Research Article

Futility Rules in Bioequivalence Trials with Sequential Designs

Anders Fuglsang^{1,2}

Received 14 August 2013; accepted 11 October 2013; published online 12 November 2013

Abstract. Health Canada, the US Food and Drug Administration, as well as the European Medicines Agency consider sequential designs acceptable for bioequivalence studies as long as the type I error is controlled at 5%. The EU guideline explicitly asks for specification of stopping rules, so the goal of this work is to investigate how stopping rules may affect type I errors and power for recently published sequential bioequivalence trial designs. Using extensive trial simulations, five different futility rules were evaluated for their effect on type I error rates and power in two-stage scenarios. Under some circumstances, notably low sample size in stage 1 and/or high variability power may be very severely affected by the stopping rules, whereas type I error rates appear less affected. Because applicants may initiate sequential studies when the variability is not known in advance, achieving sufficient power and thereby complying with certain guideline requirements may be challenging and application of optimistic futility rules could possibly be unethical. This is the first work to investigate how futility rules affect type I errors and power in sequential bioequivalence trials.

KEY WORDS: alpha; bioequivalence; futility rules; power; sequential design.

INTRODUCTION

The most common design in bioequivalence trials involves two treatments (test and reference), crossover with two sequences (test–reference or reference–test), and two periods. Blood samples are taken at regular intervals to quantify the maximum concentration (C_{max}) and the area under the concentration–time curve (AUC) by non-compartmental methods. The evaluation involves construction of a 90% confidence interval for the test/reference ratio (T/R). The general success criterion is that the 90% confidence interval must lie within 0.80–1.25. This approach corresponds to a maximum tolerable type I error rate of 5% (regulator's risk or patient's risk).

The power of such a trial depends on the sample size, the intrasubject coefficient of variation, and the test/reference ratio. 1-power is referred to as the applicant's risk, *i.e.*, the risk of not showing bioequivalence. To calculate power, one needs to know or estimate the true performance of the test product *versus* the reference product in terms of AUC and C_{max}, and the associated variability (coefficient of variation; CV). In some cases, applicants can estimate variability figures through literature studies and perform dissolution tests to estimate T/R for solid dosage forms, but in other cases, the performance of the test and reference products may not be

known, and this can make it difficult to power a traditional trial adequately. Potvin and co-workers investigated by means of simulation studies how sequential designs can be implemented in such situations under assumed test/reference ratios of 0.95; they initially published three methods. They involve evaluation in a limited number of subjects followed—if necessary—by inclusion of additional subjects and final evaluation. Frameworks for three methods (termed B, C, and D) are shown in Figs. 2, 3 and 4 in Potvin *et al.* (1). For method B, there is an initial evaluation of bioequivalence; if the bioequivalence criterion is met at $\alpha=0.0294$, the products are deemed bioequivalent; otherwise, one tests if the power was 80% or more. If it was, then we stop without concluding bioequivalence. If power was below 80%, we include more subjects for a desired power goal of 80% and evaluate bioequivalence at $\alpha=0.0294$. For method C, we first evaluate power at stage 1; if the power is 80% or better, we test for bioequivalence and pass or fail at stage 1 using $\alpha=0.05$. Otherwise, we test for bioequivalence at $\alpha=0.0294$. If the test fails, we include new subjects for a power goal of 80% and test for bioequivalence using $\alpha=0.0294$ (method C) or $\alpha=0.280$ (method D).

The methods apply the observed CV from stage 1 in the calculation of sample size in stage 2, but they do not use the observed T/R for calculation of the final sample size. Instead, an assumed value (constant) for T/R is used. Under the tested circumstances with assumed T/R=0.95, methods B and C generally performed equally. However, in some cases, method B was more conservative in terms of type 1 error rates, and method C presented slightly better power and smaller average total sample sizes. Later, Montague *et al.* (2) evaluated these methods when the test/reference ratio is 0.90.

Electronic supplementary material The online version of this article (doi:10.1208/s12248-013-9540-0) contains supplementary material, which is available to authorized users.

¹ Fuglsang Pharma, Hiort Lorenzens Vej 6C, 6100 Haderslev, Denmark.

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

Two-stage approaches seem to be acceptable for regulatory submissions in USA, Canada, and Europe. In Europe, the 2010 regulatory guidance for bioequivalence trials states the following:

It is acceptable to use a two-stage approach when attempting to demonstrate bioequivalence. An initial group of subjects can be treated and their data analyzed. If bioequivalence has not been demonstrated, an additional group can be recruited and the results from both groups combined in a final analysis. If this approach is adopted, appropriate steps must be taken to preserve the overall type I error of the experiment and **the stopping criteria should be clearly defined** prior to the study (3).

Emphasis by the author.

Potvin's and Montague's works did not involve such stopping criteria or futility rules as the authors call them. This is when Sponsor's will not be willing to conduct a second stage; this would typically be due to high sample size. Potvin *et al.* also point out that this matter deserves further attention. The purpose of this work is therefore to investigate how Potvin's methods perform in terms of power and type I error rates when stopping criteria are applied. Potvin recommended methods B and C when the test/reference ratio is 0.95, while the subsequent work by Montague *et al.* identified method D as the preferable method when the test/reference ratio is 0.90 (2). Accordingly, in this work, I will investigate the performance of methods B and C in conjunction with stopping criteria at test/reference ratios of 0.95, and I will investigate the performance of method D when the test/reference ratio is 0.90 in conjunction with stopping criteria.

MATERIALS AND METHODS

Software and Simulations

The methods invented by Potvin *et al.* (1) were implemented in software using the MinGW C compiler version 3.4.2 in the Code::Blocks environment version

12:11. The software was validated against the results published by Montague *et al.* and Potvin *et al.* (1,2) and their results can be reproduced. The software was written so as to allow expansion of Potvin's methods with a stopping criterion using a cap on maximal sample sizes. The Mersenne Twister algorithm was used for random number generation due to its long period, and a Box-Muller transform was used to derive Gaussian numbers.

In line with the previous publications (1,2), 1,000,000 trials were simulated per scenario.

There are many ways to implement a stopping criterion. One could work with a fixed maximum sample size across all scenarios, but since applicants are free to choose a starting sample size (sample size at stage 1, N_1), it might also be desirable to define stopping criteria that scale with N_1 itself. Here, I investigate stopping criteria defined as scenarios where the total sample size cannot exceed $2\times$, $3\times$, or $4\times$ the sample size at stage 1, and I also investigate fixed maximum sample sizes of 60 and 80, respectively. There are definitely additional ways of implementing stopping criteria; such other stopping schemes fall outside the scope of this work.

RESULTS

This study involves simulation of 576,000,000 trials scenarios for methods B/C/D at various T/R levels, CV-levels, and values of N_1 . The [supplementary material](#), which is available online, summarizes all of them with type I error rates, power, average sample sizes, and their fractiles (5th, 50th, 95th). It is not desirable to present all the generated data in tables here. Figure 1 shows power and type 1 error rates for method B at T/R=0.95 for CV=0.2 and CV=0.4, respectively. Figure 2 shows the corresponding graphs for method C. Figure 3 shows the corresponding graphs for method D where T/R is controlled at 0.90.

It can be seen with these examples that type I error rates do not suffer from the futility rules applied here, but power may be very negatively affected. The [supplementary material](#) (accessible online) shows the results for all the scenarios.

Fig. 1. Power (left pane) and type 1 error rate (right pane) as function of N_1 for method B at T/R=0.95 and CV=0.2 (black circle) or at CV=0.4 (white circle), when the total sample size cannot exceed $3\times N_1$. The dashed line at the right pane indicates the traditional limit of 0.05

Fig. 2. Power (left pane) and type 1 error rate (right pane) as function of N_1 for method C at $T/R=0.95$ and $CV=0.2$ (black circle) or at $CV=0.4$ (white circle), when the total sample size cannot exceed $3 \times N_1$. The dashed line at the right pane indicates the traditional limit of 0.05

$$\begin{aligned} N_{\max} &\leq 2 * N_1 \\ N_{\max} &\leq 3 * N_1 \\ N_{\max} &\leq 4 * N_1 \\ N_{\max} &\leq 60 \\ N_{\max} &\leq 80 \end{aligned}$$

These tables also give information about the sample size fractiles and averages. When reading these tables, it can be mentioned that at one 1,000,000 simulations, the type I error rate can be considered statistically significantly inflated at or above 0.0513, while Potvin *et al.* used 0.052 to signify a clinically relevant inflation (1).

An important conclusion to draw from these tables is that methods B and C perform rather equally when futility criteria are applied. Their performance was also quite similar in the absence of futility criteria (1). Differences between methods B and C are most visible when we look at high initial sample size and low variability. It is worth noting that an increase from $CV=0.2$ to 0.4, however modest it may sound, can in some cases diminish power to extremely low levels (see, *e.g.*, Fig. 1 left pane).

DISCUSSION

Some general inferences can be made from these results as follows:

1. The type I error rates do not generally suffer from the stopping criteria.
2. Power may suffer very considerably from stopping criteria; for example, at $N_1=12$ and $CV=0.3$, power in methods B and C drops below 30% when the stopping criterion is that the total sample size cannot exceed $3 \times$ the stage 1 sample size ($N_{\max}=36$). Without a stopping criterion, the power is above 78% (Ref. 1 and own simulations).
3. Methods B and C generally perform rather equally in terms of power. In terms of type I errors, method B was more conservative for CVs of 10 to 30%. The methods did not significantly exceed the goal post of 0.05.

Since the stopping criteria implemented here seem to mainly affect power rather than the overall alpha level, it can be concluded that the requirement to specify stopping criteria mainly imply a risk to sponsor/applicants rather than to patients. Thus, the stopping criteria implemented here appear generally safe.

Assuming that power should not be too much below 0.8, we can see that a sample size of $N_1=12$ does not seem to be a good idea if there is a chance that the true variability exceeds approximately 20%. In this regard, it should be noted that variability in practice is not only depending on biology and the active pharmaceutical ingredient but also on factors such as bioanalytical precision and accuracy. Furthermore, if we

Fig. 3. Power (left pane) and type 1 error rate (right pane) as function of N_1 for method B at $T/R=0.95$ and $CV=0.2$ (black circle) or at $CV=0.4$ (white circle), when the total sample size cannot exceed $3 \times N_1$. The dashed line at the right pane indicates the traditional limit of 0.05

assume T/R below 0.90, as may be necessary and realistic in some cases, we see that using $N_1=12$ does not bring power above 50%, except if $CV \leq 10\%$ which is only very rarely the case in practice.

From an ethical perspective, it might not be appropriate to initiate trials that have too low power. Due consideration should be given to the general principle of ICH guideline E9 (4) which specifies that 'the number of subjects in a clinical trial should always be large enough to provide a reliable answer to the questions addressed.' The same guideline suggests that power should be 80–90% (type II error rates 10–20%), although the wording is less explicit.

Given the results obtained here, I will not recommend that $N_1=12$ be used regardless of the stopping criterion, algorithm, and expected (assumed) T/R. It might be proposed that it is desirable from a sponsor's perspective to limit the development costs as much as possible and that one should therefore start out with a very low number of subjects in stage 1, as this would limit the costs as much as possible in case a trial has to be stopped after stage 1. The results obtained here do not completely support this type of argumentation, as a low value of N_1 often will imply a penalty in terms of power which translates into higher chance of need for additional trials. From an ethical standpoint, it might be better if sponsors could prospectively specify their desired level of power along with a reasonable estimate of maximum variability. Then it would be possible to calculate a futility rule for any given N_1 .

Example, assuming a T/R of 0.90, maximum variability of 30%, $N_1=20$, and a minimum desired power level of 75%, simulations show that the appropriate futility level is $N_{\max}=150$ with a type I error rate below 5%.

In the writing phase of this manuscript, a paper investigating if the observed test/reference ratio at stage 1 can be used to help define a futility rule was published (5); their work illustrates the performance of futility rules when the point estimate from stage 1 is taken into account for the planning of the second stage; their results are, thus, not directly comparable

to the data generated in this study. It is worth noting that under the conditions simulated in that study, power was also found to suffer considerably. The two papers—although going in different directions—are, thus, in accordance as regards with the impact of futility rules on power, and it would be desirable to look into methods that would allow futility rules without impacting power too negatively.

CONCLUSION

Introduction of the five futility rules studied in this paper may severely impact power in trials with sequential designs and under some circumstances such trials might be unethical.

REFERENCES

1. 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.
2. 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. doi:10.1002/pst.483.
3. Committee for Human Medicinal Products. Investigation of Bioequivalence. European Medicines Agency, 2010. Available via http://www.ema.europa.eu/ema/pages/includes/document/open_document.jsp?webContentId=WC500070039
4. Steering Committee (International Conference of Harmonization). Statistical Principles for Clinical Trials. International Conference of Harmonization. 1998. Available via http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf
5. Karalis V, Macheras P. An insight into the properties of a two-stage design in bioequivalence studies. *Pharm Res.* 2013;30:1824–35.