two previous sections, sample estimator $\hat{\rho}$ = 0.87 is the base value. Figure 3.5 shows the sensitivities of the inferences to ρ. The figure presents the ratios of ρ at its alternative values against its base value (take reciprocal if ratio is less than 1), and also presents the ratios of the lengths between the upper and lower margins observed at the alternative over the base value of ρ. Again the ratios of lengths are presented at three levels of SPL, those corresponding to 90% and 95% confidence intervals (SCIs) and those at 1/8 LI. The sensitivity relationship between margin length and ρ is, understandably, more complicated. An exact calculation gives the range of ρ where BE is more likely to be achieved is $\rho \geq 0.648$ for Example 1. The 90% SCI margin for ρ=0.648 is (-0.064, 0.223).

Figure 3.5. Ratios of rho and length of margin at the alternative values against base value



3.5. Comparison of Different Approaches in Crossover Design

In this section, we compare results based on the proposed likelihood, TOST, and Bayesian approaches using two data sets. Example 1 was presented in Section 3.4.1, additional descriptive statistics include $\hat{\gamma}$=0.953, $\hat{\rho}$= 0.87, $\sqrt{S_x/n}$ =0.538, $\sqrt{S_y/n}$ =0.564 and $\hat{\sigma}_{\bar{x}-\bar{y}}$ = 0.0588.

Example 2 is also a crossover BE data with

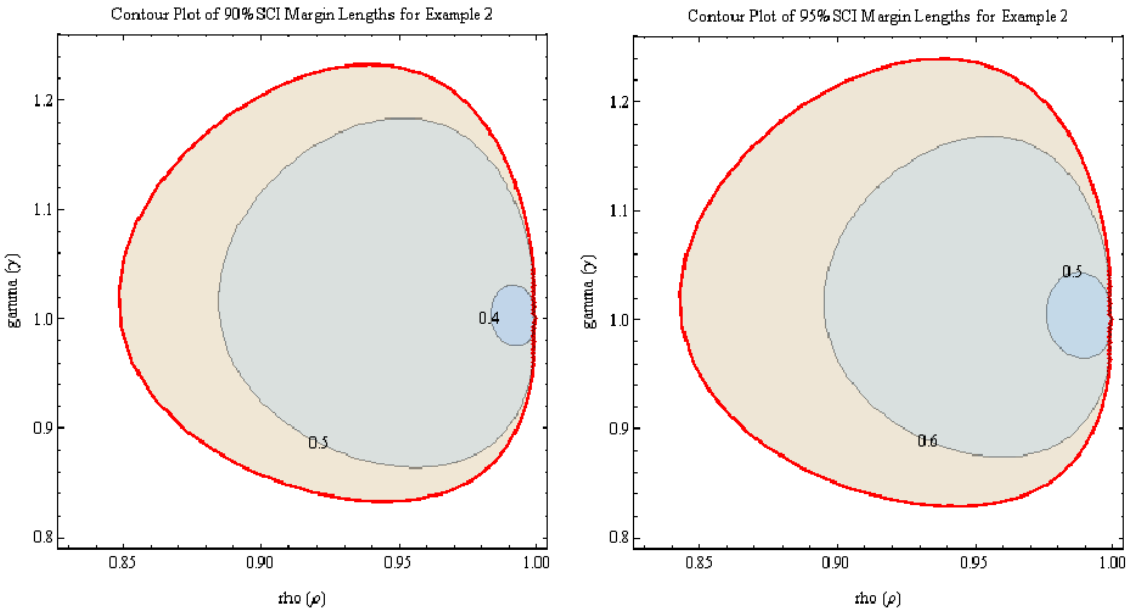
Test drug x_i: {7.85, 4.77, 6.81, 6.64, 6.59, 7.81, 8.07, 7.92, 8.09, 7.11, 7.76, 6.95},

Reference drug y_i: {8.30, 5.35, 6.35, 6.70, 6.57, 6.57, 7.47, 7.91, 8.45, 7.57, 7.72, 6.91}.

Data (ln(AUC) only) were presented in Balthasar [2] from a simulated cyclosporine bioequivalence study with 12 subjects, for a two-treatment, crossover study of 300mg of cyclosporine administered orally. Data were simulated assuming the same parameter values and variances for both study periods; consequently, there was no 'true' difference between the formulations on their rate and extent of absorption. However, these parameter values lead to moderately high values of intra-subject variability in AUC. In addition to descriptive statistics presented in Table 3.1, $\bar{x} - \bar{y}$=0.0417, $\hat{\gamma}$=1.053, $\hat{\rho}$=0.844, $\sqrt{S_x/n}$ =0.912, $\sqrt{S_y/n}$ =0.866 and $\hat{\sigma}_{\bar{x}-\bar{y}}$=0.150. The contour plots for the widths of

90% and 95% SCIs using Example 2 data is shown in Figure 3.6. The solid thick line represents the width of 90% SCI =0.5393 (on the left panel) and 95% SCI =0.6609 (on the right panel), which are larger than the distance of the standard BE margins (0.446). It represents failing to achieve BE for the two drugs based on the standard BE criteria. On the other hand, this example represents a class of drugs with high variability, for which a widened equivalence margins (e.g. ln(0.7) and ln(1/0.7)) are often considered appropriate. The implications of the three parameters ($\gamma, \sigma, and\ \rho$), and their impact on BE follows similarly to those discussed for Example 1.

Figure 3.6. Contour plots for the widths of 90% and 95% SCIs for Example 2



TOST results

Results in Table 3.1 suggest that equivalence in ln(AUC) is achieved for Example 1 and failed for Example 2.

Proposed likelihood functions of δ (5) and (7) with γ replaced by its sample estimate

Table 3.2 shows the SPL CIs of the likelihood functions in (5) and in (7) with γ and ρ replaced with the sample estimate. In both examples, the SPL CIs 90% and 95% SCIs are nearly identical results to the 90% and 95% CIs of the difference $\hat{\delta}=\bar{x}_T - \bar{x}_R$ generated using the TOST method (Table 3.1), which according to Sec 3.4.2, is theoretically correct.

Table 3.2. Results of Profile Likelihood and Fieller Pivotal

Data set	SPL at *t*(0.95, df)	SPL CI	SPL at *t*(0.975, df)	SPL CI
Example 1	SPL=0.226	Likelihood in (5) -0.0213, 0.180 Likelihood in (7) -0.0213, 0.180	SPL=0.118	Likelihood in (5) -0.0421, 0.2013 Likelihood in (7) -0.0422, 0.2013
Example 2	SPL=0.194	Likelihood in (5) -0.223, 0.3113 Likelihood in (7) -0.228, 0.3113	SPL=0.0918	Likelihood in (5) -0.2888, 0.3721 Likelihood in (7) -0.2888, 0.3721

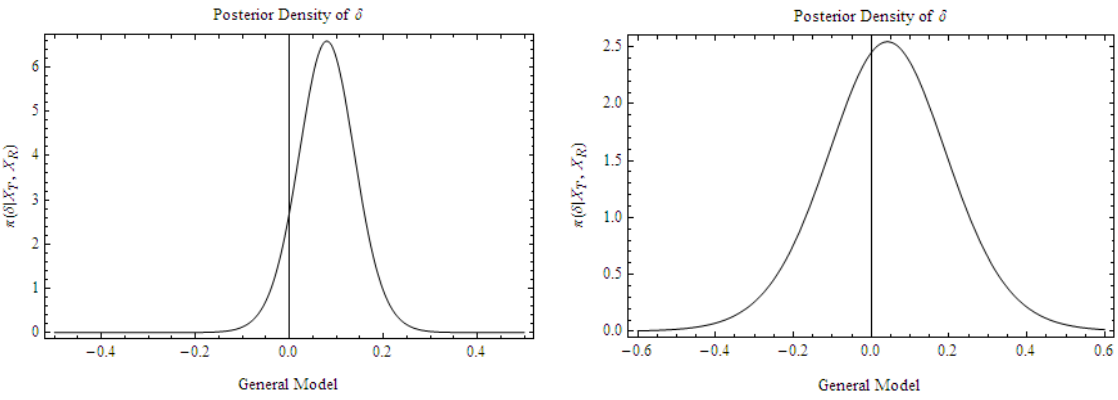
Bayesian analysis with γ, σ, ρ integrated out

Table 3.3 presents the Bayesian analysis results based on the marginal posterior distribution of δ provided in (14). The highest posterior density (HPD) interval is frequently referred to as the shortest Bayesian credible interval that satisfies the following two conditions: i) the posterior probability of that set is 100(1-α)%, ii) the minimum density of any point within that set is equal to or larger than the density of any point outside that set [10]. The results in Tables 3.3 include 90% and 95% HPDs and the probability of posterior distribution of δ falling within the equivalence margins (-0.223, 0.223) for each example. Posterior density of δ is plotted and shown in Figure 3.7 for the two examples. Results in Table 3.3 suggest that posterior distribution of δ yield lightly wider HPD intervals relative to the corresponding SPL CIs using the proposed likelihood method (Table 3.2) or the CIs produced by the TOST method (Table 3.1).

Table 3.3. Results of Bayesian Analysis

Data set	90% HPD	95% HPD	Prob of δ falls within (-0.223, 0.223)
Example 1	-0.0232, 0.1823	-0.0445, 0.2036	0.9873
Example 2	-0.2381, 0.3214	-0.3016, 0.3850	0.8072

Figure 3.7. Bayesian Posterior Density for Two Examples



Comparison

This section along with results in Sec 3.4.1 suggest that CIs of δ generated by TOST can be matched exactly with those SPL CIs generated by the likelihood function in (5) or in (7) with γ and ρ replaced with sample estimate. This suggests that likelihood functions (5) and (7) will provide the same information about δ in general case. In the special case where γ and ρ may also contain information about δ, then SPL CIs of δ generated by (5) will be narrower than that by (7) because $L_F(\rho, \gamma)$ in (6) is dependent on δ. This point is demonstrated in a ready to submit paper by the same authors. The Bayesian HPD intervals of δ generated from (14) are slightly wider than CIs generated by TOST, and SPL CIs using the likelihood function (5) or (7). On the other hand, Bayesian approach does not overcome the drawback that it conceals the extent to which inference on BE is affected by γ, ρ or σ.

3.6. Conclusions and future direction

We believe that when it comes to establish BE, one should look beyond just relying on simple hypothesis testing to reach a final conclusion. The proposed likelihood method that measures and presents strength of evidence from the data offers more information than just relying on simple hypothesis testing to reach a final conclusion on BE, even though TOST are the most commonly used decision making tools. The proposed method addresses the inadequacy of hypothesis testing by including the parameter θ (i.e., γ^2, $\sigma_R^2|\gamma^2 = c$ or ρ) to represent unspecified variances. This enables us to assess the extent that BE is affected by the variability of PK parameter θ, i.e. assess the robustness of conditional BE inference. Our analyses confirms that probability of achieving BE could be diminished when i) variances of the two drugs are heterogeneous, ii) variances are large even homogeneous is maintained and iii) data from the two drugs are less correlated in a cross over study. Furthermore, the proposed likelihood method allows us to find a range of θ where BE is more likely to be achieved. When either high heterogeneity or large variances are observed and these variances are unknown, the proposed method is particularly effective in making inference about the extent that BE is affected by the unspecified variances than methods that replace θ with sample estimate or integrate out θ.

The proposed likelihood method is well suited to assess the evidence of BE for

log-normal distributed data in both cross-over and parallel designs. Because the likelihood function in (5) is presented in two components—a function of F statistic and a function of t statistic, both 95% and 90% CLs of δ (difference in two mean ln(AUC)) generated using TOST can be produced by 95% SCI and 90% SCI respectively. They are also comparable to or better than the 95% and 90% HPDs obtained by Bayesian posterior distribution of δ. The proposed likelihood method establishes a threshold under which BE is more likely to be achieved and measure the strength of evidence using SCIs. We believe the proposed method is a better alternative than the usual TOST method. In the cases where BE is barely met (i.e., the 90% CI of δ barely falls within (-0.223, 0.223)), knowing the extent of the variances that could change the outcome from bioequivalence to failure or *vice versa*, as discussed in Sections 3.4 and 3.5, would motivate drug developers to address and control the variability, and produce improved bioequivalence results. Ultimately, bioequivalent drug products approved by the regulators with stronger evidence will not only benefit patients but also yield more satisfactory outcome to drug makers and regulatory agencies.

Going forward, in addition to the current work, we will be considering the proposed likelihood method for population bioequivalence inference, which is more complicated than average bioequivalence. Lastly, sample programs used to perform all the calculations including R codes for likelihood and profile likelihood methods, *Wolfram* Mathematica® 7 code for Bayesian model are available for request.

4. Inference of equivalence for the ratio of two normal means with unspecified variances (Third Topic)

4.1. Introduction

In section 3, we addressed the inference of bioequivalence for log-normal data with unspecified variances by expressing the entire likelihood function in the product of two components: *F*-statistic function for variances and *t*-statistic function for the difference of the parameter of interest. For bioequivalence studies, U.S. FDA [17] recommends that 90% confidence interval (CI) values for the ratio of the relative means for AUC and Cmax of the new and reference drugs should fall between 0.80 and 1.25. The distribution of PK parameters is often assumed to be log-normal, therefore bioequivalence is usually assessed on the difference of logarithmically transformed PK parameters. The proposed method in section 3 retains the unspecified variances in the likelihood function via three parameters: the ratio of two variances (γ^2), the variance of the reference drug σ_R^2, and correlation of the two drugs (ρ) for cross-over studies only. The word "unspecified" means those parameters are included in the model and allowed to

vary over a range of values, to show the robustness of bioequivalence inference conditional on different values. This method helps identify ranges of variances where bioequivalence is more likely to be achieved, which cannot be accomplished by methods that replace or remove these parameters such as two one-sided tests (TOST), or Bayesian analysis. This method applies to both cross-over and parallel designs.

The endpoint in late phase study for equivalence test is often normal distributed. When the untransformed variable is normally distributed, Hauschke, *et al*. [20] gave several examples to justify the need to assess the therapeutic equivalence for two treatments using the ratio of two normal means. Liu J-P and Weng C-S [31] and Berger and Hsu [5] showed that analyses based on the original scale are always more powerful than those based on the transformed scale when the distribution is normal. A likelihood-ratio test proposed by Sasabuchi has been suggested by several authors including Hauschke et al. [20], Berger and Hsu [5] and Hua *et al.*, [22], among others, as appropriate to test the *null hypothesis of non-equivalence* for the ratio of two normal means. Similar to the log-normal data, according to the literature, confidence interval inclusion such as Fieller confidence set [15] is more commonly used than test procedures for normal data.

In this topic, we expand the likelihood approach by Diaz-Frances and Sprott [14] and what we have covered in section 3 to address equivalence for the ratio

of two normal means (β) by expressing the entire likelihood function in the product of two components: a function of *F*-statistic for variances and the a function of the *t*-statistic for β. In essence, the findings regarding inference of bioequivalence for log-normal data that is affected by the unspecified variances as described in the beginning paragraph and what we have covered in section 3 can be applied to the inference of equivalence for the ratio of two normal means. We show that, the proposed method is effective in investigating the impact of variances on the inference of equivalence. Since this has already been addressed in section 3, we will only briefly discuss this point for normal data in this part. Our focus in this topic is to show additional benefits of the proposed method. In general case when model parameters are independent, the proposed full-likelihood function method produces results that are same as the likelihood-ratio test and comparable to Bayesian analysis. In the special case when model parameters are dependent, for example the ratio of two variances is directly proportional to the ratio of two means, the full-likelihood function yields better results in inference about equivalence than the likelihood-ratio test which relies solely on the t-statistic function or Bayesian analysis which integrates out the variances in the posterior distribution.

The remainder of this part of topic is organized as follows. Section 4.2 describes the proposed likelihood method. Section 4.3 presents the Bayesian model. The impact on equivalence by the unspecified variances is described in

section 4.4. Comparisons of different methods in general and special cases will be shown in section 4.5 followed by summary in section 4.6.

4.2. The Likelihood Method

In this section, we give details on a whole likelihood function that can be used to characterize equivalence for normal data based on bivariate normal distribution. The method addresses equivalence trials in cross-over and parallel designs.

4.2.1. Crossover Design

Suppose two variables x_i and $y_i, i = 1, \dots, n,$ follow bivariate normal (BVN) distribution

$$\binom{x_i}{y_i} \sim BVN\left(\binom{\beta\mu}{\mu}, \begin{pmatrix} \gamma^2\sigma^2 & \rho\gamma\sigma^2 \\ \rho\gamma\sigma^2 & \sigma^2 \end{pmatrix}\right) \quad (1)$$

In the context of equivalence studies, variables x_i and y_i represent individual observations of an efficacy parameter for the two treatments (test and reference)

respectively. Parameters in model (1) represent $\beta = E(x_i)/E(y_i)$, $\mu = E(y_i)$, ρ the correlation coefficient between the two drugs, and γ^2 the ratio of two variances ($\gamma^2 = Var(x_i)/Var(y_i)$) and $\sigma^2 = Var(y_i)$. The correlation coefficient ρ is to address cross-over designs where subjects start with one treatment and then are switched to the other treatment usually after a wash-out period, as such x_i and y_i are observations from the same subject, while in parallel designs ρ=0. Model (1) assumes there is no sequence or period effect.

Hypothesis and Likelihood-ratio Test

It follows from [5, 20, 31] the equivalence hypotheses for the ratio β with the pre-specified margins (θ_1, θ_2) can be stated as,

$$H_{01}: \beta < \theta_1 \text{ or } H_{02}: \beta > \theta_2 \qquad (2)$$

$$H_A: \theta_1 \leq \beta \leq \theta_2$$

These authors further indicated that the size-α likelihood ratio test originally proposed by Sasabuchi (1998) [41, 42] is appropriate for the hypotheses in (2) and rejects H_0 if

$$T_1^c \geq t_{\alpha,n-1} \text{ and } T_2^c \leq -t_{\alpha,n-1} \qquad (3)$$

where $T_h^c = \frac{\bar{x}-\theta_h\bar{y}}{\hat{\sigma}_{(\bar{x}-\theta_h\bar{y})}}, h = 1, 2.$ $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$, $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}, S_x = \sum_{i=1}^n (x_i - \bar{x})^2$, and $S_y = \sum_{i=1}^n (y_i - \bar{y})^2 . t_{\alpha,v}$ is the (1-α) percentile of the central *t*-distribution with *v* degrees of freedom.

Of note, T_h^c in (3) is in a general form, the letter C in superscript means crossover design. For convenience, we chose the equivalence margins: θ_1=0.8, θ_2=1.25 to assess equivalence for the ratio of two normal means, however, readers should always consider clinically relevant margins in their research and practice.

Fieller Confidence Set

An equivalence confidence interval (CI) for the ratio of two means can be constructed using the generalized Fieller theorem [15]. By solving the quadratic equation $T_h^{c\,2} = t_{\alpha,n-1}^2$ from (3), we obtained the confidence limits as follows:

$$\theta_{\pm}^c = \frac{(\bar{x}\bar{y}-a_c\rho\gamma)\pm\sqrt{(\bar{x}\bar{y}-a_c\rho\gamma)^2-(\bar{x}^2-a_c\gamma^2)(\bar{y}^2-a_c)}}{\bar{y}^2-a_c}, a_c = \frac{\hat{\sigma}^2}{n} t_{\alpha,n-1}^2 \,. \quad (4)$$

After some algebraic, it can be shown that inequalities in (3) are operationally equivalent to

$$\theta_-^c \geq \theta_1 \; and \; \theta_+^c \leq \theta_2 \; and \; \bar{y}^2 > a_c \quad (5)$$

Equalities in (5) can be interpreted as follows: if the 100(1-2α) per cent Fieller confidence set for β has finite length, it is given by the interval $I_F^c = (\theta_-^c, \theta_+^c)$.

Furthermore, Fieller's confidence set is an interval if and only if $\bar{y}^2 > a_c$ holds true.

Of the two one-sided tests proposed by Sasabuchi and CI based on the Fieller theorem, the former compares T_1^c (T_2^c) with critical *t*-value under a pre-specified value of θ_1 (θ_2); while the latter assumes θ is a unknown random variable and compare its range of plausible values with the margins (θ_1, θ_2).

Full Likelihood Function

The likelihood approach focuses on showing the statistical evidence instead of making decision to choose between equivalence and nonequivalence using the hypotheses (2). The likelihood function based on model (1) is in the form of

$$L(\beta, \mu, \rho, \gamma, \sigma | X, Y) \propto \frac{1}{\gamma^n \sigma^{2n} (1-\rho^2)^{n/2}} exp\left(-\frac{1}{2\sigma^2(1-\rho^2)} A\right) \quad (6)$$

where $A = \frac{\sum_{i=1}^n (x_i - \beta\mu)^2}{\gamma^2} + \sum_{i=1}^n (y_i - \mu)^2 - \frac{2\rho}{\gamma} \sum_{i=1}^n (x_i - \beta\mu)(y_i - \mu) = \frac{S_x}{\gamma^2} + S_y + \frac{n(\bar{x} - \beta\mu)^2}{\gamma^2} + n(\bar{y} - \mu)^2 - \frac{2\rho}{\gamma} \sum_{i=1}^n (x_i - \beta\mu)(y_i - \mu)$.

Without the knowledge of true μ and σ, a likelihood function of β, γ and ρ can be constructed by replacing μ in (6) with its restricted maximum likelihood estimator (rMLE) $\hat{\mu}_{rM}(\beta, \gamma, \rho) = \frac{(\beta - \rho\gamma)\bar{x} + \gamma(\gamma - \rho\beta)\bar{y}}{(\beta^2 - 2\rho\beta\gamma + \gamma^2)}$, followed by replacing σ^2 with its rMLE

$\hat{\sigma}_{rM}^2 = \frac{B}{2n(1-\rho^2)}$ where B=$\left[\frac{S_x}{\gamma^2} + S_y + \frac{n(\bar{x}-\beta\bar{y})^2(1-\rho^2)}{(\beta^2-2\rho\beta\gamma+\gamma^2)} - \frac{2\rho}{\gamma}S_{xy}\right]$, and $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$. That is

$$L_{max}(\beta, \rho, \gamma | X, Y) \propto \frac{(1-\rho^2)^{\frac{n}{2}}}{\gamma^n}\left[\frac{S_x}{\gamma^2} + S_y + \frac{n(\bar{x}-\beta\bar{y})^2(1-\rho^2)}{(\beta^2-2\rho\beta\gamma+\gamma^2)} - \frac{2\rho}{\gamma}S_{xy}\right]^{-n} \quad (7)$$

Following some algebra, the profile likelihood function in (7) can be factored into two parts:

$$L_{max}(\beta, \rho, \gamma | X, Y) = L_F(\rho, \gamma) \times L_t(\beta | \rho, \gamma) \quad (8)$$

where

$$L_F(\rho, \gamma) \propto \left(\frac{1}{FS_y\gamma\sqrt{1-\rho^2}}\right)^n \quad (9)$$

$$L_t(\beta | \rho, \gamma) \propto= \left(1 + \frac{t^2}{2(n-1)}\right)^{-n} \quad (10)$$

$F = \frac{\frac{S_x}{\gamma^2}+S_y-\frac{2\rho}{\gamma}S_{xy}}{2S_y(1-\rho^2)}$, $t = \frac{\bar{x}-\beta\bar{y}}{\hat{\sigma}_{(\bar{x}-\beta\bar{y})}} = \frac{\bar{x}-\beta\bar{y}}{\sqrt{\frac{(\beta^2-2\rho\beta\gamma+\gamma^2)}{n}\frac{\left(\frac{S_x}{\gamma^2}+S_y-\frac{2\rho}{\gamma}S_{xy}\right)}{2(n-1)(1-\rho^2)}}}$.

Derivation of (6)-(10) follows similarly to Appendix B. Note that $\hat{\sigma}^2$ is the sample estimate of σ^2 and differs slightly from $\hat{\sigma}_{rM}^2$. Taking $\beta = \theta_1$ and $\beta = \theta_2$ the corresponding t in (10) equal to T_1^c and T_2^c respectively. This means $L_t(\beta|\rho, \gamma)$ in (10) can produce the results of likelihood ratio test and more importantly the likelihood function in (8) which is the product of (9) and (10), will provide more information than likelihood ratio test alone on the inference about equivalence. A more detailed discussion can be found in Sections 4.4 and 4.5.

4.2.2. Parallel Design

In a parallel design where samples $x_i, i = 1, \dots, n$, and $y_j, j = 1, \dots, m$ are independent ($\rho = 0$), following Diaz-Frances and Sprott [14], $x_i \sim N(\beta\mu, \gamma^2\sigma^2), y_j \sim N(\mu, \sigma^2)$. The corresponding expressions of (9) and (10) are:

$$L_F(\gamma) \propto F^{\frac{m}{2}}[(n-1)+(m-1)F]^{-\frac{n+m}{2}} \tag{11}$$

$$L_t(\beta|\gamma) \propto \left(1+\frac{t^2}{n+m-2}\right)^{-\frac{n+m}{2}} \tag{12}$$

where $F = \frac{S_y/[(m-1)\sigma^2]}{S_x/[(n-1)\gamma^2\sigma^2]} = \frac{(n-1)S_y\gamma^2}{(m-1)S_x}$, $t^2 = \left(\frac{\bar{x}-\beta\bar{y}}{\hat{\sigma}_{\bar{x}-\beta\bar{y}}}\right)^2 = \frac{nm\cdot(\bar{x}-\beta\bar{y})^2}{(m\gamma^2+n\beta^2)\hat{\sigma}^2}$, and $\hat{\sigma}^2 = \frac{\frac{S_x}{\gamma^2}+S_y}{m+n-2}$, with d.f. n+m-2.

Sasabuchi (1998) demonstrated that the size-α likelihood ratio test rejects H_0 if

$$T_1^p \geq t_{\alpha,n+m-2} \; and \; T_2^p \leq -t_{\alpha,n+m-2}\,, \tag{13}$$

respectively, where $T_i^p = \frac{\bar{x}-\theta_i\bar{y}}{\hat{\sigma}_{(\bar{x}-\theta_h\bar{y})}}, i = 1, 2.$

$t_{\alpha,v}$is the (1-α) percentile of the central t-distribution with v degrees of freedom. Algebraic rearrangement shows that condition (13) is equivalent to

$$\theta_-^p \geq \theta_1 \; and \; \theta_+^p \leq \theta_2 \; and \; \bar{y}^2 > a_R$$

where

$$\theta_{\pm}^{p} = \frac{\bar{x}\bar{y} \pm \sqrt{(a_R\bar{x}^2 + a_T\gamma^2\bar{y}^2 - a_T a_R\gamma^2)}}{\bar{y}^2 - a_R}, a_T = \frac{\hat{\sigma}^2}{n} t^2_{\alpha,n+m-2} \; and \; a_R = \frac{\hat{\sigma}^2}{m} t^2_{\alpha,n+m-2}. \quad (14)$$

Condition (14) can be interpreted as follows: if the 100(1-2α) per cent Fieller confidence set for β has finite length, it is given by the interval $I_F^p = (\theta_-^p, \theta_+^p)$. Furthermore, Fieller's confidence set is an interval if and only if $\bar{y}^2 > a_R$ holds true.

Compare with method proposed by Sasabuchi and method based on the Fieller theorem, the former one plug with θ_1 (θ_2) and compare T_1^p (T_2^p) with critical t-value; while the later one assume θ is a unknown random variable and compare its range of plausible values with the margins (θ_1, θ_2). Both of them involved with the Fieller theorem, which is the same as t-statistic function $L_t(\beta|\rho,\gamma)$ in (12). Thus inference of β based on (12) should be the same as that based on the Fieller confidence set I_F^p. In other words the proposed likelihood function, which is the product of (11) and (12), is based on more information than methods only based on the Fieller theorem.

4.3. Bayesian approach

4.3.1. Crossover Design

To help derive the Bayesian posterior distribution a change of parameters will be needed. Rather than using (1) we employ the following form of distribution for x_i and y_i (which bears no deviation from the distribution (1))

$$\binom{x_i}{y_i} \sim BVN\left(\binom{\mu_1}{\mu}, \begin{pmatrix} \gamma^2\sigma^2 & \rho\gamma\sigma^2 \\ \rho\gamma\sigma^2 & \sigma^2 \end{pmatrix}\right) \tag{15}$$

Under the bivariate normal model (15), as the likelihood function for $\mu_1, \mu, \gamma, \sigma, and\ \rho$ can be expressed as

$$\mathrm{L}(\mu_1, \mu, \gamma, \sigma, \rho | X, Y) \propto \frac{1}{\gamma^n \sigma^{2n} (1-\rho^2)^{n/2}} exp\left(-\frac{1}{2\sigma^2(1-\rho^2)} C\right) \tag{16}$$

where $C = \frac{S_x}{\gamma^2} + S_y + \frac{n(\bar{x}-\mu_1)^2}{\gamma^2} + n(\bar{y}-\mu)^2 - \frac{2\rho}{\gamma}\sum_{i=1}^{n}(x_i-\mu_1)(y_i-\mu)$.

According to Berger, Bernardo, and Sun [3] and Rubio and Perex-Elizalde [40], the independence Jeffreys prior in (17) is a suitable prior.

$$\pi_{IJ}(\mu_1, \mu, \gamma, \sigma, \rho) \propto (\mu_1\mu)^{-\frac{1}{2}} \frac{1}{\gamma\sigma^2(1-\rho^2)^{\frac{3}{2}}} = (\mu_1\mu)^{-1/2}\sigma^{-2}\gamma^{-1}(1-\rho^2)^{-\frac{3}{2}} \tag{17}$$

The posterior distribution of the parameters $(\mu_1, \mu, \gamma, \sigma, \rho)$ will have the form

$$\pi(\mu_1, \mu, \gamma, \sigma, \rho | X, Y) \propto (\mu_1 \mu)^{-1/2} \gamma^{-n-1} \sigma^{-2n-2} (1-\rho^2)^{-\frac{n}{2}-\frac{3}{2}} exp\left(-\frac{1}{2\sigma^2(1-\rho^2)} C\right) \quad (18)$$

Change parameters from $(\mu_1, \mu, \gamma, \sigma, \rho)$ to (β, μ, u, v, ρ), after some algebra and integrating out parameters other than β, the marginal posterior distribution of β has the form

$$\pi(\beta | X, Y) \propto$$

$$\int_{-1}^{1} \int_{0}^{1} \beta^{-\frac{1}{2}} u^{\frac{n-1}{2}} (1-u)^{\frac{n-2}{2}} (1-\rho^2)^{\frac{n-2}{2}} \left[(1-u)\beta^2 + u - 2\rho\sqrt{(1-u)u}\beta\right]^{-1} \cdot$$

$$\mathcal{S}u,\rho 2+1-\mathrm{uu}(1-\rho 2)(\beta \mathrm{y}-\mathrm{x})2\,1-u\beta 2+u-2\rho 1-uu\beta-(n-12)\cdot 1-F(t(\beta,u,\rho))2\cdot dud\rho$$

(19)

Derivation of (19) is similar to Appendix C.

4.3.2. Parallel Design

In a parallel design where samples $x_i, i = 1, \ldots, n$, and $y_j, j = 1, \ldots, m$ are independent ($\rho = 0$), the likelihood function for $\mu_1, \mu, \gamma, and\ \sigma$ can be expressed as

$$\mathrm{L}(\mu_1, \mu, \gamma, \sigma | X, Y) \propto \frac{1}{\gamma^n \sigma^{n+m}} exp\left(-\frac{1}{2\sigma^2} D\right) \quad (20)$$

where D= $\frac{\sum_{i=1}^{n}(x_i-\mu_1)^2}{\gamma^2}+\sum_{j=1}^{m}\left(y_j-\mu\right)^2=\frac{S_x}{\gamma^2}+S_y+\frac{n(\bar{x}-\mu_1)^2}{\gamma^2}+m(\bar{y}-\mu)^2$.

According to Berger, Bernardo, and Sun [3] and Rubio and Perex-Elizalde [40], the independence Jeffreys prior in (21) is a suitable prior.

$$\pi_{IJ}(\mu_1,\mu,\gamma,\sigma)\propto(\mu_1\mu)^{-\frac{1}{2}}\frac{1}{\gamma\sigma^2}=(\mu_1\mu)^{-1/2}\sigma^{-2}\gamma^{-1} \tag{21}$$

The posterior distribution of the parameters $(\mu_1,\mu,\gamma,\sigma)$ will have the form

$$\pi(\,\mu_1,\mu,\gamma,\sigma|X,Y)\propto\ (\mu_1\mu)^{-1/2}\gamma^{-n-1}\sigma^{-n-m-2}exp\left(-\frac{1}{2\sigma^2}D\right) \tag{22}$$

Change parameters from $(\mu_1,\mu,\gamma,\sigma)$ to (β,μ,u,v), after algebra and integrating out parameters μ,u and v, the marginal posterior distribution of β has the form

$\pi(\beta|X,Y)\propto\int_0^1\beta^{-\frac{1}{2}}u^{\frac{m-1}{2}}(1-u)^{\frac{n-2}{2}}[(1-u)\beta^2+u]^{-\frac{1}{2}}\cdot\Big[(1-u)S_x^2+uS_y^2+$

1−uu(βy−x)2*1−uβ2+u−n+m2·1−F(t(β,u))·du*

(23)

where $F(t(\beta,u))$ is the student's t-distribution function with (m+n) degrees of freedom evaluated at

$$t(\beta,u)=\sqrt{\frac{(n+m)\cdot[(1-u)\beta^2+u]}{\left[(1-u)S_x^2+uS_y^2+\frac{(1-\mathrm{u})\mathrm{u}(\beta\bar{\mathrm{y}}-\bar{\mathrm{x}})^2}{(1-u)\beta^2+u}\right]}}\left[-\frac{(1-u)\beta\bar{x}+u\bar{y}}{(1-u)\beta^2+u}\right]$$

Derivation of (23) follows similarly to that of (19).

4.4. Inference on equivalence affected by unspecified variances

In this section, we illustrate the proposed likelihood method using data from a published crossover study, to assess the impact of unspecified variances on the inference about equivalence. The unspecified variances are characterized by γ^2 (the ratio of two variances), σ^2 (variance for the reference drug) and ρ (correlation coefficient). In this section, we focus on γ^2 as it is similar to derive the impact of σ^2 and ρ, which has been covered in section 3. To do so, we construct a likelihood function of β by retaining γ in the model while replacing the other two with sample estimates in the likelihood function (8).

4.4.1. A normal data set and likelihood ratio test result

Data came from the published crossover study by Balthasar [2].

Test drug Cmax: {140, 13.6, 78.8, 88.0, 54.7, 76.4, 310, 110, 182, 192, 364, 112}.

Reference drug Cmax: {226, 20.1, 51.8, 105, 40.6, 52.6, 175, 135, 337, 326, 346, 126}.

Data came from a simulated cyclosporine bioequivalence study with 12 subjects, for a two-treatment, crossover study of 300 mg of cyclosporine administered orally. Data were under the assumption of the same parameter values and variances for both study periods; consequently, there was no 'true' difference between the formulations on their rate and extent of absorption. However, these parameter values lead to moderately high values of intra-subject variability in Cmax (percent coefficient of variation: 23%). Such high variability is not appropriate for demonstrating the proposed method. For illustration purpose, we take square root of original data, and use the transformed data for analysis. After transformation, using the notations in (3), the observed mean values of Cmax are $\bar{x}$=11.25 and $\bar{y}$=11.81 for test and reference respectively; Sx=202.34, Sy=268.52, Sxy=197.22, and correlation is $\hat{\rho}$ =0.85. Results of likelihood ratio test are summarized in Table 4.1. The 90% Fieller confidence set (FCS) of $\hat{\beta}=\frac{\bar{x}}{\bar{y}}$ is (0.8496, 1.0781) which falls within the equivalence margins (0.8, 1.25), and overall p-value in rejecting the two one-sided null hypotheses in (2), *p*=0.0101 is significant at 0.05 level. These results clearly suggest sufficient evidence to conclude equivalence.

Table 4.1. Likelihood ratio test results, Fieller confidence set and equivalence assessment for example data

N	12
Ratio: $\frac{\bar{x}}{\bar{y}}$	0.9530
95% FCS Ratio	(0.8275, 1.1118)
90% FCS Ratio	(0.8496, 1.0781)
Likelihood Ratio Test (α=0.05)	
p-value (upper)	0.0101
p-value (lower)	0.0023
p-value (overall)	0.0101
Assessment	Equivalent

4.4.2. Standardized profile likelihood (SPL)

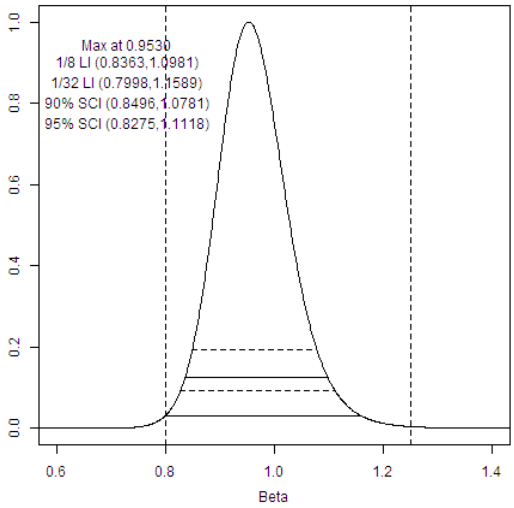
Figure 4.1 shows the plot of the standardized profile likelihood (SPL) for $L_{max}(\beta, \rho, \gamma|X, Y)$ of β as shown in (8) with $\hat{\gamma}$=0.87 and $\hat{\rho}$=0.85 (sample estimates) and β in the range of (0.6, 1.4). The standardization was taken to have the maximum value of 1 for the vertical scale of the plot:

SPL=$L_{max}(\beta|X,Y,\rho=.85,\gamma=.87)/L_{max}\left(\beta=\frac{\bar{x}}{\bar{y}}\middle|X,Y,\rho=.85,\gamma=.87\right)$ and because both ρ and γ are specified, the component $L_F(\rho,\gamma)$ in (8) is cancelled out in SPL. According to [38], SPL=1/8 and 1/32 are two common reference lines, values that fall into the likelihood interval (LI) corresponding to SPL=1/8 and 1/32 (namely 1/8 LI and 1/32 LI) indicate “moderate strong” and “strong” evidence supported by the data, respectively. Two vertical solid lines represent equivalence margins 0.8 and 1.25. The horizontal dashed line, SPL=0.194 was obtained by inserting one-sided critical value of t distribution, t(0.95,11)=1.796 into SPL. The two points on the SPL curve intersecting with SPL=0.194 are 0.8496 and 1.0781. Thus using SPL the limits is the same as Fieller confidence set. To avoid the confusion with traditional CIs, we name these two limits on the SPL curve as SPL CI, and SPL CI with t(0.95,df) as 90% SCI. Hence, with γ and ρ taking values given above, the (1-2α)% SCI matches the (1-2α)% CI using Fieller confidence set. This can be proven by letting $t^2=\left(\frac{\bar{x}-\beta\bar{y}}{\hat{\sigma}_{(\bar{x}-\beta\bar{y})}}\right)^2=t^2$(1-$\alpha$,$\upsilon$) in (10), which results in $\beta_{1,2}$ (on SPL) the same as $\theta_{\pm}^{c}$ given in (4) for crossover study.

In Figure 4.1, all 90% SCI, 95% SCI, 1/8 LI and 1/32 LI are comprised within the equivalence margins, suggesting strong evidence in favor of equivalence. Note that 1/8 LI is very close to 95% SCI (counterpart of 95% CI in SPL setting), which is expected as Royall [39] showed for normally distributed random variables

that1/8 and 1/32 LIs are approximately the same as the 95% and 99% CIs, though 1/32 LI appears to be excessive in the equivalence setting.

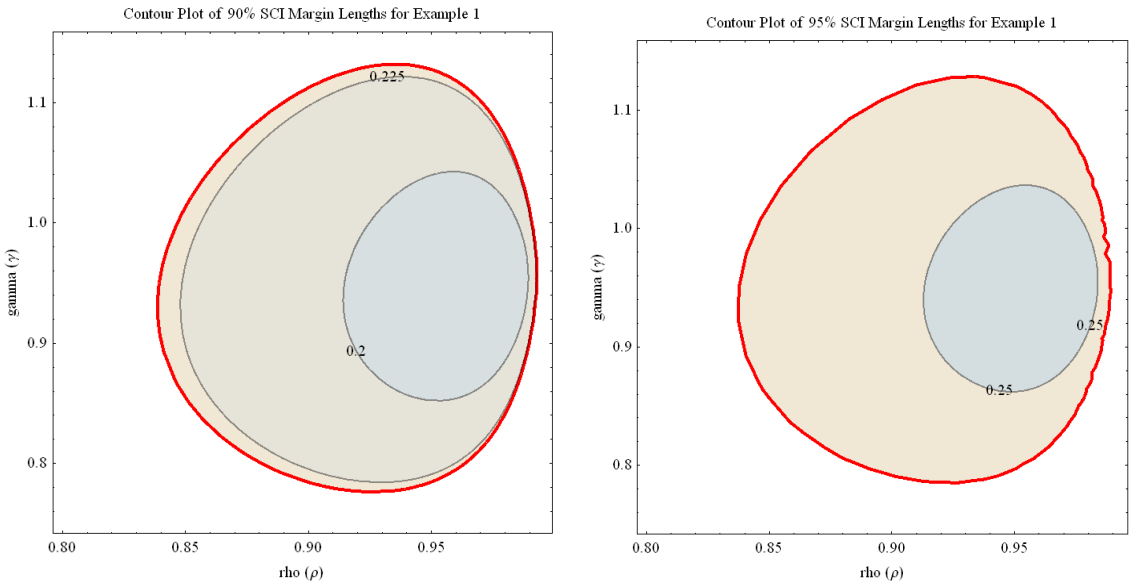
Figure 4.1. Plot of SPL of Beta



Note that the width of (1-2α)% SCI equaing to w=$\theta_+^c - \theta_-^c$, is a function containing the parameters γ and ρ according to (4). We use contour plots to characterize how widths of 90% and 95% SCIs are related to γ and ρ. Figure 4.2 shows the contour plots of widths of SCIs using example data, in terms of γ and ρ, with sample estimate $\hat{\sigma}$. The solid thick line represents the width of 90% SCI=0.2285 (on left panel) and 95% SCI =0.2843 (on right panel) which are both narrower than the width of the Equivalence margins ($= 0.45$). The data suggests adequate evidences of equivalence. Given that there are a range of pairs of γ and ρ associated with the same width of a SCI, it should help identify the source of the

variability either due to heterogeneity, weak correlation or large variances of the two samples that could reduce the probability in achieving Equivalence.

Figure 4.2. Contour plots for the widths of 90% and 95% SCIs

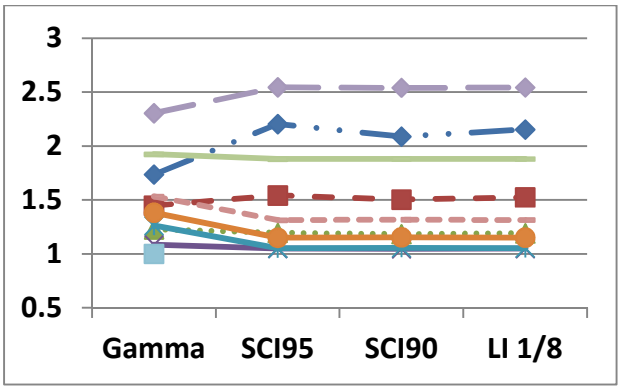


4.4.3. Inference about equivalence affected by γ

In this section, we show the impact unspecified variances have on equivalence inference. To do so, we retain γ in the likelihood function (8) but replace ρ with its sample estimate $\hat{\rho} \approx 0.85$. The use of the sample estimate for ρ allows us to explore how inference about equivalence (based on β) can be affected by γ. The distances between the upper and lower margins are considered in demonstrating

the sensitivities of the inferences to γ. Sample estimator $\hat{\gamma} = \sqrt{S_x/S_y}$ ≈0.87 is chosen as the base value. Alternative values of γ can be obtained by certain confidence intervals or other meaningful justification. Figure 4.3 presents ratios of γ at its alternative values against its base value (take reciprocal if ratio is less than 1). It also shows the ratios of lengths between the upper and lower margins observed at the alternative over the base value of γ. The ratios of lengths are presented at three levels of SPL, those corresponding to 90% and 95% confidence intervals (SCIs) and those at 1/8 LI.

Figure 4.3. Ratios of gamma and length of margin at the alternative values against base value



The ratios of margin lengths, and therefore the sensitivities, are dependent of the level of SPL. Moreover, the sensitivity relationship between margin length and γ is understandably complicated. An exact calculation based on (8) gives us a range of γ where equivalence is achieved using example data is 0.558≤ γ ≤1.305 at 90% SCI level. The 90% SCI when γ=0.558 is (0.811, 1.200), when

γ=1.305 is (0.800, 1.093). Even though the 90% SCI is still within the equivalence margins, the variances of the two drugs are no longer homogeneous raising concerns that one of the drugs carries more variability in rate of absorption than the other one. From this we have shown by methods replacing γ with its rMLE or sample estimate, it may conceal the extent to which inference about equivalence is affected by γ. This experience was also noted in [7]. We believe that, in the presence of unspecified variances and the extent of heterogeneity is a concern, the proposed model $L_{max}(\beta, \rho, \gamma|X, Y)$ (8) could help determine a range of β where equivalence is more likely to be achieved.

4.4.4. Inference about equivalence affected by σ *and* ρ

Similarly we refer to two more sensitivity plots to see inference about equivalence affected by σ and ρ. Sample estimate of each parameter is chosen as the base value. Alternative values of that parameter can be obtained by certain confidence intervals. Figure 4.4 presents ratio of σ at its alternative values against its base value (take reciprocal if ratio is less than 1). It also shows the ratios of margin lengths observed at the alternative over the base value of σ. Three levels of margin are considered, as for γ. Since the margin length is proportional to σ, the ratio of σ at alternative values and the corresponding ratios at three levels of SPL are constant. The sensitivity relationship between margin length and ρ is more

complicated. Exact calculation by the maximum likelihood function gives range of σ and ρ where BE is more likely to be achieved at 90% SCI level: $\sigma \leq 7.38$ (0.800, 1.158); $\rho \geq 0.538$ (0.800, 1.142).

Figure 4.4. Ratios of sigma and length of margin at the alternative values against base value

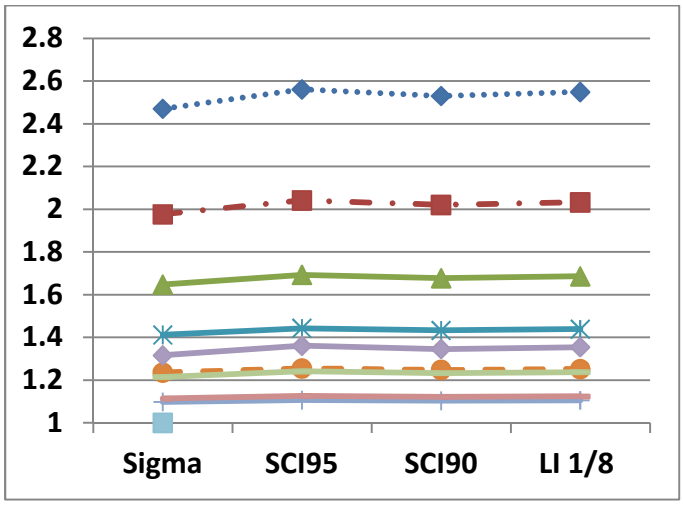
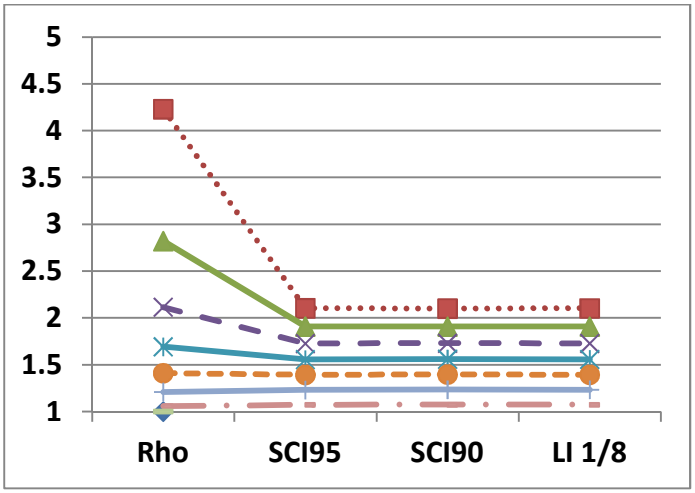


Figure 4.5. Ratios of rho and length of margin at the alternative values against base value



4.5. Comparison of Different Approaches in Crossover Design

In this section, we compare results based on the proposed likelihood method, Fieller confidence set and Bayesian approaches using the above data set. Due to the partition, when model parameters are dependent, for example the ratio of two variances is directly proportional to the ratio of two means, or the correlation between two drugs is directly proportional to the ratio of two means, L_F also can be used for equivalence inference based on the ratio of two means. Including F-statistic function L_F in this case, like the proposed likelihood method, should be able to improve the result than just solely relying on t-statistic function $L_{Fieller}$, like Fieller confidence set. To make the comparisons more clear, we listed L_F

and $L_{Fieller}$ for general case and two special cases in Table 4.2, Table 4.3, and Table 4.4 respectively. Thus we can see that in general case, L_F provides no information of β, thus the proposed likelihood method performs the same as likelihood ratio test or Fieller confidence set. However, in special cases, L_F contains information of β. Hence the proposed likelihood method should perform better. Results of different approaches are listed in the following tables 4.5, 4.6 and 4.7, with plots shown in Figure 4.6.

Table 4.2. In general case, i.e. $\gamma, \sigma \; and \; \rho$ are independent.

	Cross-over Design	Parallel Design
L_F	$\frac{(1-\rho^2)^{\frac{n}{2}}}{\gamma^n}\left[\frac{S_x}{\gamma^2}+S_y-\frac{2\rho}{\gamma}S_{xy}\right]^{-n}$	$\frac{1}{\gamma^n}\left[\frac{S_x}{\gamma^2}+S_y\right]^{-\frac{n+m}{2}}$
L_t	$\left(1+\frac{n(\bar{x}-\beta\bar{y})^2(1-\rho^2)}{\left(\frac{S_x}{\gamma^2}+S_y-\frac{2\rho}{\gamma}S_{xy}\right)(\beta^2-2\rho\gamma\beta+\gamma^2)}\right)^{-n}$	$\left(1+\frac{nm(\bar{x}-\beta\bar{y})^2}{\left(\frac{S_x}{\gamma^2}+S_y\right)(n\beta^2+m\gamma^2)}\right)^{-\frac{n+m}{2}}$

Table 4.3. $\gamma = b\beta$, where b is a known constant

	Cross-over Design	Parallel Design
L_F	$\frac{(1-\rho^2)^{\frac{n}{2}}}{(b\beta)^n}\left[\frac{S_x}{b^2\beta^2}+S_y-\frac{2\rho}{b\beta}S_{xy}\right]^{-n}$	$\frac{1}{(b\beta)^n}\left[\frac{S_x}{b^2\beta^2}+S_y\right]^{-\frac{n+m}{2}}$
L_t	$\left(1+\frac{n(\bar{x}-\beta\bar{y})^2(1-\rho^2)}{\left(\frac{S_x}{b^2\beta^2}+S_y-\frac{2\rho}{b\beta}S_{xy}\right)(\beta^2-2\rho b\beta^2+b^2\beta^2)}\right)^{-n}$	$\left(1+\frac{nm(\bar{x}-\beta\bar{y})^2}{\left(\frac{S_x}{b^2\beta^2}+S_y\right)(n\beta^2+mb^2\beta^2)}\right)^{-\frac{n+m}{2}}$

Table 4.4. $\rho = c\beta$, where c is a known constant

	Cross-over Design	Parallel Design
L_F	$\frac{(1-c^2\beta^2)^{\frac{n}{2}}}{\gamma^n}\left[\frac{S_x}{\gamma^2}+S_y-\frac{2c\beta}{\gamma}S_{xy}\right]^{-n}$	Not applicable
L_t	$\left(1+\frac{n(\bar{x}-\beta\bar{y})^2(1-c^2\beta^2)}{\left(\frac{S_x}{\gamma^2}+S_y-\frac{2c\beta}{\gamma}S_{xy}\right)(\beta^2-2c\beta^2\gamma+\gamma^2)}\right)^{-n}$	Not applicable

Table 4.5. Fieller Confidence Set:

	General Case	Special Case $\gamma = \beta$	Special Case $\rho = 0.8 * \beta$
95% FCS of β	(0.8275, 1.1118)	(0.8277, 1.1048)	(0.7854, 1.1037)
90% FCS of β	(0.8496, 1.0781)	(0.8496, 1.0738)	(0.8184, 1.0754)

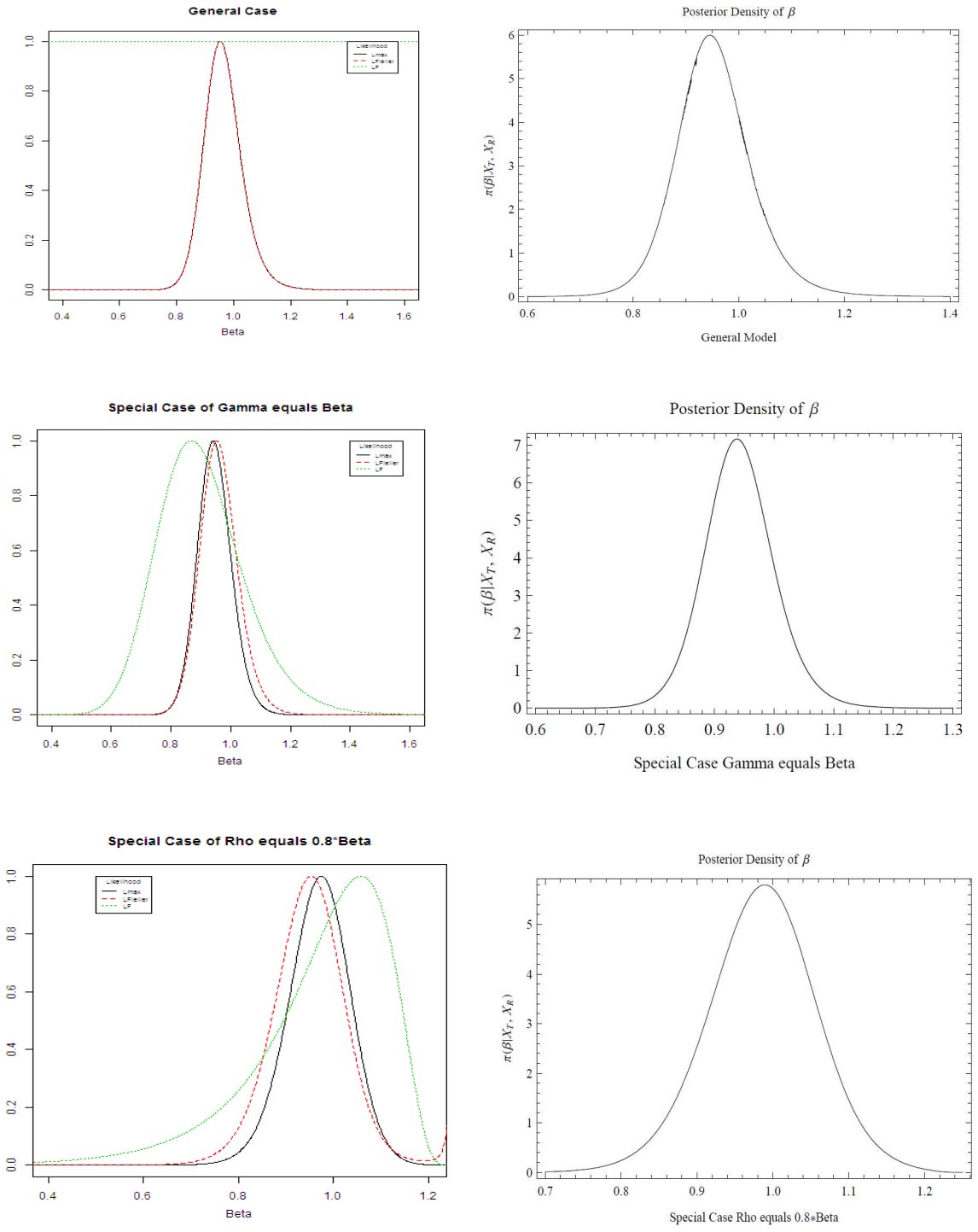
Table 4.6. Proposed Likelihood Method:

	General Case	Special Case $\gamma = \beta$	Special Case $\rho = 0.8 * \beta$
95% SPL of β	L_{max} (0.8275, 1.1118) L_t (0. 8275, 1.1118)	L_{max} (0.8249, 1.0732) L_t (0.8277, 1.1048)	L_{max} (0.8251, 1.1083) L_t (0.7854, 1.1037)
90% SPL of β	L_{max} (0.8496, 1.0781) L_t (0.8496, 1.0781)	L_{max} (0.8451, 1.0476) L_t (0.8496, 1.0738)	L_{max} (0.8536, 1.0850) L_t (0.8184, 1.0754)

Table 4.7. Bayesian Method:

	General Case	Special Case $\gamma = \beta$	Special Case $\rho = 0.8 * \beta$
95% HPD of β	(0.8001, 1.1184)	(0.8244, 1.0633)	(0.8421, 1.1268)
90% HPD of β	(0.8301, 1.0779)	(0.8444, 1.0393)	(0.8683, 1.1033)
Prob of βfalls within (0.8, 1.25)	0.9707	0. 9905	0.9900

Figure 4.6. Plots for general case and two special cases



We show that in general case, results of the proposed likelihood method are the same as Fieller confidence set, and comparable with Bayesian methods. In special case, however, the proposed likelihood method produces narrower confidence interval than Fieller confidence set. Hence the proposed likelihood method is better than Fieller confidence set or likelihood ratio test for equivalence inference. While Bayesian method gives similar length of HPD as the proposed method, both 90% and 95% HPDs of β are shifted to the left when $\gamma = \beta$ and shifted to the right when $\rho = 0.8 * \beta$.

4.6. Summary

The results in this paper suggest that the proposed likelihood method that measures and presents strength of evidence from the data offers more information than just relying on simple hypothesis testing to reach a final conclusion on equivalence of two normal means, even though 90% Fieller confidence interval and likelihood-ratio test are the most commonly used decision making tools. Similar to our findings in section 3, the proposed method can show the impact of unspecified variances on the inference of equivalence and also

help us identify a range of unspecified variances where the equivalence is more likely to be achieved.

In general case, the proposed likelihood method produces results that are same as the likelihood ratio test and comparable to Bayesian analysis. In the special case where the ratio of two means determines the ratio of variances, the proposed method yields better results (narrower CIs) in inference about equivalence than either likelihood-ratio test which relies solely on the t-statistic function or Bayesian analysis which integrates out the variances in the posterior distribution.

Our proposed method is comparable to Bayesian method. Length of HPD is almost the same as the SPL CI, regardless whether it is a general case or special case. However, the HPD is either shifted to left or right, indication the Bayesian method may be over estimate resulting from marginalizing out the unspecified variances. Our research results show that the proposed likelihood method is a better alternative than current analysis method for equivalence inference.

Appendix A. Simulation results for scenarios 1, 3 and 4 in Section 2

Scenario 1:

Response rates of taking placebo and active control are 0.18, 0.50 for HER3+/HER2+; response rates of taking placebo and active control are 0.18, 0.50 for HER3-/HER2+.

[Insert Figure A.1 here]

[Insert Figure A.2 here]

[Insert Figure A.3 here]

[Insert Table A.1 here]

[Insert Table A.2 here]

In this scenario, HER3/HER2 status is not related to effect size of active control, so change of percent of HER3+/HER2+ will not result in constancy assumption being violated, and moreover, treatment effect of reference product over putative placebo keeps the same. All metrics control type I error rate well enough, except a little bit high for metric LRR. Fixed margin method or methods based on it are conservative, while synthesis method or methods based on it control type I error rate and have higher power. It makes sense under constancy assumption. Covariate adjustment methods are stable even when we don't have any covariate effect.

Scenario 3:

Response rates of taking placebo and active control are 0.40, 0.60 within HER3+/HER2+; response rates of taking placebo and active control are 0.13, 0.45 within HER3-/HER2+.

[Insert Figure A.4 here]

[Insert Figure A.5 here]

[Insert Figure A.6 here]

[Insert Table A.3 here]

[Insert Table A.4 here]

In this scenario, HER3/HER2 status is a covariate and moderately interacts with treatment arm, so change of percent of HER3+/HER2+ in NI will result in constancy assumption being violated. Moreover, increasing the percent of HER3+/HER2+ will decrease the “difference” of effect of reference product over putative placebo, and also decrease the “difference” of effect of reference product over test product. Thus the 10% point corresponding to the biggest “Treatment effect of RP over Putative Placebo” in type I error plots. FM and methods based on it are under 0.025. Syn is under 0.025 for percent 10%-50%, and greater than 0.025 for percent 60%-70%. CovSyn is greater than 0.025, while 2sSyn is greater than 0.025 for percent 30%-70%.

Scenario 4:

Response rates of taking placebo and active control are 0.18, 0.78 within HER3+/HER2+; response rates of taking placebo and active control are 0.31, 0.38 within HER3-/HER2+.

[Insert Figure A.7 here]

[Insert Figure A.8 here]

[Insert Figure A.9 here]

[Insert Table A.5 here]

[Insert Table A.6 here]

In this scenario, HER3/HER2 status is a covariate (related with response rate), and is highly related with effect size of reference product (interacts with treatment arm). Thus change of percent of HER3+/HER2+ in NI would result in constancy assumption being violated. Moreover, the higher percent of HER3+/HER2+, the bigger difference between treatment effect of placebo and reference produce. The first three points, where constancy assumption violated most serious, both fixed margin method and synthesis method have inflated type I error rate, with synthesis method inflated more. CovFM and 2sFM always control type I error rate less than target value 0.025, CovSyn and 2sSyn always have type I error rate larger than target value 0.025. Both FM and Syn have type I error rate larger than 0.025 at the fourth point, and less than 0.025 at points five and six.

Figure A.1. Performance Characteristics Comparisons--Measure of Association: Difference.

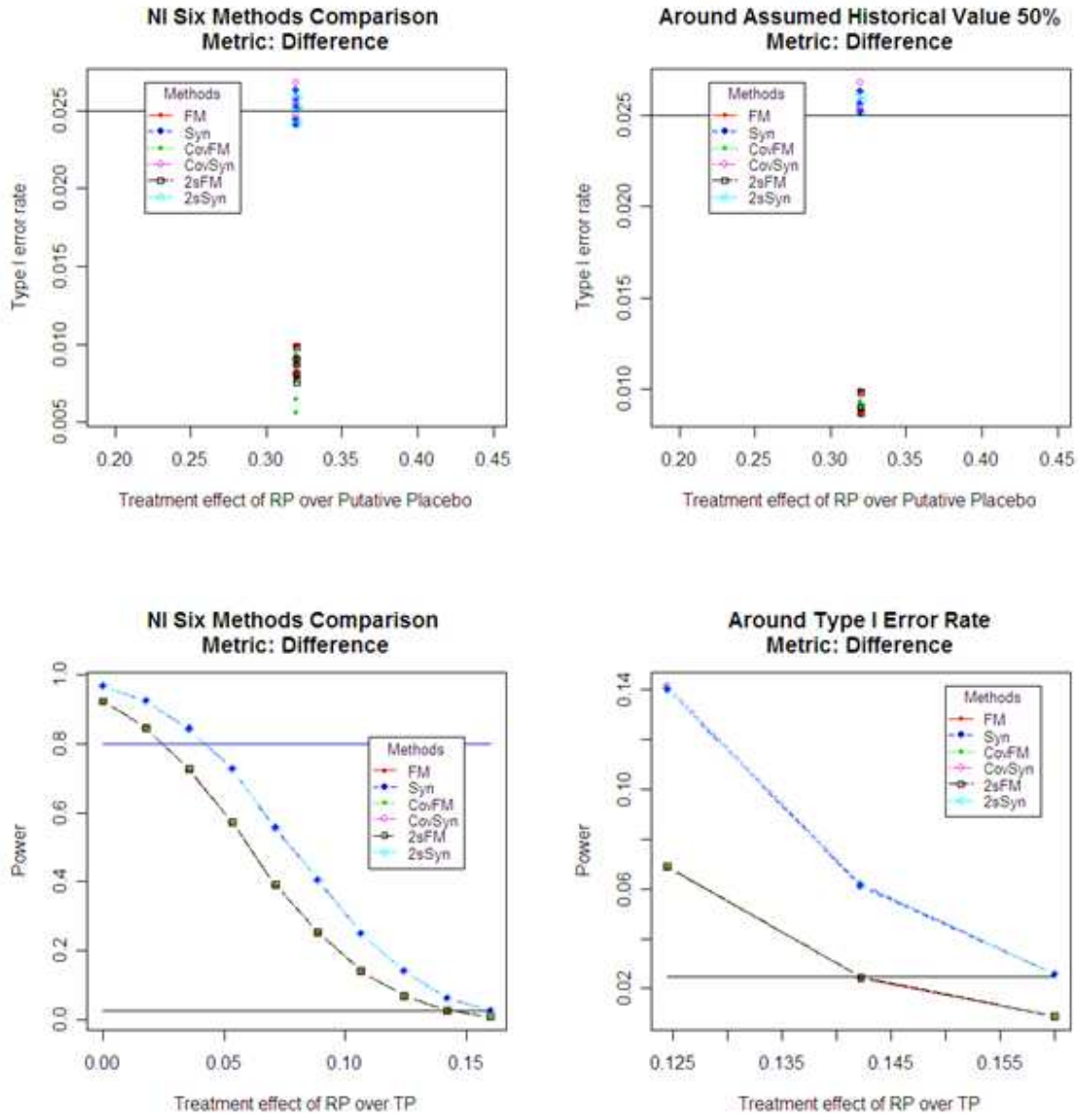


Figure A.2. Performance Characteristics Comparisons--Measure of Association: Log Relative Risk.

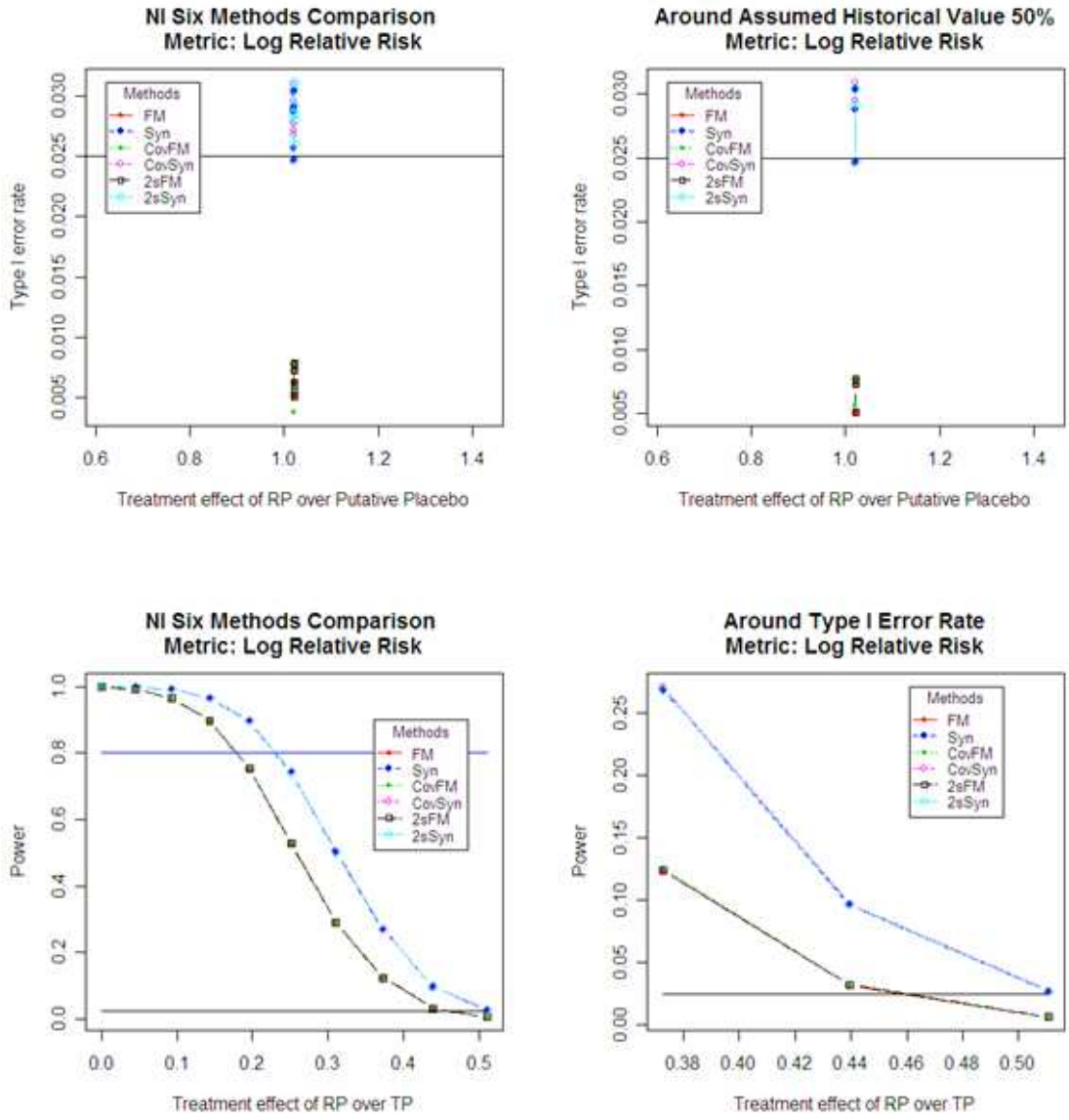


Figure A.3. Performance Characteristics Comparisons--Measure of Association: Log Odds Ratio.

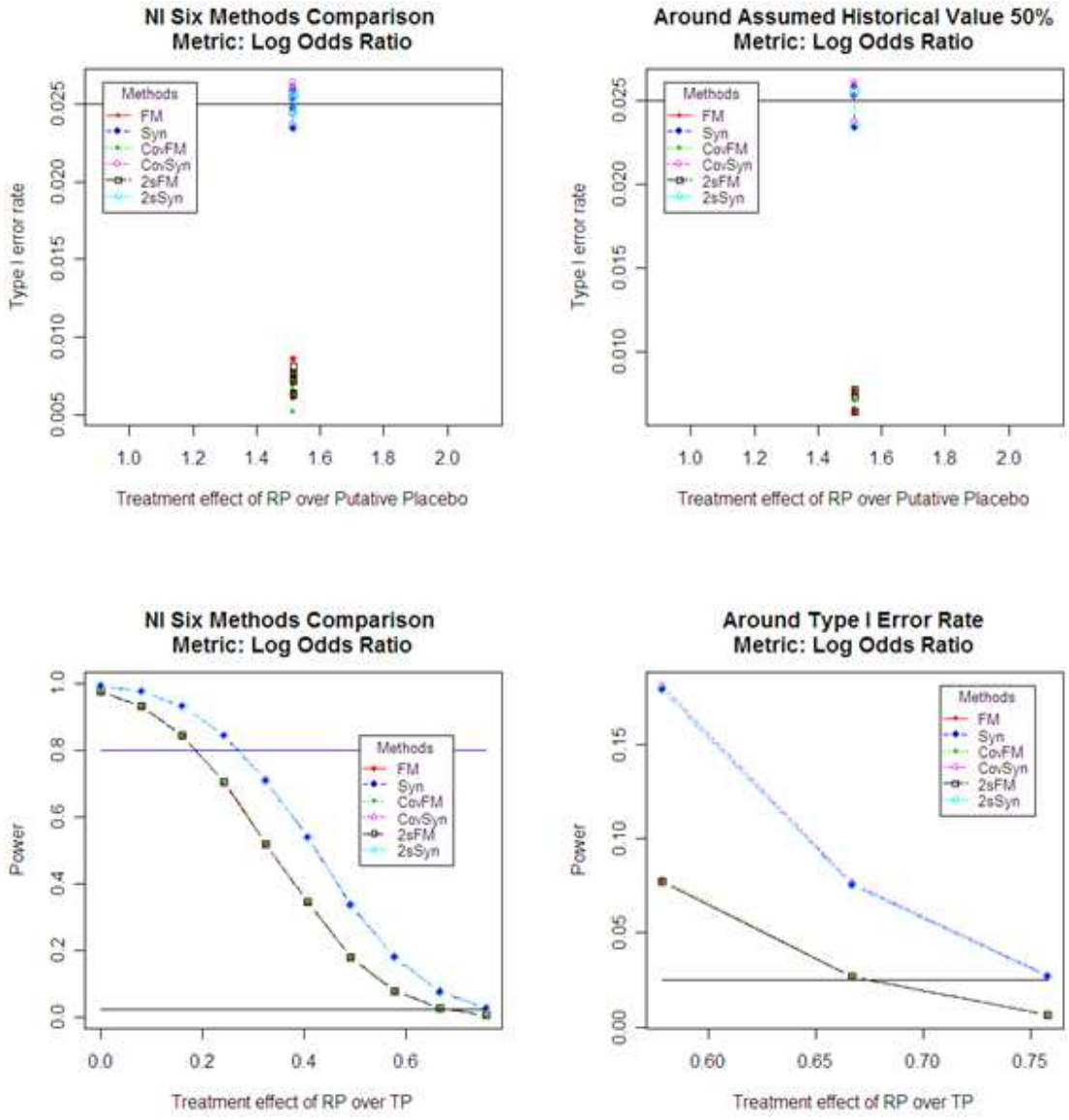


Figure A.4. Performance Characteristics Comparisons--Measure of Association: Difference.

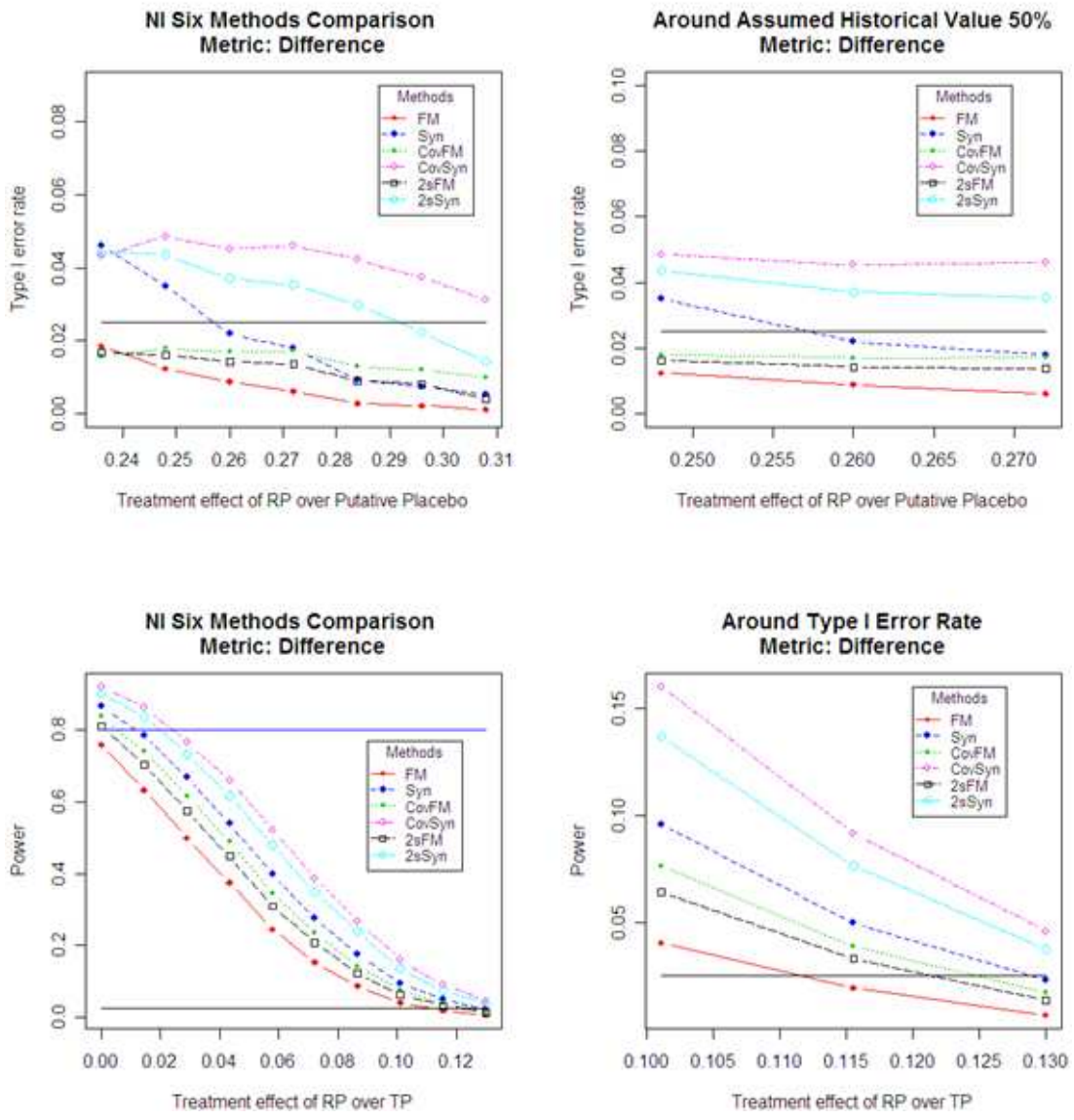


Figure A.5. Performance Characteristics Comparisons--Measure of Association: Log Relative Risk

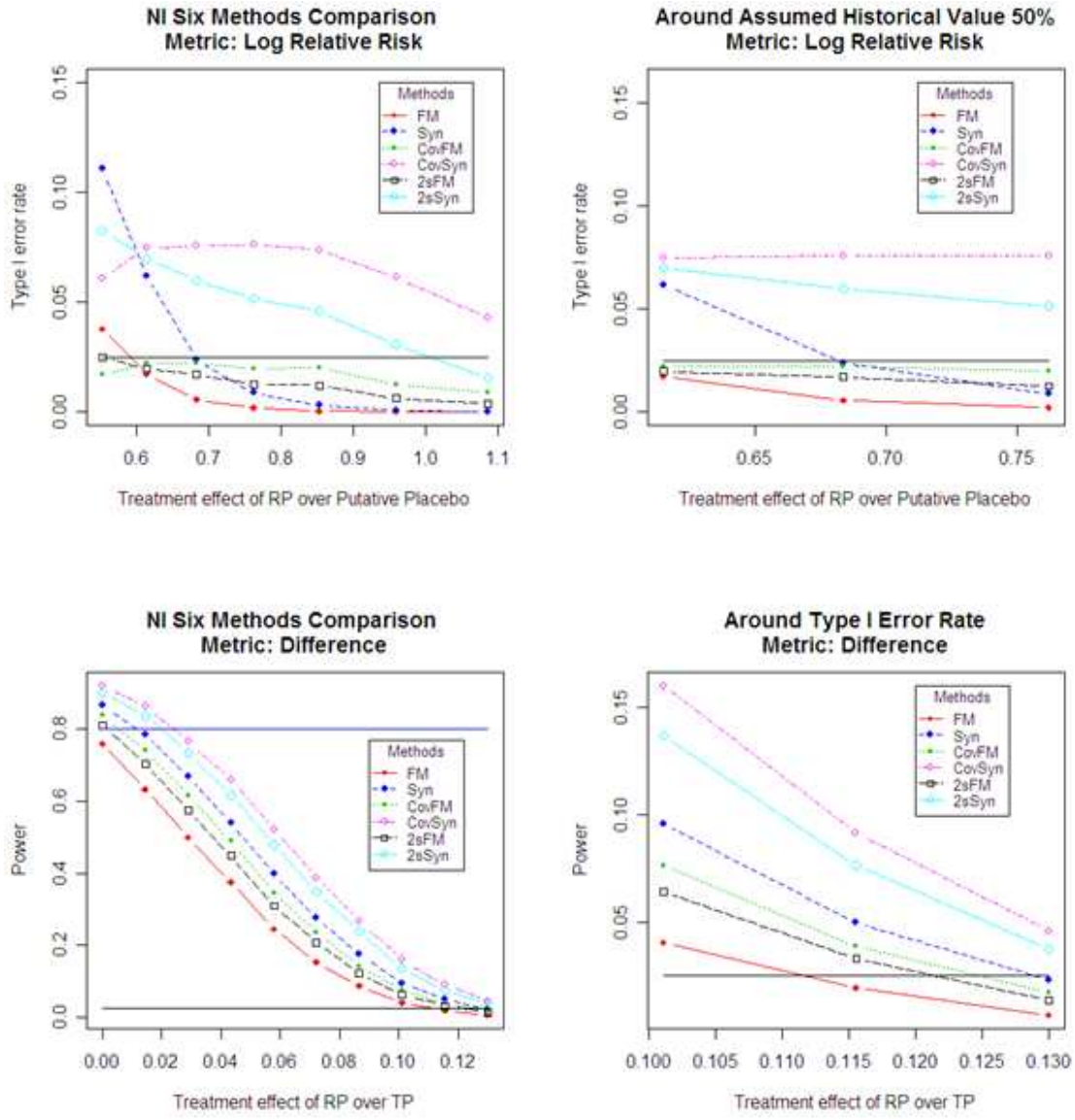


Figure A.6. Performance Characteristics Comparisons--Measure of Association: Log Odds Ratio

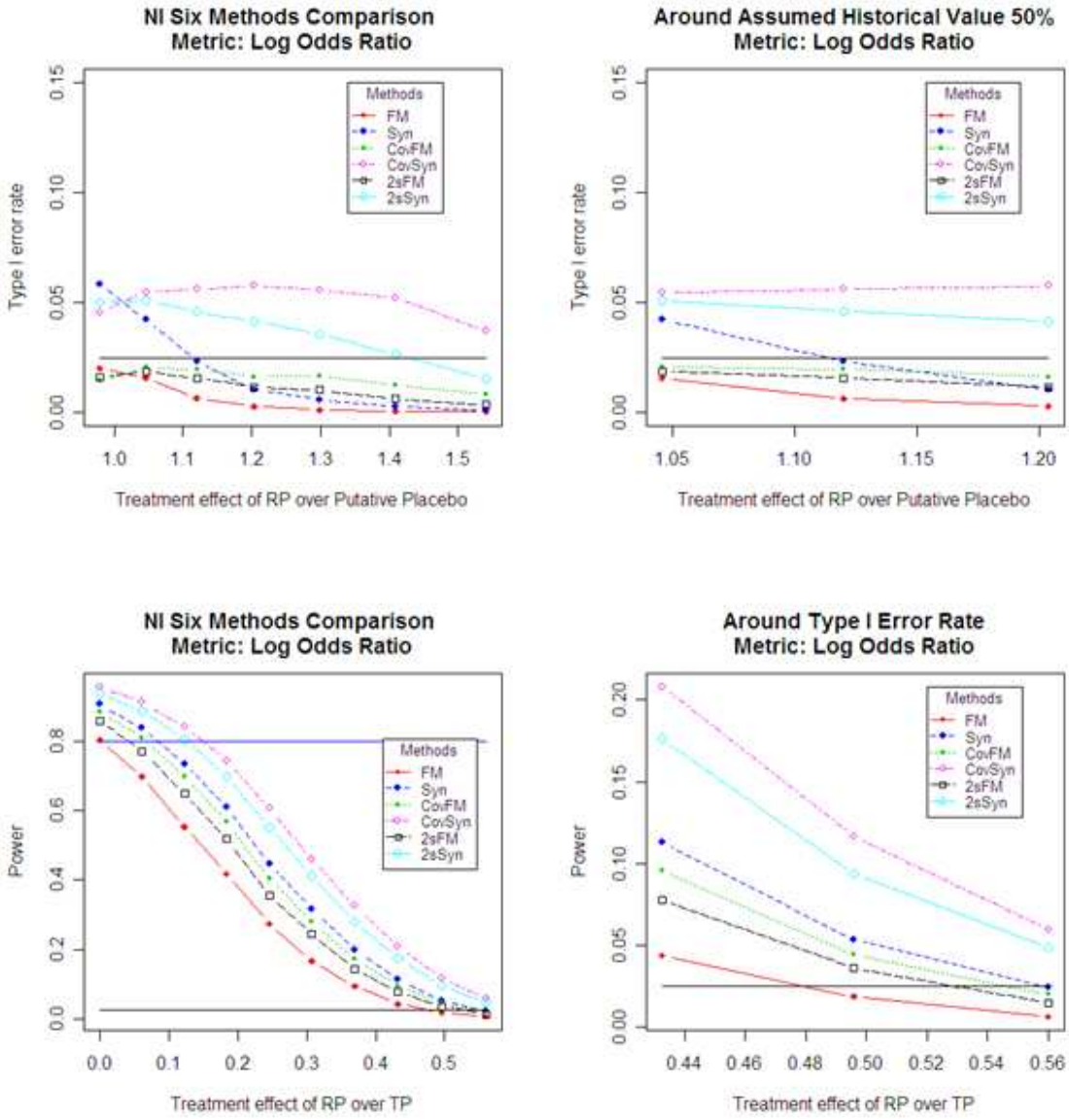


Figure A.7. Performance Characteristics Comparisons--Measure of Association: Difference.

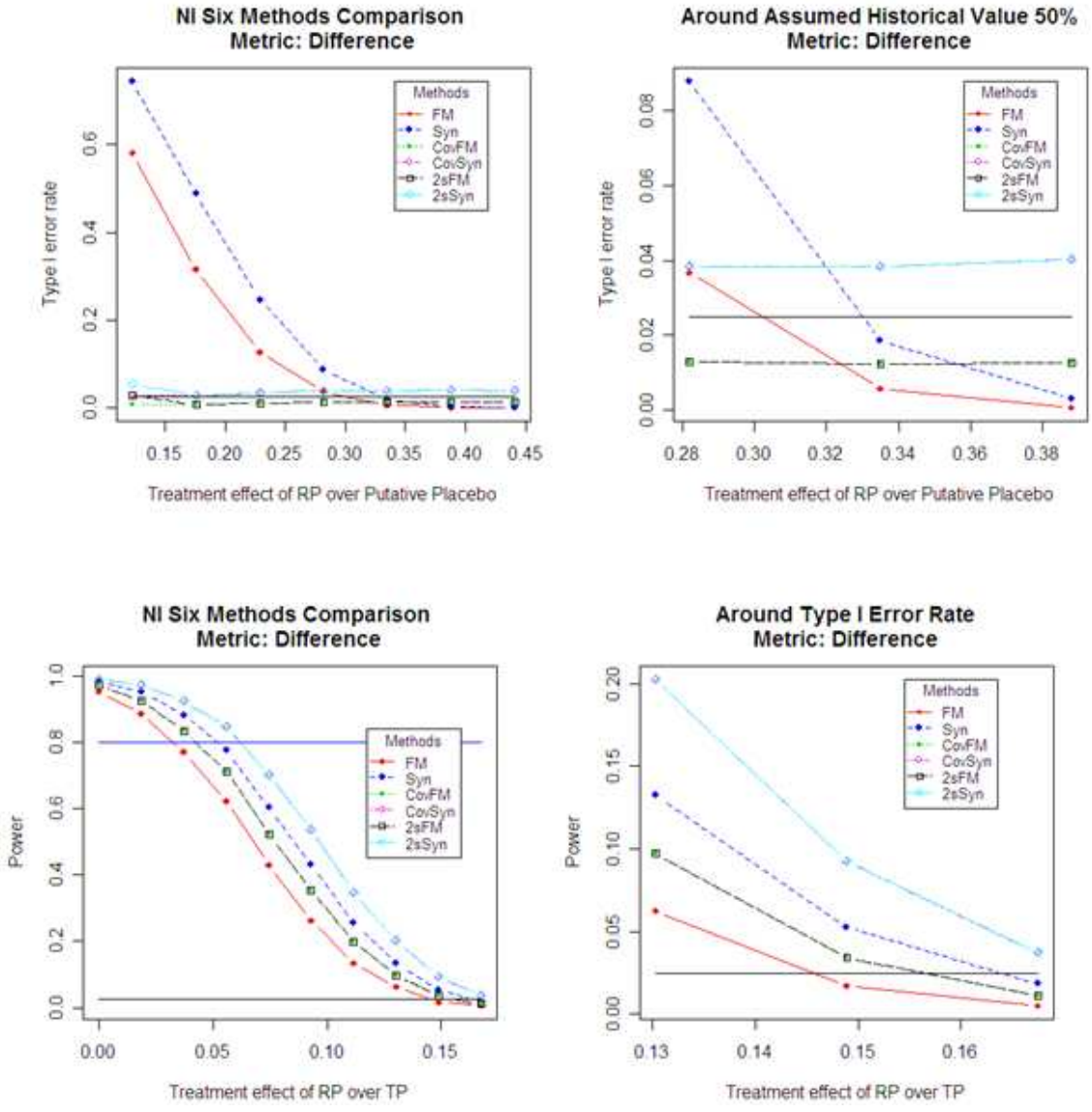


Figure A.8. Performance Characteristics Comparisons--Measure of Association: Log Relative Risk

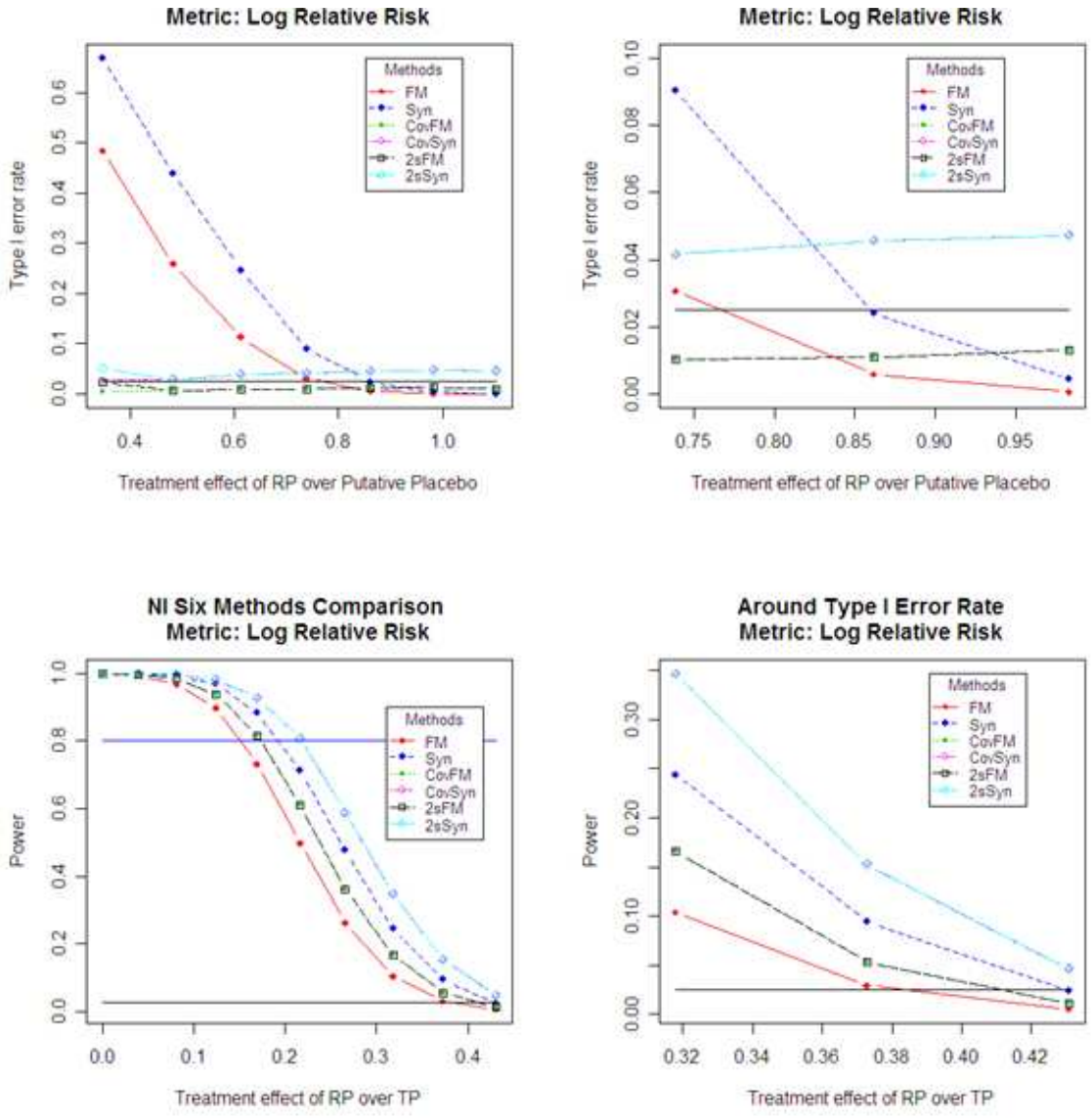


Figure A.9. Performance Characteristics Comparisons--Measure of Association: Log Odds Ratio

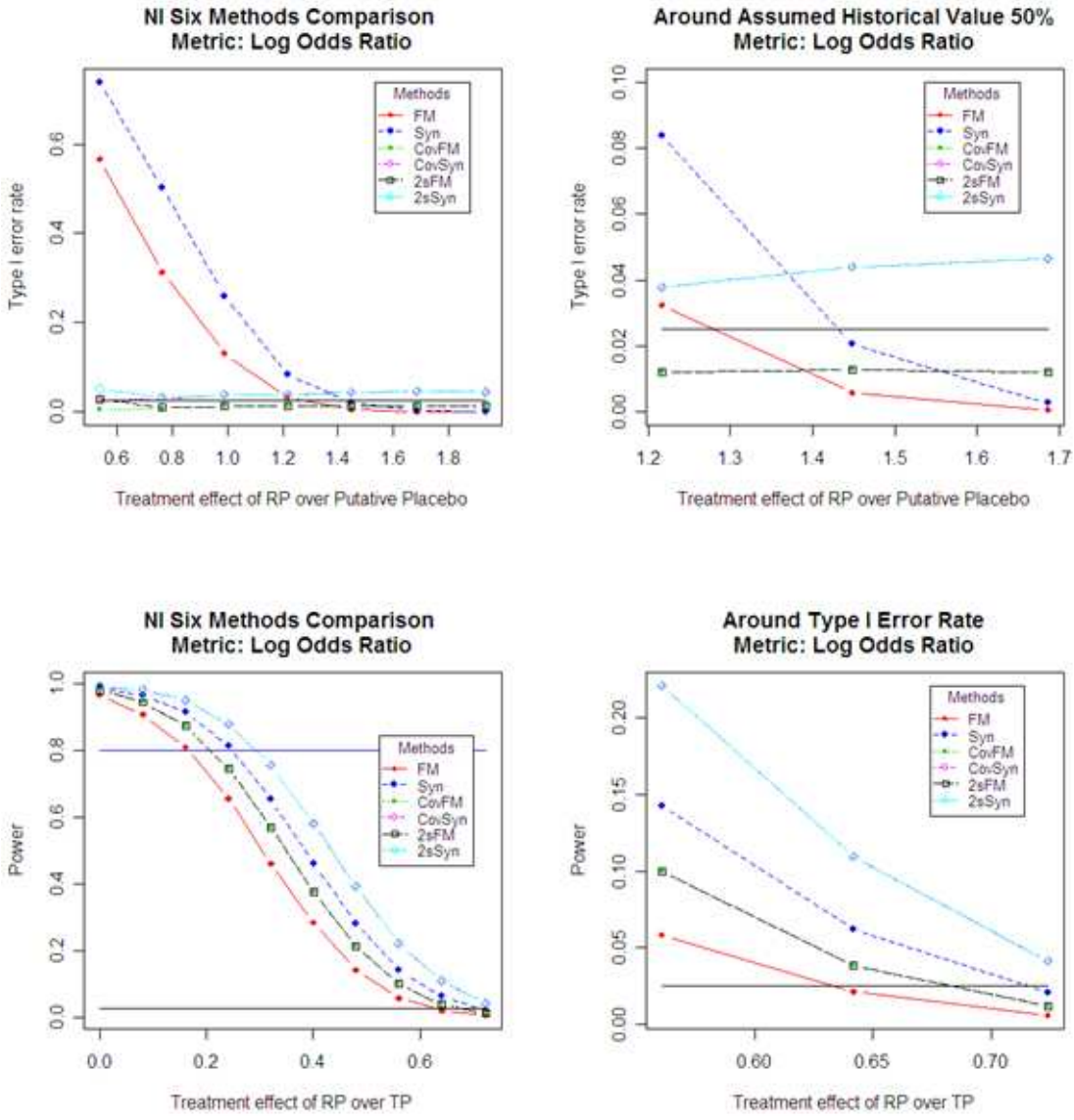


Table A.1. Scenario 1—Type I Error Rate (Bold rows are percents of HER3+/HER2+ 40%, 50% and 60%)

Metric	Perc_HER3+/HER2+	P	T	C	FM	Syn	CovFM	CovSyn	2sFM	2sSyn
Difference	0.1	0.18	0.34	0.5	0.0082	0.0254	0.0056	0.0256	0.0078	0.0253
	0.2	0.18	0.34	0.5	0.0081	0.0241	0.0065	0.0247	0.0076	0.0241
	0.3	0.18	0.34	0.5	0.008	0.0245	0.0077	0.0244	0.0081	0.0244
	0.4	0.18	0.34	0.5	**0.0099**	**0.0256**	**0.0093**	**0.0253**	**0.0098**	**0.0258**
	0.5	0.18	0.34	0.5	**0.009**	**0.0252**	**0.009**	**0.0254**	**0.009**	**0.0252**
	0.6	0.18	0.34	0.5	**0.0087**	**0.0263**	**0.0089**	**0.0268**	**0.0087**	**0.0261**
	0.7	0.18	0.34	0.5	0.0091	0.0254	0.0087	0.0258	0.009	0.0252
LRR	0.1	0.18	0.3	0.5	0.0062	0.0291	0.0038	0.0272	0.0061	0.028
	0.2	0.18	0.3	0.5	0.0054	0.0287	0.0055	0.0268	0.0055	0.0285
	0.3	0.18	0.3	0.5	0.0062	0.0257	0.0055	0.0277	0.0062	0.026
	0.4	0.18	0.3	0.5	**0.0077**	**0.0288**	**0.0076**	**0.0295**	**0.0077**	**0.0291**
	0.5	0.18	0.3	0.5	**0.0073**	**0.0304**	**0.0073**	**0.0309**	**0.0073**	**0.0304**
	0.6	0.18	0.3	0.5	**0.0051**	**0.0247**	**0.0056**	**0.0245**	**0.0051**	**0.0247**
	0.7	0.18	0.3	0.5	0.0078	0.0303	0.0078	0.0302	0.0078	0.0311
LOR	0.1	0.18	0.319043	0.5	0.0086	0.0258	0.0052	0.0264	0.0081	0.0259
	0.2	0.18	0.319043	0.5	0.0072	0.0259	0.0061	0.026	0.0072	0.0259
	0.3	0.18	0.319043	0.5	0.0077	0.0253	0.0069	0.0257	0.0076	0.025
	0.4	0.18	0.319043	0.5	**0.0077**	**0.0253**	**0.0078**	**0.0258**	**0.0077**	**0.0255**
	0.5	0.18	0.319043	0.5	**0.0064**	**0.0234**	**0.0066**	**0.0237**	**0.0064**	**0.0235**
	0.6	0.18	0.319043	0.5	**0.0074**	**0.0259**	**0.0073**	**0.0261**	**0.0073**	**0.0256**
	0.7	0.18	0.319043	0.5	0.0064	0.0247	0.006	0.0244	0.0063	0.0244

Table A.2. Scenario 1—Power.

Metric	P	T	C	FM	Syn	CovFM	CovSyn	2sFM	2sSyn
Difference	0.18	0.482222	0.5	0.8471	0.9244	0.8479	0.9251	0.8472	0.9246
LRR	0.18	0.477778	0.5	0.992	0.9985	0.9921	0.9985	0.992	0.9985
LOR	0.18	0.479894	0.5	0.9333	0.9763	0.934	0.9766	0.9336	0.9764

Table A.3. Scenario 3—Type I error rate (Bold rows are percents of HER3+/HER2+ 40%, 50% and 60%)

Metric	Perc_HER3+/HER2+	P	T	C	FM	Syn	CovFM	CovSyn	2sFM	2sSyn
Difference	0.1	0.157	0.311	0.465	0.0012	0.0052	0.0099	0.0313	0.0039	0.0142
	0.2	0.184	0.332	0.48	0.0022	0.0075	0.012	0.0373	0.0082	0.0223
	0.3	0.211	0.353	0.495	0.0029	0.0093	0.0129	0.0423	0.009	0.0298
	0.4	0.238	0.374	0.51	**0.0061**	**0.0179**	**0.0172**	**0.046**	**0.0136**	**0.0354**
	0.5	0.265	0.395	0.525	**0.0089**	**0.0219**	**0.017**	**0.0453**	**0.0142**	**0.0372**
	0.6	0.292	0.416	0.54	**0.0124**	**0.035**	**0.0178**	**0.0486**	**0.0161**	**0.0437**
	0.7	0.319	0.437	0.555	0.0185	0.046	0.0159	0.0432	0.0168	0.0443
LRR	0.1	0.157	0.270194	0.465	0	1.00E-04	0.0089	0.043	0.0037	0.0155
	0.2	0.184	0.297187	0.48	2.00E-04	7.00E-04	0.0123	0.0611	0.006	0.0306
	0.3	0.211	0.32318	0.495	3.00E-04	0.0029	0.02	0.0739	0.0121	0.0458
	0.4	0.238	0.348396	0.51	**0.002**	**0.0087**	**0.0198**	**0.0761**	**0.0124**	**0.0515**
	0.5	0.265	0.372995	0.525	**0.0055**	**0.0237**	**0.0222**	**0.0757**	**0.0168**	**0.0598**
	0.6	0.292	0.397089	0.54	**0.0171**	**0.0618**	**0.022**	**0.0746**	**0.0197**	**0.0697**
	0.7	0.319	0.420767	0.555	0.0376	0.1108	0.0169	0.0608	0.025	0.0826
LOR	0.1	0.157	0.286903	0.465	3.00E-04	0.001	0.0085	0.0373	0.0035	0.0153
	0.2	0.184	0.313295	0.48	5.00E-04	0.0026	0.0123	0.0521	0.006	0.0264
	0.3	0.211	0.338619	0.495	0.0011	0.0058	0.0167	0.0557	0.0101	0.0355
	0.4	0.238	0.363123	0.51	**0.0028**	**0.0104**	**0.0162**	**0.0575**	**0.0116**	**0.0414**
	0.5	0.265	0.386979	0.525	**0.0064**	**0.0233**	**0.0196**	**0.056**	**0.0156**	**0.0459**
	0.6	0.292	0.410312	0.54	**0.0156**	**0.0421**	**0.0205**	**0.0548**	**0.0185**	**0.0509**
	0.7	0.319	0.433217	0.555	0.0199	0.0584	0.0143	0.0456	0.0162	0.0501

Table A.4. Scenario 3—Power.

Metric	P	T	C	FM	Syn	CovFM	CovSyn	2sFM	2sSyn
Difference	0.265	0.510556	0.525	0.6322	0.7849	0.7401	0.8623	0.7019	0.8369
LRR	0.265	0.508111	0.525	0.8664	0.9549	0.946	0.9874	0.9198	0.9758
LOR	0.265	0.509664	0.525	0.6962	0.8374	0.808	0.9118	0.7707	0.8856

Table A.5. Scenario 4—Type I error rate (Bold rows are percents of HER3+/HER2+ 40%, 50% and 60%)

Metric	Perc_HER3+/HER2+	P	T	C	FM	Syn	CovFM	CovSyn	2sFM	2sSyn
Difference	0.1	0.297	0.3585	0.42	0.5813	0.744	0.0069	0.0284	0.0275	0.0532
	0.2	0.284	0.372	0.46	0.3157	0.4882	0.0075	0.0274	0.0081	0.0284
	0.3	0.271	0.3855	0.5	0.1267	0.2463	0.0106	0.0345	0.0106	0.0345
	0.4	0.258	0.399	0.54	**0.0367**	**0.0877**	**0.0129**	**0.0385**	**0.0129**	**0.0385**
	0.5	0.245	0.4125	0.58	**0.0058**	**0.0186**	**0.0125**	**0.0383**	**0.0125**	**0.0383**
	0.6	0.232	0.426	0.62	**7.00E-04**	**0.0031**	**0.0128**	**0.0403**	**0.0128**	**0.0403**
	0.7	0.219	0.4395	0.66	0	4.00E-04	0.0138	0.0395	0.0138	0.0395
LRR	0.1	0.297	0.353186	0.42	0.4834	0.6683	0.0052	0.0263	0.0233	0.0507
	0.2	0.284	0.361442	0.46	0.2584	0.4391	0.0076	0.029	0.0078	0.0293
	0.3	0.271	0.368103	0.5	0.1133	0.2466	0.0088	0.0386	0.0089	0.0387
	0.4	0.258	0.373256	0.54	**0.0307**	**0.0904**	**0.0104**	**0.0416**	**0.0104**	**0.0416**
	0.5	0.245	0.376962	0.58	**0.0059**	**0.0239**	**0.011**	**0.0457**	**0.011**	**0.0457**
	0.6	0.232	0.379262	0.62	**6.00E-04**	**0.0044**	**0.0132**	**0.0473**	**0.0132**	**0.0473**
	0.7	0.219	0.380184	0.66	0	4.00E-04	0.0102	0.0468	0.0102	0.0468
LOR	0.1	0.297	0.35613	0.42	0.5666	0.7386	0.0058	0.0248	0.0282	0.052
	0.2	0.284	0.367601	0.46	0.3134	0.5031	0.0088	0.0304	0.0094	0.0315
	0.3	0.271	0.378769	0.5	0.1313	0.2601	0.0119	0.0379	0.0119	0.038
	0.4	0.258	0.389831	0.54	**0.0324**	**0.0837**	**0.0121**	**0.0377**	**0.0121**	**0.0377**
	0.5	0.245	0.40099	0.58	**0.0058**	**0.0205**	**0.0128**	**0.0439**	**0.0128**	**0.0439**
	0.6	0.232	0.412473	0.62	**6.00E-04**	**0.0027**	**0.012**	**0.0465**	**0.012**	**0.0465**
	0.7	0.219	0.424554	0.66	2.00E-04	5.00E-04	0.0123	0.0446	0.0123	0.0446

Table A.6. Scenario 4—Power.

Metric	P	T	C	FM	Syn	CovFM	CovSyn	2sFM	2sSyn
Difference	0.245	0.561389	0.58	0.8883	0.9544	0.9268	0.9748	0.9268	0.9748
LRR	0.245	0.55744	0.58	0.9939	0.9992	0.9968	0.9997	0.9968	0.9997
LOR	0.245	0.56011	0.58	0.9086	0.9662	0.9451	0.9826	0.9451	0.9826

Appendix B. Partition the entire likelihood function into *F*-function and *t*-function (3)-(7)

Using the following steps, a maximum likelihood function of δ, γ and ρ can be constructed without the knowledge of true μ and σ.

1. Replacing μ in (3) by its restricted maximum likelihood estimate (rMLE):

$$\hat{\mu}_{rM}(\delta,\gamma,\rho) = \frac{(1-\rho\gamma)\bar{x}+\gamma(\gamma-\rho)\bar{y}-(1-\rho\gamma)\delta}{(1-2\rho\gamma+\gamma^2)} \quad \text{(B1)}$$

it leads to the maximum likelihood function of $\delta, \rho, \gamma, and\ \sigma$,

$$L_{max}(\delta,\rho,\gamma,\sigma|X,Y) \propto \frac{1}{\gamma^n\sigma^{2n}(1-\rho^2)^{n/2}} exp\left(-\frac{1}{2\sigma^2(1-\rho^2)}B\right) \quad \text{(B2)}$$

where B=$\left[\frac{S_x}{\gamma^2}+S_y+\frac{n(\bar{x}-\bar{y}-\delta)^2(1-\rho^2)}{(1-2\rho\gamma+\gamma^2)}-\frac{2\rho}{\gamma}S_{xy}\right]$, $S_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$.

2. Replacing σ^2 in (A2) by its rMLE, $\hat{\sigma}^2_{rM}=\frac{B}{2n(1-\rho^2)}$, the resulting likelihood function is

$$L_{max}(\delta,\rho,\gamma|X,Y) \propto \left(\frac{\gamma\cdot B}{\sqrt{1-\rho^2}}\right)^{-n} = \frac{(1-\rho^2)^{\frac{n}{2}}}{\gamma^n}\left[\frac{S_x}{\gamma^2}+S_y+\frac{n(\bar{x}-\bar{y}-\delta)^2(1-\rho^2)}{(1-2\rho\gamma+\gamma^2)}-\frac{2\rho}{\gamma}S_{xy}\right]^{-n} \quad \text{(B3)}$$

3. Consider the following *F* pivotal quantity:

$$F=\frac{\left(\frac{S_x}{\gamma^2}+S_y-\frac{2\rho}{\gamma}S_{xy}\right)/((2n-2)(1-\rho^2)\sigma^2)}{S_y/((n-1)\sigma^2)}=\frac{\frac{S_x}{\gamma^2}+S_y-\frac{2\rho}{\gamma}S_{xy}}{2S_y(1-\rho^2)} \quad \text{(B4)}$$

Following some algebra, (A3) can be partitioned into two parts:

$$L_{max}(\delta,\rho,\gamma|X,Y) = L_F(\rho,\gamma)\times L_{Fieller}(\delta|\rho,\gamma) \quad \text{(B5)}$$

where $L_F(\rho,\gamma) \propto \frac{(1-\rho^2)^{\frac{n}{2}}}{\gamma^n}\left[\frac{S_x}{\gamma^2}+S_y-\frac{2\rho}{\gamma}S_{xy}\right]^{-n} = \left(\frac{1}{FS_y\gamma\sqrt{1-\rho^2}}\right)^n$,

$$L_{Fieller}(\delta|\rho,\gamma) \propto \left(1+\frac{n(\bar{x}-\bar{y}-\delta)^2(1-\rho^2)}{\left(\frac{S_x}{\gamma^2}+S_y-\frac{2\rho}{\gamma}S_{xy}\right)(1-2\rho\gamma+\gamma^2)}\right)^{-n} = \left(1+\frac{t^2}{2(n-1)}\right)^{-n}$$

and $=\frac{\bar{x}-\bar{y}-\delta}{\hat{\sigma}_{(\bar{x}-\bar{y})}}=\frac{\bar{x}-\bar{y}-\delta}{\sqrt{\frac{(1-2\rho\gamma+\gamma^2)}{n}\hat{\sigma}^2}}$, $\hat{\sigma}^2=\frac{\left(\frac{S_x}{\gamma^2}+S_y-\frac{2\rho}{\gamma}S_{xy}\right)}{2(n-1)(1-\rho^2)}$.

Appendix C. Derivation of the posterior distribution of δ in (14)

The derivation is outlined in the following steps.

1. Consider the change of parameters $(\mu_1, \mu, \gamma, \sigma, \rho)$ in (13) to $(\delta, \mu, u, v, \rho)$ as follows

$$\delta = \mu_1 - \mu, u = \frac{\frac{n}{\sigma^2}}{\frac{n}{\gamma^2\sigma^2} + \frac{n}{\sigma^2}}, v = \frac{n}{\gamma^2\sigma^2} + \frac{n}{\sigma^2}$$

it leads to

$$\mu_1 = \delta + \mu, \gamma = \sqrt{\frac{u}{1-u}}, \sigma = \sqrt{\frac{n}{uv}} \propto (uv)^{-\frac{1}{2}}$$

With some algebra, the posterior distribution of $(\delta, \mu, u, v, \rho)$ can be shown as

$$\pi(\delta, \mu, u, v, \rho|X, Y) \propto u^{\frac{n-1}{2}}(1-u)^{\frac{n-1}{2}} v^n (1-\rho^2)^{-\frac{n+3}{2}} \cdot exp\left\{-\frac{v}{2(1-\rho^2)}\left[S(u,\rho)^2 + (1-\right.\right.$$
$$\left.\left.2\rho(1-u)u)(\mu - M1)2 + M2\right.\right. \quad \text{(C1)}$$

Where $S_x^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$, $S_y^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2$, $S_{xy}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i -$ $\mathcal{Y}$,

$$S(u,\rho)^2 = (1-u)S_x^2 + uS_y^2 - 2\rho\sqrt{(1-u)u}S_{xy}^2,$$

$$M_1 = \frac{\left[(1-u)(\bar{x}-\delta) + u\bar{y} - \rho\sqrt{(1-u)u}\bar{y} - \rho\sqrt{(1-u)u}(\bar{x}-\delta)\right]}{\left[1-2\rho\sqrt{(1-u)u}\right]},$$

and $$M_2 = \frac{(1-\mathrm{u})\mathrm{u}(\bar{\mathrm{x}}-\delta-\bar{\mathrm{y}})^2 + 4\rho^2(1-\mathrm{u})\mathrm{u}(\bar{\mathrm{x}}-\delta)\bar{\mathrm{y}} - \rho^2(1-\mathrm{u})\mathrm{u}(\bar{\mathrm{x}}-\delta+\bar{\mathrm{y}})^2}{1-2\rho\sqrt{(1-u)u}}.$$

Then we have

$$\pi(\delta, \mu, u, v, \rho|X, Y) \propto u^{\frac{n-1}{2}}(1-u)^{\frac{n-1}{2}} v^{I-1}(1-\rho^2)^{-\frac{n+3}{2}} \cdot \exp\left\{-\frac{J}{2}v\right\} \quad \text{(C2)}$$

where $I = n + 1$, and $J = \frac{1}{(1-\rho^2)}\left[S(u,\rho)^2 + (1 - 2\rho\sqrt{(1-u)u})(\mu - M_1)^2 + M_2\right]$.

2. Integrate (B2) with respect to v $(0,+\infty)$, and after some algebra we obtain,

$$\pi(\delta,\mu,u,\rho|X,Y) \propto u^{\frac{n-1}{2}}(1-u)^{\frac{n-1}{2}}(1-\rho^2)^{\frac{n-1}{2}}\left[1-2\rho\sqrt{(1-u)u}\right]^{-1} \cdot (S(u,\rho)^2+M_2)^{-n}$$

$$\cdot\left\{\left(\frac{S(u,\rho)^2+M_2}{1-2\rho\sqrt{(1-u)u}}\right)^{-\frac{1}{2}}\cdot\left[1+\left(\frac{\sqrt{n}(\mu-M_1)}{\sqrt{\frac{S(u,\rho)^2+M_2}{1-2\rho\sqrt{(1-u)u}}}}\right)^2\Big/n\right]^{-\frac{n+1}{2}}\right\}^2$$

3. Integrating $\pi(\delta,\mu,u,\rho|X,Y)$ with respect to μ $(0,+\infty)$, we obtain

$\pi(\delta,u,\rho|X,Y) \propto u^{\frac{n-1}{2}}(1-u)^{\frac{n-1}{2}}(1-\rho^2)^{\frac{n-1}{2}}\left[1-2\rho\sqrt{(1-u)u}\right]^{-1} \cdot (S(u,\rho)^2 + M2)^{-n} \cdot 1-F(t(\delta,u,\rho))2$

where $F(t(\delta,u,\rho))$ is the student's t-distribution function with n degrees of freedom with

$$t(\delta,u,\rho) = \sqrt{\frac{n\cdot\left[1-2\rho\sqrt{(1-u)u}\right]}{[S(u,\rho)^2+M_2]}}\cdot(-M_1)$$

Finally the marginal posterior distribution of δ is

$\pi(\delta,|X,Y) \propto \int_{-1}^{1}\int_{0}^{1} u^{\frac{n-1}{2}}(1-u)^{\frac{n-1}{2}}(1-\rho^2)^{\frac{n-1}{2}}\left[1-2\rho\sqrt{(1-u)u}\right]^{-1} \cdot (S(u,\rho)^2 + M2)^{-n} \cdot 1-F(t(\delta,u,\rho))2 \; dud\rho$

Bibliography

1. Anderson S, Hauck WW. A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics—Theory and Methods* 1983; **12**:2663–2692.
2. Balthasar JP. Bioequivalence and Bioequivalency Testing, *American Journal of Pharmaceutical Education* 1999; **63**: 194-198.
3. Berger JO, Bernardo JM, Sun D. The formal definition of reference priors. *Annals of Statistics* 2009; **37**: 905-938.
4. Berger JO, Sun D. Objective Priors For The Bivariate Normal Model, *Annals of Statistics* 2008; **36(2)**: 963–982, DOI: 10.1214/07-AOS501.
5. Berger RL, Hsu JC. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science* 1996; **11**:283–302.
6. Blackwelder W.C. “Proving the null hypothesis” in clinical trials. *Controlled Clinical Trials* 1982; **3**: 345-353.
7. Blume JD. How to choose a working model for measuring the statistical evidence about a regression parameter. *International Statistical Review* 2005; **73**:351–363.
8. Broeze KA, Opmmer BC, Van Der Veen F, Bossuyt PM, Bhattachwrya S, Mol BW. Inidividual patient data meta-analysis: a promising approach for evidence synthesis in reproductive medicine. *Human Reproduction Update* 2010; **16**: 561-567.

9. CBER/FDA Memorandum. Summary of CBER considerations on selected aspects of active controlled trial design and analysis for the evaluation of thrombolytics in acute MI, June 1999.

10. Chen MH, Shao QM, Ibrahim JG. Monte Carlo Methods in Bayesian Computation. *New York: Springer-Verlag 2000*.

11. Choi L, Caffo B, Rohde C. A survey of the likelihood approach to bioequivalence trials. *Statistics in Medicine* 2008; **27**:4874–4894.

12. Committee for Medicinal Products for Human Use (CHMP) (2006). Guideline on the choice of non-inferiority margin. *Statistics in Medicine* 2006; **25:** 1628-1638

13. D'Agostino RB, Massaro JM, Sullivan L. Non-Inferiority Trials: design concepts and issues-the encounters of academic consultants in statistics. *Statistics in Medicine* 2003; **22:** 169-186.

14. Díaz-Francés E, Sprott, DA. Inference for the ratio of two normal means with unspecified variances. *Biometrical Journal* 2004; **46**: 83–89.

15. Fieller EC. Some problems in interval estimation. *Journal of Royal Statistical Society, Series B* 1954; **16**: 175-185.

16. Fisher LD. Advances in clinical trials in the twentieth century. *Annual Review of Public Health* 1999; **20**: 109-124.

17. FDA Guidance for Industry Bioavailability and Bioequivalence Studies for Orally Administered Drug Products — General Considerations. *2003*

(http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm070124.pdf)

18. Food and Drug Administration. Guidance for industry: non-inferiority clinical trials. *March 2010*.
19. Hasselblad V, Kong DF. Statistical methods for comparison to placebo in active-control trials. *Drug Information Journal* 2001; **35**: 435-449.
20. Hauschke D, Kieser M, Diletti E, Burke, M. Sample size determination for proving equivalence based on the ratio of two means for normally distributed data. *Statistics in Medicine* 1999; **18**: 93-105
21. Holmgren EB. Establishing equivalence by showing that a prespecified percentage of the effect of the active control over placebo is maintained. *Journal of Biopharmaceutical Statistics* 1999; **9**: 651-659.
22. Hua SY, Hawkins DL, Zhou J. Statistical considerations in bioequivalence of two area under the concentration-time curves obtained from serial sampling data, *Journal of Applied Statistics* 2013; **40(5)**: 1140-1154.
23. Hung HMJ, Wang SJ, Tsong Y., Lawrence J., O'Neill RT. Some fundamental issues with non-inferiority testing in active controlled clinical trials. *Statistics in Medicine* 2003; **22**: 213-225.
24. Hung HMJ, Wang SJ, O'Neill R. A regulatory perspective on choice of margin and statistical inference issues in non-inferiority trials. *Biometrical Journal* 2005; **47**: 28-36.

25. Hung HMJ, Wang SJ, O'Neill Robert. Issues with statistical risks for testing methods in noninferiority trial without a placebo arm. *Journal of Biopharmaceutical Statistics* 2007; **17**: 201-213.

26. ICHE-9. International Conference on Harmonisation: Statistical Principles for Clinical Trials. *1998.* http://www.ich.org/LOB/media/MEDIA485.pdf.

27. Julious SA. The ABC of non-inferiority margin setting from indirect comparisons. *Pharmaceutical Statistics* 2011; **10**: 448-453.

28. Julious SA, Wang SJ. Issues with indirect comparisons in clinical trials particularly with respect to non-inferiority trials. *Drug Information Journal* 2008; **42**: 625-633.

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

30. Laster LL, Johnson MF, Kotler ML. Non-inferiority trials: the 'at least as good as' criterion with dichotomous data. *Statistics in Medicine* 2006; **25**: 1115-1130.

31. Liu, J-P, Weng, C-S. Evaluation of log-transformation in assessing bioequivalence. *Communications in Statistics-Theory and Methods* 1994; **23**: 421-434.

32. Locke CS. An exact confidence interval for untransformed data for the ratio of two formulation means. *Journal of Pharmacokinetics and Biopharmaceutics* 1984; **12**:649–655.

33. Marzo A, Bo LD, Rusca A, Zini P. Bioequivalence Of Ticlopidine Hydrochloride Administered In Single Dose To Healthy Volunteers, *Pharmacological Research* 2002; **46(5)**: 401-407.

34. Metzler CM. Bioavailability—a problem in equivalence. *Biometrics* 1974; **30**: 309-317.

35. Nie L, Soon GX. A covariate-adjustment regression model approach to noninferiority margin definition. *Statistics in Medicine* 2010; **29**: 1107-1113.

36. Pocock SJ. Clinical trials: a practical approach. *Wiley, Chichester 1983*.

37. Rothmann M, Li N, Chen G, Chi GYH, Temple R, Tsou HT. Design and analysis of non-inferiority mortality trials in oncology. *Statistics in Medicine* 2003; **22**: 239-264.

38. Royall RM. Statistical Evidence: A Likelihood Paradigm. *Chapman & Hall/CRC: London/Boca Raton, 1997.*

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

40. Rubio FJ and Pérez-Elizalde S. Letter to the Editor. *Biometrical Journal* 2009, **51**: 736–738.

41. Sasabuchi, S. A multivariate one-sided 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, Series A, Mathematics* 1988; **42**: 9-19.

42. Sasabuchi, S. A multivariate one-sided test with composite hypotheses when the covariance matrix is completely unknown. *Memoirs of the Faculty of Science, Kyushu University, Series A, Mathematics* 1988; **42**: 37-46.

43. Schuirmann DJ. 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* 1987; **15**:657–680.

44. Simmonds MC, Higgins JP, Stewart LA, Tierney JF, Clarke MF, Thompson SG. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *Clinical Trials* 2005; **2**: 209-217.

45. Snapinn SM. Alternatives for discounting in the analysis of noninferiority trials. *Journal of Biopharmaceutical Statistics* 2004; **14**: 263-273.

46. Tsong Y, Wang SJ, Hung HMJ, Cui L. Statistical issues on objective, design, and analysis of noninferiority active-controlled clinical trial. *Journal of Biopharmaceutical Statistics* 2003; **13**: 29-41.

47. Wang SJ, Hung HMJ. Assessing treatment efficacy in noninferiority or equivalence studies. *Controlled Clinical Trials* 2002; **23**: 2-14

48. Wang SJ, Hung HMJ. TACT method for non-inferiority testing in active controlled trials. *Statistics in Medicine* 2003; **22**: 227-238.

49. Wang SJ, Hung HMJ, Tsong Y., Utility and pitfalls of some statistical methods in active controlled clinical trials. *Controlled Clinical Trials* 2002; **23:** 15-28.

50. Westlake WJ. Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Sciences* 1972; **61**: 1340-1341.

51. Westlake WJ. Symmetric confidence intervals for bioequivalence trials. *Biometrics* 1976; **32**:741–744.

52. Wiens BL. Choosing an equivalence limit for noninferiority or equivalence studies. *Controlled Clinical Trials* 2002; **23**: 2-14.

Curriculum Vitae