Research Article

Sequential Bioequivalence Approaches for Parallel Designs

Anders Fuglsang^{1,2}

Received 17 March 2013; accepted 22 January 2014; published online 14 February 2014

Abstract. Regulators in EU, USA and Canada allow the use of two-stage approaches for evaluation of bioequivalence. The purpose of this paper is to evaluate such designs for parallel groups using trial simulations. The methods developed by Diane Potvin and co-workers were adapted to parallel designs. Trials were simulated and evaluated on basis of either equal or unequal variances between treatment groups. Methods B and C of Potvin *et al.*, when adapted for parallel designs, protected well against type I error rate inflation under all of the simulated scenarios. Performance characteristics of the new parallel design methods showed little dependence on the assumption of equality of the test and reference variances. This is the first paper to describe the performance of two-stage approaches for parallel designs used to evaluate bioequivalence. The results may prove useful to sponsors developing formulations where crossover designs for bioequivalence evaluation are undesirable.

KEY WORDS: bioequivalence; parallel; power; sequential designs; type I errors.

INTRODUCTION

Approval of generics often depends on *in vivo* proof of bioequivalence (BE). The most common way to demonstrate BE is to compare the pharmacokinetics of the generic formulation (test, T) to that of the originator (reference, R). The accepted primary pharmacokinetic metrics are area under the concentration–time curve up to the last sampling point (AUC_t) and the maximum observed concentration ($C_{\max}$). Typically, a 90% confidence interval for the T/R on the observed scale (geometric mean ratio) is constructed *via* two one-sided *t* tests for the T–R difference for these metrics on the logarithmic scale, and the general acceptance range is 80–125% in EU, USA and Canada (1–3). Exceptions involve narrow therapeutic index drugs when the acceptance range will sometimes need to be tightened.

Power, defined as the developer's chance of showing bioequivalence for a given sample size, depends on the ratio of the population geometric means of the test and reference products, as well as their population variances. These values can, however, be difficult to estimate prior to conduct of a pivotal trial, and consequently, the sample size may be difficult or impossible to calculate. In such cases, it may be desirable to use a two-stage approach, that is, to evaluate a limited number of subjects first and then use the initially acquired information to include additional subjects and do a final pooled analysis. In such cases, regulators require that the

type I error rate (overall alpha; risk of concluding bioequivalence when the two formulation are not bioequivalent) be controlled. The traditional limit for overall alpha is 5%. Potvin *et al.* developed two algorithms which were shown to protect against inflation of the overall type I error rate under most circumstances for two-sequence, two-treatment crossover designs at test/reference ratios of 0.95 (4). Later, Montague *et al.* extended the methodology to situations where the test/reference ratio is 0.90 (5). To date, no publication has described two-stage approaches for parallel designs which are also commonly used and acceptable to regulators. The purpose of this paper is therefore to adapt Potvin's method to parallel designs and to evaluate the performance in terms of type I error rates and power.

MATERIALS AND METHODS

Software

Software implementing parallel design adaptations of methods B and C from Potvin *et al.* were programmed by the author in the programming language C. The executable uses the Mersenne–Twister algorithm to generate pseudo-random numbers with uniform distribution (6) and relies on a further Box–Muller transform for generation of pseudo-random Gaussian numbers used as the simulated log-transformed pharmacokinetic responses.

The software uses equal sample sizes of test and reference at stage 1, and equal sample sizes of the two formulations at stage 2 but sample sizes at stage 1 and stage 2 are not necessarily equal. No stage effect was simulated.

Note that throughout this paper, the term N_1 denotes the sample size at stage 1, where $\frac{1}{2}N_1$ are dosed the test formulation and the other $\frac{1}{2}N_1$ the reference treatment.

Electronic supplementary material The online version of this article (doi:10.1208/s12248-014-9571-1) contains supplementary material, which is available to authorized users.

¹ Hiort Lorensens Vej 6c st. tv., 6100, Haderslev, Denmark.

² To whom correspondence should be addressed. (e-mail: a.fuglsang@ymail.com)

Along the same lines, $N_1 + N_2 = N_{\text{tot}}$ denotes the total number of subjects in the trial, half of which receive the test treatment and the other half the reference treatment.

Adaptation to Parallel Designs

The decision trees/algorithms for methods B and C in the work of Potvin *et al.* are applied without modification, but since this work involves simulations and evaluation for parallel designs, another set of equations for sample size calculation and evaluation of bioequivalence must be applied. Potvin's methods involve a single variability estimate for the dimensioning of stage 2 due to the crossover nature of the studied designs. There is to the best of my knowledge no method available which dimensions a bioequivalence trial following a parallel design on the basis of unequal variances for the test and reference treatments. For the purpose of calculating power after stage 1 and/or dimensioning the second stage, the software therefore relies on calculation of a pooled variance estimate from estimates of variances associated with the two treatments at stage 1:

$$\hat{V}_{\text{pool}} = \frac{(^{1/2}N_1 - 1)\ddot{V}_T + (^{1/2}N_1 - 1)\ddot{V}_R}{N_1 - 2}$$

where $\ddot{V}_T$ and $\ddot{V}_R$ are the sample variance estimates associated with test and reference. Note that when $N_T = N_R$ as throughout this work, the pooled variance is simply the average of the two variance estimates.

Power and sample size after stage 1 is then calculated *via*

$$1 - \beta = F_t \left(\frac{\ln(1.25/\Theta)}{s\sqrt{2/\left(\frac{1}{2}N_1\right)}} - t_{1-\alpha, df} \right) - F_t \left(\frac{-\ln(1.25\Theta)}{s\sqrt{2/\left(\frac{1}{2}N_1\right)}} + t_{1-\alpha, df} \right)$$

where Θ is the point estimate, β is the type II error, s is the estimated standard deviation (square root of the pooled error variance estimate), F_t is the cumulative t distribution function and $t_{1-\alpha, df}$ denotes the critical value of the t distribution with df degrees of freedom at the $1-\alpha$ probability level. The equation for power is same as the one given in the paper by Potvin *et al.* with adaptation to a parallel design *via* a design constant as given the Power.TOST package for R (7). The total sample size required (stages 1 and 2 combined) for the desired 80% power was estimated from the stage 1 data by increasing the value of N (and the corresponding value of df) iteratively, until the smallest even N yielding at least the desired power was found.

One way to evaluate bioequivalence would be to evaluate the residual on basis of a normal linear model with the fixed terms *treatment* and *stage*, the latter only being relevant where two stages are actually carried out. The underlying assumption of this approach is that the variabilities are equal. The software implements this option, using the following equations:

Let Y_{ij} denote the logarithmised pharmacokinetic data point from the i th subject ($i \in \{1, 2, 3 \dots \frac{1}{2}(N_1 + N_2)\}$); it follows that any subject indexed by $i \leq \frac{1}{2}N_1$ is in stage 1; otherwise,

the subject is in stage 2) in the j th treatment group ($j \in \{1, 2\}$); in the following, $j=1$ corresponds to the test treatment and $j=2$ to the reference treatment).

The grand mean is:

$$\bar{Y}_{..} = \frac{\sum_{i=1}^{^{1/2}(N_1 + N_2)} \sum_{j=1}^2 Y_{ij}}{(N_1 + N_2)}$$

The total sum of squares is

$$SS_{\text{total}} = \sum_{i=1}^{^{1/2}(N_1 + N_2)} \sum_{j=1}^2 \left(\bar{Y}_{..} - Y_{ij} \right)^2$$

Mean of the test treatment ($j=1$) is:

$$\bar{Y}_{.1} = \frac{\sum_{i=1}^{^{1/2}(N_1 + N_2)} Y_{i1}}{^{1/2}(N_1 + N_2)}$$

Mean of the reference treatment ($j=2$) is:

$$\bar{Y}_{.2} = \frac{\sum_{i=1}^{^{1/2}(N_1 + N_2)} Y_{i2}}{^{1/2}(N_1 + N_2)}$$

Mean of stage 1 is given by:

$$\bar{Y}_{\text{stg1}} = \frac{\sum_{i=1}^{^{1/2}N_1} \sum_{j=1}^2 Y_{ij}}{N_1}$$

Mean of stage 2 is given by:

$$\bar{Y}_{\text{stg2}} = \frac{\sum_{i=1+^{1/2}N_1}^{^{1/2}(N_1 + N_2)} \sum_{j=1}^2 Y_{ij}}{N_2}$$

The sums of squares accounted for treatment as a fixed factor becomes:

$$SS_{\text{trt}} = SS_{\text{total}} - \sum_{i=1}^{^{1/2}(N_1 + N_2)} \left(\left(\bar{Y}_{.1} - Y_{i1} \right)^2 + \left(\bar{Y}_{.2} - Y_{i2} \right)^2 \right)$$

The sum of squares accounted for stage as a fixed factor becomes:

$$SS_{\text{stg}} = SS_{\text{total}} - \sum_{i=1}^{^{1/2}N_1} \sum_{j=1}^2 \left(\bar{Y}_{\text{stg1}} - Y_{ij} \right)^2 - \sum_{i=1+^{1/2}N_1}^{^{1/2}(N_1 + N_2)} \sum_{j=1}^2 \left(\bar{Y}_{\text{stg2}} - Y_{ij} \right)^2$$

The residual sum of squares in the ANOVA having treatment and stage as factors is then:

$$SS_{\text{res}} = SS_{\text{total}} - SS_{\text{trt}} - SS_{\text{stg}}$$

and is associated with $(N_1 + N_2 - 3)$ degrees of freedom when two stages are employed, or $(N_1 + N_2 - 2)$ degrees of freedom when only one stage is conducted, in which case $N_2 = 0$ and calculation of SS_{stg} is omitted. Note that the equations apply only to the balanced case when equal numbers receive test and reference at stage 1 and when equal numbers receive test and reference at stage 2. The confidence interval for the test-reference difference on the log scale (lower limit, L ; upper limit, U) at a given alpha is then given by:

$$L = (\bar{Y}_{.1} - \bar{Y}_{.2}) - t_{(1-\alpha, df)} \times SE$$

$$U = (\bar{Y}_{.1} - \bar{Y}_{.2}) + t_{(1-\alpha, df)} \times SE$$

where SE is

$$SE = \sqrt{(2\hat{V}/n)}$$

with $n = \frac{1}{2}(N_1 + N_2)$ and where $\hat{V}$ is the pooled variance estimate.

Finally, L and U are exponentiated to obtain the confidence interval for the geometric mean ratio on the observed scale. With parallel designs, it may be desirable to take into account that the test and reference formulations are not associated with the same level of variability when the evaluation of bioequivalence takes place. In fact it is an explicit requirement in EU and US not to assume equal variances. Since evaluation of bioequivalence is based on t test approaches, this can be achieved using the Welch-Satterthwaite correction for degrees of freedom (8,9), but one potential drawback of this method is that stage as a factor cannot be readily included in the model for reduction of the variability used to derive the confidence interval for the primary metrics. The software also implements the Welch-Satterthwaite approach as an option where the adjusted degrees of freedom becomes:

$$df = \frac{\left(\left(\hat{V}_T/n \right) + \left(\hat{V}_R/n \right) \right)^2}{\left(\hat{V}_T^2 / (n^2(n-1)) \right) + \left(\hat{V}_R^2 / (n^2(n-1)) \right)}$$

($n = \frac{1}{2}(N_1 + N_2)$) and where $\hat{V}_T$ and $\hat{V}_R$ are the sample variance estimates associated with test and reference, respectively, and where the SE is:

$$SE = \sqrt{\left(\hat{V}_T/n \right) + \left(\hat{V}_R/n \right)}$$

Bioequivalence evaluations were validated against the `lm`, `anova` and `t.test` functions of the statistical package R version

3.0.2, and power/dimensioning functions were validated against R's add-on PowerTOST package (7).

Simulated Scenarios

One million trials were simulated for each scenario. In all cases, stage 1 power and stage 2 sample size estimates were based on an assumed geometric mean ratio of 0.95, and a target power of 80% as was done in the original paper by Potvin *et al.*; furthermore, the same scenarios were tested with the higher power target of 90%.

The normal linear model approach with *stage* and *treatment* as fixed factors was compared to the Welch-Satterthwaite approach. Since the results indicate that the two evaluation methods perform almost identically (see next section), the subsequently simulated scenarios were based on the Welch-Satterthwaite approach. Performance was evaluated using CVs from 0.1 (10%) to 1.0 (100%) for initial sample sizes (N_1) of 12, 24, 36, 48, 60, 72, 84, 96, 120 and 180. In these scenarios, the CV for test was set to be equal to the CV for reference. A range of simulations were carried out to test if this is appropriate. This was done by using various levels of (CV_T , CV_R) corresponding to an overall CV (backcalculated from the pooled variance) of exactly 0.4 (40%) and 0.7 (70%). To illustrate this principle, the (CV_T , CV_R) vectors for CV=0.4 (40%) were:

(0.28, 0.49777)
(0.32, 0.46969)
(0.36, 0.43729)
(0.40, 0.40)
(0.44, 0.35686)
(0.48, 0.30599)
(0.52, 0.24330)

Examples: When both CV_T and CV_R are 0.40, the corresponding variances V_T and V_R are both 0.14848, so the pooled variance is 0.14828. When $CV_T = 0.28$ and $CV_R = 0.49777$ (here rounded to five decimals; un-truncated double precision floats were used throughout), then $V_T = 0.075478$ and $V_R = 0.22136$ so the pooled variance is also 0.14828 corresponding to a CV of 0.40 (40%). The (CV_T , CV_R) vectors for a CV of 0.7 were:

(0.4, 0.95597)
(0.5, 0.88095)
(0.6, 0.79525)
(0.7, 0.7)
(0.8, 0.59474)
(0.9, 0.47500)
(1, 0.33174)

The values were calculated *via*

$$V_R = 2V - V_T$$

To convert between a variance V on the log scale and a CV on the observed scale, the following relation applies:

$$CV = \sqrt{(e^V - 1)} \text{ or } V = \log(1 + CV^2)$$

where V is the pooled variance. Example: We aim for a CV of 0.7 corresponding to a pooled variance of $V = \log(1 + 0.7^2) =$

Fig. 1. Type I error rate (*left*) and power (*right*) for Potvin's method C adapted to parallel two-stage designs at $N_1=48$ at various levels of V_T/V_R ratios where the pooled variance is controlled so as to yield a CV of 40%. *Filled circles* indicate evaluations where unequal variances are assumed; *open circles* indicate evaluations where equal variances are assumed. Two important conclusions are drawn: The two types of evaluations perform rather similarly, and the performance in terms of type I error rate and power appears rather constant across the range of simulated V_T/V_R values

0.39878, and we wish to keep CV_T at 0.4 corresponding to $V_T=0.14842$. By insertion, we get $V_R=2V-V_T=0.95597$.

RESULTS AND DISCUSSION

Figure 1 graphs the performance of method C in terms of type I error rate and power at $N_1=48$ for evaluations based on the normal linear model with stage and treatment as fixed factors where equal variances are assumed, as well as for the model with Welch-Satterthwaite correction where unequal variances are assumed. Under these scenarios, the two assumptions performed similarly, and, equally important, performance was rather constant across the range of V_T/V_R values tested. Similarly, the power was roughly constant and near the desired 80% level across the range of V_T/V_R values tested. This means, the type I error rate is roughly constant and at or below the 0.05 limit when the test product is associated with higher variability than the reference product, when the test product is associated with the same variability as the reference product and also when the test product is associated with lower variability as the reference product. The same is observed for power.

Figure 2 shows similar results for $N_1=120$, and the same conclusions are reached. The same conclusions are drawn when method B is evaluated at $N_1=48$ and $N_1=120$ (see

Supplementary material). The same conclusions are reached for methods B and C when the CV is raised to 0.7 (data not shown), suggesting—but not necessarily exhaustingly proving—that the conclusion is generalisable across low and high variabilities and low and high sample sizes.

There are two major methodological implications of these findings:

1. The distinction between the evaluation methods (equal variances or unequal variances) is not important. The Welch-Satterthwaite approach was used to generate the data presented here. This is due to the fact that the guidelines suggest that equal variances should not be assumed for parallel designs.
2. For any given level of CV calculated from the pooled variances, the individual variabilities of test and reference are not crucial as long as we control the pooled variance. It is therefore reasonable to evaluate and report performance at $CV_T=CV_R$; the tables reporting the performance of the methods are therefore constructed on that basis. It should be emphasized that it cannot strictly be ruled out that scenarios exist where this would not hold true. This is, however, impossible to systemically test since an infinite number of combinations of N_1 , CV_T , CV_R exist.

Fig. 2. Type I error rate (*left*) and power (*right*) for Potvin's method C adapted to parallel two-stage designs at $N_1=120$ at various levels of V_T/V_R ratios where the pooled variance is controlled so as to yield a CV of 40%. *Filled circles* indicate evaluations where unequal variances are assumed; *open circles* indicate evaluations where equal variances are assumed. The conclusions drawn at $N_1=120$ are the same as for $N_1=48$ (see Fig. 1): The two types of evaluations perform rather similarly, and the performance in terms of type I error rate and power appears rather constant across the range of simulated V_T/V_R values

Table I. Power, Type I Error Rates and Sample Size Characteristics for Method B Applied to Parallel Bioequivalence Designs

CV	N_1	Power	Type I error rate	AvgN	F5	F50	F95
0.1	48	1.000	0.0294	48.0	48	48	48
0.2	48	0.870	0.0305	48.4	48	48	50
0.3	48	0.826	0.0476	74.5	48	72	122
0.4	48	0.805	0.0413	148.6	48	152	212
0.5	48	0.791	0.0311	231.6	160	228	316
0.6	48	0.787	0.0297	318.6	218	314	434
0.7	48	0.785	0.0296	412.2	282	406	562
0.8	48	0.785	0.0294	510.9	350	504	696
0.9	48	0.784	0.0295	612.3	420	604	836
1.0	48	0.783	0.0297	714.8	490	704	976
1.0	84	0.793	0.0295	714.7	542	708	906
0.1	120	1.000	0.0297	120.0	120	120	120
0.2	120	0.998	0.0297	120.0	120	120	120
0.3	120	0.902	0.0294	120.0	120	120	120
0.4	120	0.826	0.0413	132.3	120	120	178
0.5	120	0.824	0.0478	185.8	120	200	278
0.6	120	0.815	0.0467	280.9	120	308	388
0.7	120	0.804	0.0368	401.0	280	410	504
0.8	120	0.797	0.0303	509.9	406	508	624
0.9	120	0.796	0.0298	611.9	488	608	748
1.0	120	0.795	0.0294	714.6	570	710	874

The evaluation for bioequivalence is based on an assumption of unequal variances. Both test and reference were simulated with the CVs indicated

N_1 sample size at stage 1, AvgN average total sample size, F5 5th percentile of total sample sizes, F50 50th percentile of total sample sizes, F95 95th percentile of total sample sizes

Table I shows the performance characteristics of method B at $N_1=48$ and $N_1=120$, while Table II shows the corresponding performance characteristics for method C; the entire set of results is uploaded as [Supplementary material](#). Generally, type I error rates do not significantly exceed the limit of 0.05 although a few exceptions are noted. At

1,000,000 simulations per scenario the type I error rate is statistically significantly inflated when it exceeds 0.05036. Power ranges from about 70% and upwards depending on N_1 and variability. The standard errors for the estimates of the type I error rate and power did not exceed 0.0003 and 0.0005, respectively. This is in line with the results drawn from

Table II. Power, Type I Error Rates and Sample Size Characteristics for Method C Applied to Parallel Bioequivalence Designs

CV	N_1	Power	Type I error rate	AvgN	F5	F50	F95
0.1	48	1.000	0.0504	48.0	48	48	48
0.2	48	0.908	0.0499	48.2	48	48	48
0.3	48	0.827	0.0479	74.4	48	72	122
0.4	48	0.806	0.0410	148.5	48	152	212
0.5	48	0.790	0.0311	231.6	160	228	316
0.6	48	0.787	0.0298	318.5	218	314	434
0.7	48	0.785	0.0295	412.4	282	406	562
0.8	48	0.784	0.0297	511.0	350	504	696
0.9	48	0.784	0.0297	612.2	420	604	836
1.0	48	0.784	0.0297	714.7	490	704	976
0.1	120	1.000	0.0502	120.0	120	120	120
0.2	120	0.999	0.0502	120.0	120	120	120
0.3	120	0.938	0.0506	120.0	120	120	120
0.4	120	0.830	0.0454	131.2	120	120	178
0.5	120	0.824	0.0478	185.8	120	200	278
0.6	120	0.815	0.0467	280.9	120	308	388
0.7	120	0.804	0.0368	401.0	280	410	504
0.8	120	0.797	0.0303	509.9	406	508	624
0.9	120	0.796	0.0298	611.9	488	608	748
1.0	120	0.795	0.0294	714.6	570	710	874

The evaluation for bioequivalence is based on an assumption of unequal variances. Both test and reference were simulated with the CVs indicated

N_1 sample size at stage 1, AvgN average total sample size, F5 5th percentile of total sample sizes, F50 50th percentile of total sample sizes, F95 95th percentile of total sample sizes

crossover bioequivalence studies. Methods B and C performed similarly under the tested scenarios. Potvin's methods B and C when adapted to parallel designs would thus seem to be appropriate to use.

It should be emphasized that these simulations involve the assumption that the sample sizes for the test and reference treatments are equal at stage 1 and that the sample sizes for test and reference are equal at stage 2. Hence, results may differ for designs with unbalanced sample sizes, and this aspect might merit further studies. There is plenty of room for further work in this field. For example, it would be relevant to investigate how the methods perform when the population geometric mean ratio is 0.90 rather than 0.95, when futility criteria (criteria for stopping after stage 1) are applied or when there is imbalance between treatment groups stage 1, stage2 or both.

CONCLUSION

1. Potvin's methods B and C adapted to parallel designs seem to generally protect well against inflation of type I error rates.
2. It does not seem to matter in practice if one assumes unequal variances (Welch–Satterthwaite correction of degrees of freedom) or if one assumes equal variances (a normal linear model with stage and treatment as fixed factors) for the evaluation of bioequivalence.
3. Methods B and C generally perform quite similarly in terms of type I errors, power and sample sizes.
4. Due to points 1–3, methods B and C appear to be appropriate for studies where parallel groups in conjunction with a two-stage approach are desirable.

ACKNOWLEDGMENTS

Thanks to Helmut Schütz and Detlew Labes who provided valuable input.

REFERENCES

1. Committee for Human Medicinal Products. Investigation of bioequivalence. CHMP. CPMP/EWP/QWP/1401/98 Rev. 1. 2010. http://www.ema.europa.eu/ema/pages/includes/document/open_document.jsp?webContentId=WC500070039. Accessed 1 Mar 2013.
2. United States Food and Drug Administration, Center for Drug Evaluation and Research. Statistical approaches to establishing bioequivalence. Guidance for industry: statistical approaches to establishing bioequivalence. 2001. <http://www.fda.gov/downloads/Drugs/Guidances/ucm070244.pdf>. Accessed 1 Mar 2013.
3. Therapeutic Products Directorate, Health Canada. Conduct and analysis of comparative bioavailability studies. 2012. http://www.hc-sc.gc.ca/dhp-mpps/alt_formats/pdf/prodpharma/applic-demande/guide-ld/bio/gd_cbs_etc_ld-eng.pdf. Accessed 1 Mar 2013.
4. Potvin D, DiLiberti CE, Hauck WW, Parr AF, Schuirmann DJ, Smith RA. Sequential design approaches for bioequivalence studies with crossover designs. *Pharm Stat*. 2008;7:245–62.
5. Montague TH, Potvin D, DiLiberti CE, Hauck WW, Parr AF, Schuirmann DJ. Additional results for 'sequential design approaches for bioequivalence studies with crossover designs'. *Pharm Stat*. 2012;11:8–13.
6. Matsumoto M, Nishimura T. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans Model Comput Simul*. 1998;8:3–30.
7. Labes D. Power and sample size based on two one-sided *t*-tests (TOST) for (bio)equivalence studies. 2013. <http://cran.r-project.org/web/packages/PowerTOST/PowerTOST.pdf>. Accessed 1 Mar 2013.
8. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biom Bull*. 1946;2:110–4.
9. Welch BL. The generalization of "Student's" problem when several different population variances are involved. *Biometrika*. 1947;34:28–35.