

A more powerful average bioequivalence analysis for the 2×2 crossover

Catalina Stefanescu^{1,*,\dagger} and Devan V. Mehrotra²

¹ *London Business School, Regent's Park, London NW1 4SA, UK*

² *Merck Research Laboratories, Blue Bell, PA 19422, USA*

SUMMARY

The 2×2 crossover is commonly used to establish average bioequivalence of two treatments. In practice, the sample size for this design is calculated under an implicit belief that the true average bioavailabilities of the two treatments are (almost) identical. However, the "standard" average bioequivalence analysis does not reflect this prior belief and this leads to a loss in efficiency. The reason for this loss is that if the true average bioavailabilities are indeed the same, then data from the 2×2 crossover contain additional information that the standard analysis fails to utilize. We propose an alternate average bioequivalence analysis in order to rectify the deficiency in the standard analysis. The validity and significant power advantages of our proposed method are illustrated with simulations and a numerical example with real data is provided.

KEY WORDS: crossover trials; ANCOVA; bootstrap

*Correspondence to: Catalina Stefanescu, London Business School, Regent's Park, London NW1 4SA, UK

\daggerE-mail: cstefanescu@london.edu. Telephone: +44 (0)20 7262 5050. Fax: +44 (0)20 7724 7875

1. INTRODUCTION

The two-treatment, two-period (2×2) crossover trial is routinely used to establish average bioequivalence of two treatments (drugs). In this trial, subjects are randomly assigned to two groups, usually of equal size. Subjects in the first group receive treatment A followed by treatment B (sequence AB), and vice versa for the other group (sequence BA). A suitable washout period is imposed between treatments in order to eliminate potential carryover effects of the first treatment. After administration of each treatment, blood samples are collected at fixed time points, and the concentration of the drug in the blood is quantified. The typical primary endpoint of interest is the area under the drug concentration versus time curve (AUC), which represents the bioavailability of the drug. The two treatments are declared bioequivalent if their true relative average bioavailability is estimated to be within pre-specified "bioequivalence limits" with high confidence.

At the design stage of a 2×2 crossover average bioequivalence trial, there is usually no reason to presume that the true average bioavailability of treatment A is either less or greater than that of treatment B . Accordingly, the required sample size is calculated under an implicit belief that the true average bioavailabilities of the two treatments are (almost) identical. This is often reasonable, but at the end of the trial how should the AUC data be analyzed? Over the last three decades, a plethora of statistical methods for establishing average bioequivalence have been published (see Senn [1] for an up-to-date bibliography). However, all the well known methods, including the current "standard" approach that we describe later, make inefficient use of the available AUC data. The reason for the inefficiency is that if the true average bioavailabilities are indeed the same as presumed at the design stage, then the AUC data contain additional information that existing methods of analysis fail to utilize.

In this paper we develop a novel analytic approach for extracting the additional information contained in the AUC data when the true average bioavailabilities are the same. Our method relies on using the second period response as a covariate, and thereby yields a more powerful analysis.

The rest of this paper is structured as follows. In Section 2 we briefly describe the current standard average bioequivalence analysis for the 2×2 crossover, and develop our proposed analysis. In Section 3 we illustrate the utility of our approach using real data from a clinical trial for which our method supports a conclusion of average bioequivalence while the standard method does not. Results of a simulation study to illustrate the validity and significant power advantages of our proposed method are described in Section 4. Finally, summary remarks and directions for future research are outlined in Section 5.

2. THE 2×2 CROSSOVER TRIAL: AVERAGE BIOEQUIVALENCE

Let n be the number of subjects in each sequence group. Denote by y_{ijk} the $\log_e(AUC)$ under treatment k for subject j within sequence i ($i = 1, 2; j = 1, \dots, n; k = A, B$). We shall assume the following crossover model:

Sequence 1 (AB)

$$\begin{aligned} \text{Period 1 : } y_{1jA} &= \mu + \pi_1 + \mu_A + S_{j(1)} + \varepsilon_{1jA} \\ \text{Period 2 : } y_{1jB} &= \mu + \lambda_A + \pi_2 + \mu_B + S_{j(1)} + \varepsilon_{1jB} \end{aligned} \tag{1}$$

Sequence 2 (BA)

$$\begin{aligned} \text{Period 1 : } y_{2jB} &= \mu + \pi_1 + \mu_B + S_{j(2)} + \varepsilon_{2jB} \\ \text{Period 2 : } y_{2jA} &= \mu + \lambda_B + \pi_2 + \mu_A + S_{j(2)} + \varepsilon_{2jA} \end{aligned}$$

Here λ_A and λ_B represent the carryover effects of treatments A and B respectively, π_i denotes the period effect, and μ_A and μ_B are the treatment effects. We assume that the random subject effects $S_{j(i)}$ are i.i.d. $N(0, \phi_1)$, the random residual effects ε_{ijk} are i.i.d. $N(0, \phi_0)$, and $S_{j(i)}$ and ε_{ijk} are mutually independent. In the following sections we shall also assume that there exists no differential carryover effect, so that $\lambda_A = \lambda_B$.

Let $\delta = \mu_A - \mu_B$ denote the true mean difference between treatments. For $j = 1, \dots, n$, let $d_{1j} = y_{1jA} - y_{1jB}$ for subjects in sequence 1 and $d_{2j} = y_{2jB} - y_{2jA}$ for subjects in sequence 2. Let \bar{d}_i and V_i be the sample mean and variance, respectively, of d_{ij} , $i = 1, 2$. Denote $\hat{\delta} = (\bar{d}_1 - \bar{d}_2)/2$. Note that $\hat{\delta}$ is an unbiased estimator of the true mean difference δ and $V(\hat{\delta}) = \phi_0/n$.

Let δ_0 be the bioequivalence bound. We wish to test the null hypothesis of average bioequivalence

$$H_0 : \delta \leq -\delta_0 \text{ or } \delta \geq \delta_0 \tag{2}$$

versus the alternative of average bioequivalence

$$H_a : -\delta_0 < \delta < \delta_0.$$

2.1. Standard analysis

The standard analysis is given by the two one-sided tests (TOST) approach (Schuirmann [2]).

Let

$$t^{(l)} = \frac{\hat{\delta} + \delta_0}{\sqrt{\widehat{V}(\hat{\delta})}}, \quad t^{(u)} = \frac{\hat{\delta} - \delta_0}{\sqrt{\widehat{V}(\hat{\delta})}}$$

and

$$p^{(l)} = \Pr(t_{2n-2} > t^{(l)}), \quad p^{(u)} = \Pr(t_{2n-2} < -t^{(u)}),$$

where t_n is the central t distribution with n degrees of freedom and $\widehat{V}(\widehat{\delta}) = (V_1 + V_2)/4n$ is an unbiased estimator of ϕ_0/n . Then H_0 is rejected at significance level α if $\max(p^{(l)}, p^{(u)}) < \alpha$.

This is equivalent to rejecting H_0 if

$$\widehat{\delta} > -\delta_0 + t_{2n-2}^\alpha \sqrt{\widehat{V}(\widehat{\delta})} \quad \text{and} \quad \widehat{\delta} < \delta_0 - t_{2n-2}^\alpha \sqrt{\widehat{V}(\widehat{\delta})} \quad (3)$$

simultaneously, where t_n^α is the $100 \cdot (1 - \alpha)$ percentile of the t_n distribution. In practice, α is typically 0.05.

2.2. Proposed analysis

Let x_{ij} be the response in period 2, that is $x_{ij} = y_{ijB}$ for sequence AB and $x_{ij} = y_{ijA}$ for sequence BA . Let

$$\rho = \phi_1(\phi_1 + \phi_0)^{-1} = \text{Corr}(y_{ijA}, y_{ijB}).$$

Observe that for the assumed crossover model (1) we have $\text{Corr}(d_{ij}, x_{ij}) = -\sqrt{0.5(1 - \rho)}$ and $V(d_{ij}|x_{ij}) = 0.5(1 + \rho)V(d_{ij}) < V(d_{ij})$. The fact that $V(d_{ij}|x_{ij}) < V(d_{ij})$ suggests using x_{ij} as a covariate. We therefore fit the following analysis of covariance (ANCOVA) model

$$d_{ij}|x_{ij} = \mu^* + \delta_i + \beta x_{ij} + \varepsilon_{ij}^*, \quad (4)$$

where $\delta_i = (-1)^i \delta$, $\beta = \rho - 1$, and $\varepsilon_{ij}^* \sim N(0, \phi_0(1 + \rho))$. Let

$$\widehat{\delta}^* = \frac{1}{2}(\bar{d}_1 - \bar{d}_2) - \frac{1}{2}\widehat{\beta}(\bar{x}_1 - \bar{x}_2),$$

where $\widehat{\beta}$ is the ordinary least squares estimator given by

$$\widehat{\beta} = \frac{\sum_{i=1}^2 \sum_{j=1}^n (d_{ij} - \bar{d}_i)(x_{ij} - \bar{x}_i)}{\sum_{i=1}^2 \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2}.$$

Under the crossover model (1) we have $E(\widehat{\delta}^*|x_{ij}) = \delta(1 + \rho)/2 = E(\widehat{\delta}^*)$, and

$$V(\widehat{\delta}^*|x_{ij}) = \frac{1}{4}\phi_0(1 + \rho) \left(\frac{2}{n} + \frac{(\bar{x}_1 - \bar{x}_2)^2}{\sum_{i=1}^2 \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2} \right).$$

Note that $E(\widehat{\delta}^*) \approx \delta$ if $\delta \approx 0$ or $\rho \approx 1$. Thus $\widehat{\delta}^*$ is a biased estimator of δ , but the bias is negligible when δ is close to zero (as presumed at the design stage) or when the correlation ρ is large (as is common in practice). From chapter 7 in Fleiss [3], the estimated conditional variance of the ANCOVA estimator is given by

$$\widehat{V}(\widehat{\delta}^* | x_{ij}) = \frac{1}{4(2n-3)} \cdot \frac{SS_{xx}SS_{dd} - SS_{dx}^2}{SS_{xx}} \left(\frac{2}{n} + \frac{(\bar{x}_{1.} - \bar{x}_{2.})^2}{SS_{xx}} \right), \quad (5)$$

where $SS_{xx} = \sum_{i=1}^2 \sum_{j=1}^n (x_{ij} - \bar{x}_{i.})^2$, $SS_{dd} = \sum_{i=1}^2 \sum_{j=1}^n (d_{ij} - \bar{d}_{i.})^2$, and $SS_{dx} = \sum_{i=1}^2 \sum_{j=1}^n (d_{ij} - \bar{d}_{i.})(x_{ij} - \bar{x}_{i.})$.

We develop an adaptive method of testing the null hypothesis (2). Specifically, we propose a variation of the TOST approach in which the test statistic employs the ANCOVA estimator $\widehat{\delta}^*$ when $\widehat{\delta}$ is “close” to zero, and the standard estimator $\widehat{\delta}$ when $\widehat{\delta}$ is “far” from zero. The motivation behind this method is the following: since $\widehat{\delta}$ is unbiased for δ , values of $\widehat{\delta}$ close to zero suggest that δ is close to zero as well. In this case the estimator $\widehat{\delta}^*$ could be used for testing because its bias is very small and its variance is smaller than the variance of $\widehat{\delta}$, leading to increased power. On the other hand, values of $\widehat{\delta}$ far from zero suggest that δ is far from zero. Then the bias of $\widehat{\delta}^*$ is too large and the standard estimator $\widehat{\delta}$ should instead be used for testing.

In order to control the test size and maximize the power, we propose use of “tuning parameters” ε , ψ_1 and ψ_2 , where $0 \leq \varepsilon, \psi_1, \psi_2 \leq \delta_0$. The adaptive method leads to the following rejection rule:

IF $|\widehat{\delta}| > \varepsilon$ THEN reject H_0 if

$$\widehat{\delta} > -\delta_0 + t_{2n-2}^\alpha \sqrt{\widehat{V}(\widehat{\delta})} + \psi_2$$

and

$$\widehat{\delta} < \delta_0 - t_{2n-2}^\alpha \sqrt{\widehat{V}(\widehat{\delta})} - \psi_2$$

ELSE reject H_0 if

$$\widehat{\delta}^* > -\delta_0 + t_{2n-3}^\alpha \sqrt{\widehat{V}(\widehat{\delta}^*|x_{ij})} - \psi_1$$

and

$$\widehat{\delta}^* < \delta_0 - t_{2n-3}^\alpha \sqrt{\widehat{V}(\widehat{\delta}^*|x_{ij})} + \psi_1$$

The parameter ε quantifies what "close to zero" means in the previous paragraph and serves as the threshold between using $\widehat{\delta}$ or $\widehat{\delta}^*$ for testing. The parameter ψ_2 has the role of controlling the type I error rate. If $\psi_2 > 0$, it is harder to reject H_0 when H_0 is true and $\widehat{\delta}$ is used for testing. Conversely, the parameter ψ_1 has the role of increasing power. If $\psi_1 > 0$, it is easier to reject H_0 when H_0 is false and $\widehat{\delta}^*$ is used for testing. Thus ψ_1 and ψ_2 have the joint effect of maximizing power while preserving the test size.

Note that, for any ψ_1 , the choice $(\varepsilon = 0, \psi_1, \psi_2 = 0)$ leads to the standard analysis for which the test size is α . Therefore there will always exist triplets $(\varepsilon, \psi_1, \psi_2)$ for which the test size of the adaptive method is controlled at α . From among all such triplets, we determine $(\varepsilon, \psi_1, \psi_2)$ for which the power of the adaptive method is maximum, using a parametric bootstrap procedure (Efron and Tibshirani [4]). Resampling is done under the null hypothesis $\delta = \delta_0$ for test size, and under the alternative $\delta = 0$ for power, from a bivariate normal distribution

with the observed correlation structure of the data. Let $(\varepsilon, \psi_1, \psi_2)$ be such that, based on B bootstrap samples, the estimated test size at $(\varepsilon, \psi_1, \psi_2)$ is smaller than $\alpha + \sqrt{B^{-1}\alpha(1-\alpha)}$, and the power at $(\varepsilon, \psi_1, \psi_2)$ is maximum. We recommend using $B = 5000$. Denote

$$\begin{aligned} t_{\psi_2}^{(l)} &= \frac{\widehat{\delta} + \delta_0 - \psi_2}{\sqrt{\widehat{V}(\widehat{\delta})}}, & p_{\psi_2}^{(l)} &= \Pr(t_{2n-2} > t_{\psi_2}^{(l)}), \\ t_{\psi_2}^{(u)} &= \frac{\widehat{\delta} - \delta_0 + \psi_2}{\sqrt{\widehat{V}(\widehat{\delta})}}, & p_{\psi_2}^{(u)} &= \Pr(t_{2n-2} < -t_{\psi_2}^{(u)}), \\ t_{\psi_1}^{*(l)} &= \frac{\widehat{\delta}^* + \delta_0 + \psi_1}{\sqrt{\widehat{V}(\widehat{\delta}^*|x_{ij})}}, & p_{\psi_1}^{*(l)} &= \Pr(t_{2n-3} > t_{\psi_1}^{*(l)}), \\ t_{\psi_1}^{*(u)} &= \frac{\widehat{\delta}^* - \delta_0 - \psi_1}{\sqrt{\widehat{V}(\widehat{\delta}^*|x_{ij})}}, & p_{\psi_1}^{*(u)} &= \Pr(t_{2n-3} < -t_{\psi_1}^{*(u)}). \end{aligned}$$

Also, let $\lambda_\varepsilon = 1$ if $|\widehat{\delta}| \leq \varepsilon$, and 0 otherwise. Then the p-value of the test corresponding to $(\varepsilon, \psi_1, \psi_2)$ is given by

$$\lambda_\varepsilon \max(p_{\psi_1}^{*(l)}, p_{\psi_1}^{*(u)}) + (1 - \lambda_\varepsilon) \max(p_{\psi_2}^{(l)}, p_{\psi_2}^{(u)}). \quad (6)$$

The conclusion of average bioequivalence is reached if the p-value in (6) is less than α .

3. ILLUSTRATIVE EXAMPLE

To illustrate the adaptive method proposed in Section 2.2, we analyze data from a bioequivalence clinical trial (Bradstreet [5]). Table I contains $\log_e(AUC)$ data from a 2×2 crossover study with $n = 13$ subjects per sequence. For this data set we have $\widehat{\delta} = 0.111$ and $\widehat{V}(\widehat{\delta}) = 0.005$. Using the bioequivalence bound $\delta_0 = \log_e(1.25)$, the p-value for the TOST approach in Section 2.1 is $0.0672 > 0.05$. Hence using the standard method we fail to reject

Table I. $\log_e(AUC)$ data from a 2×2 crossover study (Bradstreet [5]).

Sequence 1 (<i>AB</i>)			Sequence 2 (<i>BA</i>)		
<i>A</i>	<i>B</i> (<i>x</i>)	<i>A</i> – <i>B</i> (<i>d</i>)	<i>B</i>	<i>A</i> (<i>x</i>)	<i>B</i> – <i>A</i> (<i>d</i>)
7.506	7.263	0.243	7.301	7.561	-0.260
6.868	7.198	-0.330	7.625	7.800	-0.175
5.714	5.764	-0.050	5.532	5.898	-0.366
6.433	6.075	0.358	6.841	7.074	-0.233
6.749	6.594	0.155	6.974	7.355	-0.381
6.198	6.311	-0.113	8.591	8.740	-0.149
6.720	6.827	-0.107	7.028	6.363	0.665
6.892	7.322	-0.430	6.590	6.993	-0.403
6.285	6.792	-0.507	6.191	6.210	-0.019
6.361	6.286	0.075	5.906	6.452	-0.546
7.181	6.340	0.841	6.854	7.013	-0.159
6.242	6.632	-0.390	5.702	6.694	-0.992
7.537	7.619	-0.082	5.980	6.187	-0.207

H_0 and don't conclude bioequivalence.

The ANCOVA estimator for this data set is $\hat{\delta}^* = 0.100$ with estimated conditional variance $\hat{V}(\hat{\delta}^* | x_{ij}) = 0.005$ calculated using (5). For different values of $(\varepsilon, \psi_1, \psi_2)$, Table II reports the test size and power estimated from 5000 bootstrap samples. The test size is included in parentheses if it is greater than .053, i.e. more than one standard error larger than .05 based on 5000 bootstrap samples. The values $\varepsilon = .113$, $\psi_1 = .02$ and $\psi_2 = .02$ lead to the maximum power .8560 while controlling for the test size. From (6) it follows that the p-value of the

Table II. Estimated test size and power of the adaptive test for the data in Table 1, at different values of ε , ψ_1 and ψ_2 . Estimation based on $B = 5000$ bootstrap samples. Optimal values of $(\varepsilon, \psi_1, \psi_2)$ are in boldface.

ε	ψ_1	ψ_2	Estimated test size	Estimated power
.000	.00	.00	.0472	.8216
⋮	⋮	⋮	⋮	⋮
.113	.02	.01	(.0544)	.8566
.113	.02	.02	.0528	.8560
⋮	⋮	⋮	⋮	⋮
.113	.03	.01	(.0586)	.8686
.113	.03	.02	(.0570)	.8680
⋮	⋮	⋮	⋮	⋮
.114	.01	.01	.0522	.8390
.114	.01	.02	.0506	.8388
⋮	⋮	⋮	⋮	⋮
.224	.00	.00	(.0738)	.8344

test corresponding to this choice of tuning parameters is $0.0328 < 0.05$. Hence the adaptive approach enables us to reject the null hypothesis and conclude bioequivalence.

4. SIMULATION RESULTS

We conducted a simulation study in order to compare the performance of our proposed adaptive method with that of the standard approach. Data were generated for a 2×2 crossover trial

Table III. Probability of concluding bioequivalence (%) for the standard and adaptive methods at different values of the true treatment difference δ and intrasubject correlation ρ . Nominal $\alpha = 5\%$.

True δ	Standard method	Adaptive method	
		$\rho = 0.2$	$\rho = 0.8$
$.224 = \delta_0$	5.0	4.8	5.2
$.112 = \delta_0/2$	41.4	48.5	49.5
$.056 = \delta_0/4$	66.9	70.6	73.6
$.028 = \delta_0/8$	75.1	80.0	81.0
.000	77.9	81.9	82.7

with $n = 8$ subjects per sequence, from bivariate normal populations with $\phi_0 = .044$ and correlation coefficients $\rho = 0.2$ and $\rho = 0.8$. Since $\rho = \phi_1(\phi_1 + \phi_0)^{-1}$, these two values of ρ correspond to $\phi_1 = 0.011$ and $\phi_1 = 0.176$ respectively. The true treatment difference δ was assigned five values between 0 and $\delta_0 = .224(\approx \log_e(1.25))$.

Table III reports the probabilities of concluding bioequivalence with the standard and adaptive approaches, for different values of δ and ρ . For the standard method these are exact probabilities and they do not depend on the correlation ρ . For the adaptive method these are empirical probabilities based on 1000 simulated samples. Both methods control the test size at α . However, as theorized, the power of the adaptive method is notably higher than that of the standard approach.

5. DISCUSSION AND FUTURE RESEARCH

In this paper, we have explained why the standard method for establishing average bioequivalence is inefficient, proposed a principled alternate method of analysis, and quantified the gains in efficiency (power) of our method using simulations. Our proposed adaptive method relies on computing the test statistics with the ANCOVA estimator when, based on the unbiased standard estimator, the data suggest that the true treatment difference is close to zero. The attractiveness of the ANCOVA estimator in this case stems from the fact that its variance can be substantially smaller than the variance of the standard estimator, thereby yielding an increase in power. Conversely, in order to control the test size, the adaptive approach uses the standard estimator for testing whenever the data suggest that the true treatment difference is far from zero. We showed empirically that this adaptive approach controls the type I error rate and yields relatively high power for hypothesis testing, as expected.

The proposed adaptive testing methodology is easy to implement. The program for the bootstrap sampling procedure is currently written in SAS and the computation time is about three minutes for the crossover data discussed in Section 3. Although bootstrap sampling is usually computationally time consuming, in our experience the computation time remains reasonable even for larger values of the sample size for each sequence group. A SAS program for the proposed adaptive analysis is available from the authors upon request.

It is interesting to note that Longford [6] has described a novel approach for combining an unbiased and a biased estimator of a parameter in the context of 2×2 crossover trials with potential carryover effects. A similar combination of the unbiased standard estimator and the (potentially) biased ANCOVA estimator may lead to a synthetic estimator with improved

performance, which we are currently investigating.

Finally, the methodology of adaptive testing developed in this paper can be extended to cover higher order crossover designs, as well as population and individual bioequivalence. These and other generalizations are the subject of further research.

ACKNOWLEDGEMENTS

The authors are grateful to Tom Bradstreet for providing the crossover data used in Section 3.

REFERENCES

1. Senn SJ. *Cross-over Trials in Clinical Research*. Wiley, 2002.
2. Schuirmann DJ. A comparison of the 2 one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics* 1987; **15**, 657–680.
3. Fleiss JL. *The Design and Analysis of Clinical Experiments*. Wiley, 1986.
4. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman and Hall, 1994.
5. Bradstreet TE. (1994). Favorite data sets from early phases of drug research, Part 3. *Proceedings of the Statistical Education Section of the American Statistical Association*, Toronto, Canada, 247–252.
6. Longford NT. Synthetic estimators with moderating influence: the carry-over in cross-over trials revisited. *Statistics in Medicine* 2001; **20**, 3189–3203.