Research Article

Pilot and Repeat Trials as Development Tools Associated with Demonstration of Bioequivalence

Anders Fuglsang^{1,2}

Received 10 August 2014; accepted 30 December 2014; published online 4 March 2015

Abstract. The purpose of this work is to use simulated trials to study how pilot trials can be implemented in relation to bioequivalence testing, and how the use of the information obtained at the pilot stage can influence the overall chance of showing bioequivalence (power) or the chance of approving a truly bioinequivalent product (type I error). The work also covers the use of repeat pivotal trials since the difference between a pilot trial followed by a pivotal trial and a pivotal trial followed by a repeat trial is mainly a question of whether a conclusion of bioequivalence can be allowed after the first trial. Repeating a pivotal trial after a failed trial involves dual or serial testing of the bioequivalence null hypothesis, and the paper illustrates how this may inflate the type I error up to almost 10%. Hence, it is questioned if such practice is in the interest of patients. Tables for power, type I error, and sample sizes are provided for a total of six different decision trees which allow the developer to use either the observed geometric mean ratio (GMR) from the first or trial or to assume that the GMR is 0.95. In cases when the true GMR can be controlled so as not to deviate more from unity than 0.95, sequential design methods *ad modum* Potvin may be superior to pilot trials. The tables provide a quantitative basis for choosing between sequential designs and pivotal trials preceded by pilot trials.

KEY WORDS: bioequivalence; pilot trials; power; type I error.

INTRODUCTION

Demonstration of bioequivalence is a requirement for approval of generics in many countries, as well as in certain types of interaction studies and line extensions, and is also used, *e.g.*, to prove similarity of formulations used during development. Bioequivalence is commonly evaluated by comparing rate and extent of absorption for drugs intended to be systemically absorbed before action as well as for some locally acting drugs. The most common design is a 2-treatment, 2-period, 2-sequence randomized crossover trial. The primary metrics are most often the maximum observed concentration (C_{max}, typically used as an indicator of rate of absorption) on the time-concentration curve and the area under this curve until last sampling point (AUC_t, typically used as an indicator of extent of absorption). Bioequivalence is usually declared when the 90% confidence intervals for the ratios of geometric means (GMRs) fall within 80.00 to 125.00% (*e.g.*, 1–3).

Electronic supplementary material The online version of this article (doi:10.1208/s12248-015-9744-6) contains supplementary material, which is available to authorized users.

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

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

In order to calculate sample size for a desired level of power in a pivotal bioequivalence trial, it is necessary to estimate or guess the *in vivo* geometric mean ratio (GMR) and the associated variability. The true GMR is a formulation issue and can sometimes be estimated from *in vitro* trials such as dissolution studies in appropriate media in the case of solid oral dosage forms. The variability, however, cannot be reliably estimated by *in vitro* trials but may sometimes be derived from public assessment reports or scientific literature. Occasionally, developers will find themselves in a situation where it is desirable to conduct a pilot trial in order to estimate missing information. There is not much literature which has evaluated how pilot studies should be conducted and how their resulting information should be used towards the pivotal trial. One purpose of this work is therefore to evaluate different modes of using pilot trials. The issue of determining sample size in a pilot is outside the scope of this paper since such determination is usually not based on hard scientific evidence but often in practice rely on budgetary constraints and gut feeling.

In a sense, the use of a repeat pivotal trial after a pivotal trial can be considered the same as the use of a pivotal trial after a pilot trial—this comes down to the semantics (“pilot” *versus* “pivotal”) and the way the result obtained from the first trial is used, in particular whether or not bioequivalence can be declared after the first trial. In practice, it does happen that applicants plan and execute a pilot trial in the anticipation of using information from the pilot trial to plan and execute a

Table I. Graphical Illustration of Important Terms and their Definitions Used in this Manuscript

Conclusion	Confidence interval (%)
	<div><div></div><div>80100125</div><div></div></div>
Bioequivalent	<div><div></div><div></div><div></div></div>
Bioinequivalent (BE not shown)	<div><div></div><div></div><div></div></div>
Inconclusive (BE not shown)	<div><div></div><div></div><div></div></div>

Two products are said to be bioequivalent if the confidence interval is contained within the acceptance limits of 80.00 to 125.00% (examples: *green bars*). Two products are said to be bioinequivalent and BE has not been shown if the confidence interval is entirely outside the acceptance limits of 80.00 to 125.00%. The two products have been shown not to be bioequivalent (examples: *orange bars*). A BE trial is inconclusive and BE has not been shown if a part of the confidence interval has an overlap with the acceptance range of 80.00 to 125.00% but is not entirely contained within it (examples: *violet bars*). It is impossible to tell from the trial result if the products should be considered either bioequivalent or bioinequivalent

pivotal trial and find proof of bioequivalence already at the pilot stage. These cases are rare, and from a regulatory perspective, it is not completely clear if such trials (when nominated “pilot” trials in the protocols) can be submitted for approval. On one hand, ethics might be considered; if a pilot trial has shown bioequivalence, it may be unethical to conduct a new trial since this could be seen as an unnecessary risk associated with additional human exposure to investigative medicinal products *cf.* clause 16 of the Helsinki Declaration: “Medical research involving human subjects may only be conducted if the importance of the objective outweighs the risks and burdens to the research subjects.” (4) A similar wording is found in ICH E6 (Good Clinical Practice) clause 2.2.: “Before a trial is initiated, foreseeable risks and inconveniences should be weighed against the anticipated benefit for the individual trial subject and society. A trial should be initiated and continued only if the anticipated benefits justify the risks” (5).

On the other hand, when a pilot trial is conducted, then, the purpose, in contrast to a pivotal trial, is often exploratory in nature (for example, a pilot can be used to determine

product similarity or the residual variability of the similarity), and accordingly, some regulators might insist that such a conclusion—effectively changing the trial objective from exploratory to confirmatory—could not be allowed. Moreover, if applicants work with the traditional alpha of 5% in their trials, then, an opportunity to conclude bioequivalence after a pilot trial as well as after a pilot trial’s resulting pivotal trial involves multiplicity in terms of equivalence testing and this could lead to an overall inflation type I error (risk of approving a product that is not bioequivalent).

The entire issue of using pilot or repeat trials, ways to use the information generated by them, and what such use implies for total sample size, power, and overall type I error is not well studied.

The purpose of this paper is therefore to introduce various ways to work with pilot trials (or pivotal trials with repeating options) and investigate what the implications are in terms of sample size, power, and overall type I error.

The scope of this work is limited to the serial application of the standard 2-treatment, 2-sequence, 2-period trials.

Table II. Characteristics of the Eight Different Methods Examined in this Paper

Characteristic	Method							
	1	2	3	4	5	6	7	8
Allow conclusion of bioequivalence after first/pilot trial		✓		✓		✓		✓
Conclude failure if first/pilot trial is bioinequivalent					✓	✓	✓	✓
Use observed GMR from first/pilot for sample size in second trial ^a			✓	✓			✓	✓

^a Otherwise, assume GMR=0.95

Parallel designs or (semi-)replicated designs incl. reference scaled bioequivalence are outside the scope of this work as is investigation of alpha levels other than the traditional 0.05.

MATERIALS AND METHODS

Methods 1–6

Central to this manuscript is the way the information from the first trial (pilot trial) is used after this has been conducted. As mentioned above, one important choice is whether to allow the conclusion of bioequivalence at this point. Another choice to make is whether the observed GMR (GMR_{obs}) from the first trial should be used for the calculation of sample size in the second trial when one such is necessary, or if the assumed GMR should be 0.95 (or another fixed value) as was the principle behind the two-stage methods introduced by Diane Potvin and coworkers (6,7). Third, in a regulatory sense, there one can make an important distinction between the terms “bioequivalent,” “bioinequivalent,” and “inconclusive” (8), see Table I for an explanation. Basically, concluding that two products are bioequivalent for the given endpoint implies that the confidence interval is contained entirely within the acceptance range (usually 0.8000 to 1.2500). *Bioequivalent* is when the confidence interval is entirely outside the acceptance range. *Inconclusive* is when part of the confidence interval is inside the acceptance range and part of the confidence interval is outside the acceptance range. While regulatory acceptance of a product requires the conclusion of *bioequivalence*, a sponsor’s wish to proceed with a second trial after a trial that does not show bioequivalence may hinge on whether the first trial is actually inconclusive or bioinequivalent. If the first trial proves bioinequivalent, the prospects for a second trial may be considered minimal and the development could simply stop at that point *cf.* ICH E6 clause 2.2. The matter is not covered well by specific regulatory guidance.

These considerations give rise to a total of $2 \times 2 \times 2 = 8$ different decision trees (hereafter called methods 1, 2, 3... 8) whose key features are summarized in Table II. In practice, though, method 7 gives results that are identical to the results obtained with method 3, and method 8 gives results that are identical to the results obtained with method 4. The reason is that those studies that are deemed bioinequivalent after the first trial must by definition have a point estimate that is outside the acceptance range, and for these trials in methods 3 and 4, no sample size will give a power above 5%. These situations will thus lead to failure. Therefore methods 7 and 8 will not be discussed further in this manuscript.

Figures 1 and 2 show a graphical representations of methods 1 and 6, respectively, which illustrate the range of complexity studied here; method 1 is relatively simple, method 6 is more involved *cf.* Table II. Graphical representations of all methods are uploaded as [supplementary material](#).

It should be emphasized that these methods do not involve pooling of data of the two trials.

Software and Trial Scenarios

The software used to evaluate methods 1–6 was programmed in C and compiled with MinGW 4.7.1 using the

Fig. 1. Decision tree for method 1

Code::Blocks 12.11 environment under Windows 8. Mersenne Twister was used as the generator for random numbers due to its long period (9). A Box-Muller transform was used to generate pseudorandom numbers with a normal distribution. Equations implemented by the software are given in the paper by Potvin *et al.* (6). Since these simulations are computationally intensive and involve memory allocation from the stack, there was a cap of 10,000 subjects in all trial simulations. Trials exceeding this sample size would be declared failed.

Scenarios with $N_1=12, 24, 36, 48$, and 60 as in the paper by Potvin *et al.*, and with coefficients of variation (CVs) ranging from 0.1 (10%) to 0.8 (80%). In all cases, alpha levels of 0.05 were used, and the target power for the second trial was 80%.

Type I errors were derived from simulations at $GMR=0.8$.

RESULTS

The amount of data generated for this manuscript is vast and cannot be reasonably presented by tables here. The supplementary material which is available online holds all performance results generated (uploaded as “Method 1.pdf,” “Method 2.pdf,” *etc.*). Here, I shall illustrate the performance of the eight methods using scenarios derived from a sample size of 24 in the first trial which in my experience is a fairly common choice in the industry.

Figure 3 illustrates how the type I error rate varies with the coefficient of variation for all methods. It is noted that methods 2, 4, and 6 (those methods that allow concluding bioequivalence (BE) after the first trial; the decision tree in itself can be seen as a kind of group

Fig. 2. Decision tree for method 6

Fig. 3. Type I error rate as function of coefficient of variation when an initial sample size of 24 subjects is applied at GMR=0.95. Note that for methods 2, 4, 6, and 8, the type I error rate exceeds 0.05. The maximum standard error was 3.0×10^{-4} , observed for method 1 at CV=0.1

sequential trial) may lead to inflated overall type I errors. This can be seen as a confirmation of the theory of multiplicity; thus, there is no clear benefit even though power may appear relatively high.

Figure 4 shows the power as function of the coefficient of variation at a GMR of 0.95. At low CVs, the methods 2, 4, and 6 (methods that allow the BE conclusion after the first trial) are superior to methods 1, 3, and 5. At high CVs, methods 1, 2, 5, and 6 (those that use an assumed GMR of 0.95 for the calculation of sample size in the second trial) perform better than methods 4 and 3. The same methods are quite superior in terms of sample size at a GMR of 0.95 as illustrated by Fig. 5.

To illustrate how the methods perform when there is deviation from a GMR of 0.95, consider Fig. 6 which shows power as function of the (true) GMR. Generally, as the true GMR departs downwards from 0.95 towards the BE limit, power drops off. Throughout most of the interval, method 4 is superior while methods 1 and 5 are inferior.

DISCUSSION

On basis of the results obtained here with methods 2, 4, and 6, I am inclined to conclude that allowing repeat trials when an initial pivotal trial has not shown bioequivalence is not a practice that protects the patient's interest as it leads to overall type I errors above 0.05, even though power may appear relatively high, so in terms of regulatory science, a guideline change is suggested. Obviously, the inflation of type I errors might be remedied by applying adjusted (decreased) alphas for the construction of confidence intervals after the first or after both trials. This might on the other hand involve a trade-off in terms of power or sample size. Investigation of alpha adjustment is outside the scope of this paper but is a relevant future task. A possible alternative would perhaps be to never repeat trial on "the same" formulation.

The results obtained here do not allow us to conclude that any single method is superior to the others. In terms of type I errors, methods 1, 3, and 4 are obviously better *cf.*

Fig. 4. Power as function of the coefficient of variation when an initial sample size of 24 subjects is applied at GMR=0.95. The maximum standard error was 4.9×10^{-4} , observed for method 3 (=7) at CV=0.1

Fig. 5. Average sample size as function of coefficient of variation when an initial sample size of 24 subjects is applied at GMR=0.95. Note that methods 3, 4, 7, and 8 are associated with a distinctly higher sample size than methods 1, 2, 5, and 6 since the latter assume a GMR of 0.95 for the planning of the second trial

Fig. 3, since these methods do not involve serial testing for bioequivalence.

As regards power, methods 2 and 6 are better *cf.* Fig. 4. When the true geometric ratio deviates from 0.95, method 4 seems better. If a developer feels reasonably sure (through, *e.g.*, dissolution tests or other appropriate methods, where available) that the true GMR is not deviating more from unity than 0.95 (or $1/0.95$) and assuming the developer wishes to use an approach that protects against inflation of type I errors, it becomes a matter of prioritizing power *versus* sample size as well as possibly the risk of not matching the GMR of 0.95 in case the choice is between methods 1, 2, 5 and 6.

Method 3 is the most conservative in terms of type I error, and the power is relatively high but this does come at the cost of a high sample size.

The results obtained here can be appropriately compared to the results obtained by Potvin *et al.* in their studies of

two-stage approaches. The original methods introduced by Potvin involve studies on true GMRs of 0.95 and the use of the same (assumed) GMR for the planning of the second trial stage. Potvin's methods B and C by and large protect against type I error rate inflation and so they should be compared at least with methods 1, 3, and 5. See, *e.g.*, the case of CV=30% and an initial sample size of $N=24$: in Potvin's work, methods B and C would give rise to average total sample sizes of around $N=40$. In this work, with the same initial sample size, we get average total sample sizes of approximately $N=64$ for methods 1 and 5 or approximately $N=171$ for method 3. This implies that when an applicant is certain that the true GMR is not worse (meaning, not farther from unity) than 0.95, then, Potvin's methods B and C offer a natural advantage in terms of sample size over the methods studied here.

In practice, it is my impression that the use of pilot trials is still much more widespread than the use of the sequential approaches. With these results, the occasional advantage

Fig. 6. Power as function of the GMR for the eight methods when an initial sample size of 24 subjects is applied and when the CV is 0.3. Methods 2, 4, 6 and 8 are generally associated with the highest power, but those are methods that can inflate the type I error rate due to serial testing for bioequivalence *cf.* table II. The maximum standard error was 5.0×10^{-4} , observed for methods 5 and 7 at GMR=0.90

associated with the use of Potvin's approaches in certain cases has been quantified but the paper also importantly puts a quantity to the inflation of type I errors that repeat trials in themselves cause.

This work also establishes a bit of insight into situations when the assumption about GMR do not hold true; similar investigations have not been undertaken with Potvin's methods. Specifically, it would be interesting to expand this work by investigating how Potvin's methods perform when the true GMR is 0.90 and the assumed GMR for the planning of the second stage is 0.95 and to compare the results with those obtained here. After all, testing of BE aims at gaining confidence in the estimate of the GMR, the true quantity of which can never be known.

This work has studied 2-treatment, 2-sequence, 2-period designs with CVs up to 80%. With highly variable drug products (true intrasubject CV at or above 30%), applicants may have a chance to apply scaled average bioequivalence for submissions in at least EU and USA. In such cases, the applicable designs would involve repeated administration of the reference product and would imply three or more periods. These designs fall outside the scope of this work.

CONCLUSIONS

This is the first work to systematically investigate the performance of pilot or repeat trials in specific relation to bioequivalence.

The main conclusions are the following:

1. The principle of allowing a repeat trial after a pivotal trial, both with an alpha level of 0.05, easily leads to inflation of the overall type I errors.
2. When the true GMR is controlled and is not deviating more from unity than 0.95, methods 1, 3, and 4 are useful; in particular, method 1 may have some appeal: It is not associated with type I error inflation, power is 72% or better, and it leads to low average sample sizes.
3. It is not possible to identify a method that consistently gives *relatively* high power and low type I error rate while keeping the sample size relatively low when GMR is not controlled to 0.95 or better.
4. When the true GMR is controlled and is not deviating more from unity than 0.95, the sequential design approaches introduced by Potvin *et al.* have the ethical

advantage of better power and/or lower total sample size.

5. When there is uncertainty about the GMR, notably when the true GMR might deviate more from unity than 0.95, at least, method 3 would offer a reasonable way forward which protects against type I error rate inflation.

This work could be supplemented with studies of stopping rules as well of changes to select methods to assume, *e.g.*, GMR=0.90 for the planning of the second trial.

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 10 Aug 2014.
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 10 Aug 2014.
3. Therapeutic Products Directorate, Health Canada. Conduct and Analysis of Comparative Bioavailability Studies. 2012. http://www.hc-sc.gc.ca/dhp-mps/alt_formats/pdf/prodpharma/applic-demande/guide-ld/bio/gd_cbs_ebc_ld-eng.pdf. Accessed 10 Aug 2014.
4. World Medical Association. Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects. 2013. <http://www.wma.net/en/30publications/10policies/b3/>. Accessed 10 Aug 2014.
5. 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.
6. 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.
7. 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.
8. Garcia-Arieta A. The failure to show bioequivalence is not evidence against generics. *Br J Clin Pharmacol*. 2010;70:452–3.
9. Matsumoto M, Nishimura T. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans Model Comput Simul*. 1998;8:3–30.