Published in final edited form as:

Stat Med. 2011 June 30; 30(14): 1637–1647. doi:10.1002/sim.4230.

Maximum inflation of the type 1 error rate when sample size and allocation rate are adapted in a pre-planned interim look.¹

Alexandra C. Graf and Peter Bauer²

Center for Medical Statistics, Informatics and Intelligent Systems, Spitalgasse 23, 1090 Vienna, Austria

Abstract

We calculate the maximum type 1 error rate of the pre-planned conventional fixed sample size test for comparing the means of independent normal distributions (with common known variance) which can be yielded when sample size and allocation rate to the treatment arms can be modified in an interim analysis. Thereby it is assumed that the experimenter fully exploits knowledge of the unblinded interim estimates of the treatment effects in order to maximize the conditional type 1 error rate. The "worst case" strategies require knowledge of the unknown common treatment effect under the null hypothesis. Although this is a rather hypothetical scenario it may be approached in practice when using a standard control treatment for which precise estimates are available from historical data. The maximum inflation of the type 1 error rate is substantially larger than derived by [1] for design modifications applying balanced samples before and after the interim analysis. Corresponding upper limits for the maximum type 1 error rate are calculated for a number of situations arising from practical considerations (e.g. restricting the maximum sample size, not allowing sample size decreases, allowing only increase of the sample size in the experimental treatment). The application is discussed for a motivating example.

Keywords

Sample size reassessment; changing allocation rates; interim analysis; maximum type 1 error rate; z-test; conditional error function

1. Introduction

Sample size reassessment in an adaptive interim analysis based on an interim estimate of the effect size can considerably increase the type 1 error rate if the pre-planned level α test is applied for the final analysis. For sample sizes balanced between treatments Proschan and Hunsberger [1] investigated the maximum type 1 error rate inflation of the one-sided test for the comparison of two normal means (variance known). They showed that the type 1 error rate may be inflated from 0.05 to 0.1146 if in an interim analysis the second stage sample size is chosen to maximize the conditional error function of the final one-sided z-test.

¹This article was published 2011 in Statistics in Medicine, Volume 30, Issue 14, <http://onlinelibrary.wiley.com/doi/10.1002/sim.4230/full>

²peter.bauer@meduniwien.ac.at

Adaptive combination tests (see [2], [3], [4], [5]) and tests based on the conditional error function (see [1], [6], [7]) have been proposed which allow for such interim design modifications without inflating the type 1 error rate. For a short review of sample size re-estimation in clinical trials see e.g. [8].

The literature to date for the z-test has assumed that the allocation ratios between the control and the selected treatment groups are the same before and after the interim analysis. Under this assumption the conditional error function does not depend on the unknown nuisance parameter, the common mean under the null hypothesis. This paper investigates how much the type 1 error rate may be inflated if conventional tests are used when not only the sample size but also the individual allocation ratios for the control and treatment group can be adapted in an interim analysis. In the following we investigate the impact on the maximum overall type 1 error rate by including this type of adaptation in flexible two-stage designs. This problem is important because the application of the conditional error principal to control the overall type 1 error rate is complicated when nuisance parameters are involved (for examples see [9], [10], [11]). Note that in the regulated environment of pharmaceutical drug development strict control of the type 1 error rate is one of the key issues (e.g. [12], [13]). We acknowledge, that in the current European guidance on adaptive designs for clinical trials the problem of treatment effect heterogeneity across stages is emphasized (see [13]). Changing the randomization ratio might lead to the recruitment of slightly different patient populations which in turn might lead to differences in the treatment effects. One way to deal with the problem is to explore the interaction between treatment and stage by appropriate test statistics (see e.g. [14]). However, in the following we will not account for such pre-test procedures in our methodology.

The paper is organized as follows: Section 2 starts with a motivating example. In Section 3 we calculate the maximum overall type 1 error rate for the two-sample z-test allowing full flexibility for the choice of the second stage sample sizes in the treatment and control group. In Section 4 we investigate the impact of adaptation constraints on the maximum overall type 1 error rate and present the applications to the motivating example. Section 5 considers a scenario with a historical control group. Section 6 follows with concluding remarks.

2 Motivating Example

Assume that a randomized clinical trial to compare an experimental and a standard treatment has been pre-planned with an interim analysis to be performed when the continuous primary outcome variable has become available for one half of the total pre-planned number of patients. A one-sided test for superiority of the experimental treatment at the level 0.025 is planned for the final statistical analysis. (We assume that the sample sizes are large enough so that the distributional assumption for the test statistics in this manuscript apply to a good approximation.) Restricted randomization is applied to ensure that sample sizes are balanced between treatments at both stages. The only purpose of the interim analysis is to take the opportunity of stopping for futility because of lack of evidence for a sufficient treatment difference, which is a scenario of increasing practical importance. Since such decisions generally account also for aspects other than efficacy, like patient safety and burden, no formal stopping rule for the primary outcome variable has been pre-defined. The plan

excludes reducing the pre-planned total sample size if the study proceeds to the second stage because a certain amount of information on safety of the experimental treatment in comparison with the information on safety of the standard treatment is needed for a reasonably accurate assessment of the relationship of risk to benefit.

In the interim analysis concern arises if a signal of a possible unexpected adverse effect occurs in the experimental treatment group. The data safety monitoring board of the study strongly suggests an extension of the second stage sample size in order to collect more information on this emerging issue. Sufficient information on the safety profile of the standard treatment is available from the past so that increasing the sample size of the control group is not necessary. Moreover, financial resources limit the increase in sample size to no more than 50%.

At the interim analysis, it is decided to recruit up to a tripling of the planned sample size for the second stage for the experimental treatment but to leave the second stage sample size in the standard treatment group unchanged.

One of the reviewers pointed out another scenario that fits the paradigm of this paper: Response-adaptive randomization (see [15]) is based on two essential features, seeing ongoing results and changing probabilities accordingly which can change the allocation ratio substantially. Although response-adaptive randomization does this continually, the motivating example can be thought of as specific version of response-adaptive randomization where the adaptation does not occur until a substantial amount of data is available. As we will see in Section 3 below, some of the strategies of sample size reallocation to maximize the type 1 error rate would also result from a response-adaptive reasoning: e.g. if at interim a poor outcome has been observed in the control group patients should be further treated preferably with the superior treatment.

3 The maximum overall type 1 error rate for the two-sample z-test

We consider the comparison of the means of two normal distributions with common known variance σ^2 ($= 1$ without loss of generality) in a one-sided test of the null hypothesis

$$H_0 : \mu_1 = \mu_0 \text{ versus the alternative } H_1 : \mu_1 > \mu_0,$$

where the index 0 indicates the control group. The assumption of known variance is frequently applied in practice to hold at least asymptotically when using consistent variance estimators. After $n_0^{(1)}$ observations in the control and $n_1^{(1)} = r^{(1)} n_0^{(1)}$ observations in the treatment group an interim analysis is performed. In the following we use bracketed upper indices for denoting stages if necessary. In the interim look the experimenter may adapt the second stage sample sizes to $n_1^{(2)} = r_1^{(2)} n_1^{(1)} = r_1^{(2)} r^{(1)} n_0^{(1)}$ in the treatment and $n_0^{(2)} = r_0^{(2)} n_0^{(1)}$ in the control group with $0 \leq r_0^{(2)}, r_1^{(2)} \leq \infty$. Then the conditional type 1 error rate can be written as (see Appendix A.1):

$$CE_{\alpha}(r^{(1)}, r_0^{(2)}, r_1^{(2)}, Z_0^{(1)}, Z_1^{(1)}) = 1 - \Phi \left[\frac{z_{1-\alpha} \sqrt{\frac{1}{r^{(1)}(1+r_1^{(2)})} + \frac{1}{1+r_0^{(2)}}} - \frac{Z_1^{(1)}}{\sqrt{r^{(1)}(1+r_1^{(2)})} + \frac{Z_0^{(1)}}{1+r_0^{(2)}}}}{\sqrt{\frac{r_1^{(2)}}{r^{(1)}(1+r_1^{(2)})^2} + \frac{r_0^{(2)}}{(1+r_0^{(2)})^2}}}} \right]$$

(1)

where $Z_1^{(1)} = (\bar{X}_1^{(1)} - \mu) \sqrt{n_1^{(1)}}$ and $Z_0^{(1)} = (\bar{X}_0^{(1)} - \mu) \sqrt{n_0^{(1)}}$ have independent standard normal distributions, $\bar{X}_1^{(1)}$ and $\bar{X}_0^{(1)}$ are the first stage sample means in the treatment and control group respectively. Here μ denotes the common true mean under the null hypothesis, Φ denotes the normal cumulative distribution function, $z_{1-\alpha}$ the $(1 - \alpha)$ -quantile of the standard normal distribution and α the unadjusted nominal significance level to be applied in the final analysis. The notation in formula (1) indicates that all the test statistics in the following are defined as standardized differences of sample means to the true common mean μ under the global null hypothesis. Consequently, when calculating the maximum overall error rate we pretend that the experimenter has knowledge of the common mean μ , without loss of generality we assume $\mu = 0$. Although this assumption may appear unrealistic in practice, consider a clinical trial comparing a new treatment with a standard treatment. This standard treatment has been applied in large sets of patients so that the expected effect size is known fairly accurately from historical data. This situation mimics the knowledge of the common unknown mean under the null hypothesis. Such an assumption that the effect of the control group is well established is made, e.g. when demonstrating superiority against a putative placebo by showing non-inferiority against an active control treatment (see e.g. [16]).

If $r^{(1)} = 1$ and $r_1^{(2)} = r_0^{(2)} = r^{(2)}$ formula (1) reduces to the formula given by [1]. The maximum overall type 1 error rate achievable by sample size reassessment and reallocation is given by:

$$E_{\alpha}^*(r^{(1)}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \widetilde{CE}_{\alpha}(r^{(1)}, Z_0^{(1)}, Z_1^{(1)}) \phi(Z_0^{(1)}) \phi(Z_1^{(1)}) dZ_1^{(1)} dZ_0^{(1)} \quad (2)$$

where $\widetilde{CE}_{\alpha}(r^{(1)}, Z_0^{(1)}, Z_1^{(1)}) = \max_{0 \leq r_0^{(2)}, r_1^{(2)} \leq \infty} [CE_{\alpha}(r^{(1)}, r_0^{(2)}, r_1^{(2)}, Z_0^{(1)}, Z_1^{(1)})]$. Here ϕ denotes the density function of the standard normal distribution. Let in the following $\hat{r}_i^{(2)}$, $i = 0, 1$ denote the optimal sample size inflation factors to yield the maximal conditional type 1 error rate $\widetilde{CE}_{\alpha}$ for a given $Z_0^{(1)}$ in the control and $Z_1^{(1)}$ in the treatment group at the interim analysis, respectively.

The maximum overall type 1 error rates possible are equal for $r^{(1)} = r$ and $r^{(1)} = \frac{1}{r}$. Calculations of the function show that its maximum occurs when $r^{(1)} = 1$, see Figure 2 at the end of this section. For the behaviour of the function for $r^{(1)} \rightarrow 0$ or by symmetry $r^{(1)} \rightarrow \infty$ see Section 5. Therefore, in the following we focus on the overall worst case scenario $r^{(1)} = 1$ (noting that this is also the one which will be most likely pre-planned in practice) and for simplicity in this case we skip $r^{(1)} = 1$ in the notations. For $r^{(1)} = 1$ it can easily be seen that the maximum conditional error function $\widetilde{CE}_\alpha$ is symmetric around the line $Z_1^{(1)} = -Z_0^{(1)}$. At the end of the section we also indicate how the calculations change for $r^{(1)} \neq 1$.

In the following we try to maximize the type 1 error rate allowing second stage sample sizes tending to 0 or to ∞ . We are aware that these cases are of rather theoretical interest, however, it is our intention to show the most extreme inflation of the type 1 error rate. In the next section and also in the example we will consider constraints on the second stage sample sizes referring to potentially realistic situations in practice.

To simplify the integration (2) we can split it in several parts (see Figure 1).

I. For $Z_1^{(1)} > z_{1-\alpha}$ the strategy to yield a conditional type 1 error rate of 1 is given by $\tilde{r}_0^{(2)} = \infty$ and $\tilde{r}_1^{(2)} = 0$. In this case the first stage standardized treatment mean would suffice to reach significance in a test against the fixed value $\mu = 0$. For $r_0^{(2)} \rightarrow \infty$ and $r_1^{(2)} \rightarrow 0$ the numerator of the argument of the function Φ in formula (1) tends to a negative number and the denominator tends to 0 (see formula (1)). Hence the conditional error tends to 1. Similar arguments apply for $Z_0^{(1)} < -z_{1-\alpha}$, where the strategy $\tilde{r}_1^{(2)} = \infty$ and $\tilde{r}_0^{(2)} = 0$ would lead to a conditional type 1 error rate of 1 (see areas I in Figure 1). The regions $\{Z_0^{(1)} < -z_{1-\alpha}\}$ and $\{Z_1^{(1)} > z_{1-\alpha}\}$ are overlapping. The contribution to the maximum type 1 error rate over the two overlapping regions I is therefore $2\alpha - \alpha^2$.

In the subspace of region I where $Z_1^{(1)} > Z_0^{(1)} + \sqrt{2}z_{1-\alpha}$ the maximum conditional type 1 error rate $\widetilde{CE}_\alpha(Z_0^{(1)}, Z_1^{(1)}) = 1$ can also be yielded by setting $\tilde{r}_0^{(1)} = \tilde{r}_1^{(1)} = 0$ because the test already rejects in the interim analysis. This is the corresponding subspace in region I above the dashed line in Figure 1 which indicates the rejection boundary in the interim analysis.

Note that in this region I the maximization of the conditional error in (1) with $r^{(1)} = 1$ leads to the same solution since for $r_0^{(2)} \rightarrow \infty$ and $r_1^{(2)} \rightarrow 0$ or $r_1^{(2)} \rightarrow \infty$ and $r_0^{(2)} \rightarrow 0$ the conditional error function does not depend on $r^{(1)}$.

II. For $Z_1^{(1)} \leq 0$ and $Z_0^{(1)} \geq 0$ (see rectangle II in Figure 1) obviously the largest conditional type 1 error rate achievable is $\widetilde{CE}_\alpha(Z_0^{(1)}, Z_1^{(1)}) = \alpha$ by letting $r_0^{(2)} \rightarrow \infty$ and $r_1^{(2)} \rightarrow \infty$ (see Appendix A.2). This means that the first stage negative treatment effect will not play any role in the final test decision since it is overruled by the new large experiment with type 1

error rate α started at interim. Hence the contribution to the maximum overall type 1 error rate in the lower right quadrant of Figure 1 will be $\frac{1}{4}\alpha$.

Again the same solution arises for $r^{(1)} = 1$.

III. We next look only at the half open strip $Z_0^{(1)} \geq 0$ and $0 \leq Z_1^{(1)} < z_{1-\alpha}$ (the same result will apply for the strip $Z_1^{(1)} \leq 0$ and $-z_{1-\alpha} < Z_0^{(1)} \leq 0$ because of symmetry arguments, see areas III in Figure 1). Obviously in the strategy to maximize the type 1 error rate we take $\tilde{r}_0^{(2)} = \infty$. Thus the optimization does not depend on the value of $Z_0^{(1)}$ if $Z_0^{(1)} \geq 0$ (see Appendix A.3). For $\tilde{r}_0^{(2)} = \infty$ the value $\tilde{r}_1^{(2)}$ maximizing the conditional error rate is

$$\tilde{r}_1^{(2)} = \left(\frac{z_{1-\alpha}}{Z_1^{(1)}} \right)^2 - 1.$$

The corresponding worst case conditional error function is the circular conditional error function of [1]:

$$\widetilde{CE}_\alpha(Z_0^{(1)}, Z_1^{(1)}) = 1 - \Phi \left(\sqrt{z_{1-\alpha}^2 - (Z_1^{(1)})^2} \right).$$

Hence, integrating over $Z_1^{(1)}$ between 0 and $z_{1-\alpha}$, the total contribution to the overall maximum type 1 error rate in the two stripes III is given by

$$\frac{1}{4} \left(e^{-\frac{z_{1-\alpha}^2}{2}} - 2\alpha \right).$$

The same solution holds for $r^{(1)} = 1$, the arguments of the proof in the Appendix A.3 apply to this case as well.

IV. For the remaining white area IV in Figure 1 we used pointwise numerical optimization and numerical integration (using R-packages `optim` and `integrate`) to find the values of the worst case conditional type 1 error rate. Note that setting the derivatives of $\widetilde{CE}_\alpha$ with respect to $r_0^{(2)}$ and $r_1^{(2)}$ to zero leads to equations which cannot be solved analytically.

This is the only part of the sample space where the calculations differ for different values of the first stage allocation rate $r^{(1)}$. For example, the boundary for early rejection changes to $Z_1^{(1)} = z_{1-\alpha} \sqrt{r^{(1)} + 1} + \sqrt{r^{(1)}} Z_0^{(1)}$ and $\widetilde{CE}_\alpha$ is not anymore symmetric around $Z_0^{(1)} = -Z_1^{(1)}$ in this area. Figure 2 shows $E_\alpha^*(r^{(1)}) / E_\alpha^*(1)$ for varying values of $r^{(1)}$ and different α . Large differences between the solutions are only found for very low and very large values of $r^{(1)}$. The case of $r^{(1)} = 0$ is discussed in Section 5 below.

Summing up the above results for $r^{(1)} = 1$ the maximal conditional type 1 error rate is

$$E_{\alpha}^* = 2\alpha - \alpha^2 + \frac{\alpha}{4} + \frac{1}{4} \left(e^{-\frac{z_{1-\alpha}^2}{2}} - 2\alpha \right) + \int_{-z_{1-\alpha}}^0 \int_0^{z_{1-\alpha}} \widetilde{CE}_{\alpha} \left(Z_0^{(1)}, Z_1^{(1)} \right) \phi \left(Z_1^{(1)} \right) \phi \left(Z_0^{(1)} \right) dZ_1^{(1)} dZ_0^{(1)}$$

where $\widetilde{CE}_{\alpha} \left(Z_0^{(1)}, Z_1^{(1)} \right) = 1$ in the triangle

$\{0 \geq Z_0^{(1)} > -z_{1-\alpha}, z_{1-\alpha} > Z_1^{(1)} \geq 0, Z_1^{(1)} > Z_0^{(1)} + \sqrt{2}z_{1-\alpha}\}$ which is the small triangle above the dashed line in the white area IV in Figure 1.

In summary, for $\alpha = 0.05$ the maximum overall type 1 error rate is $E_{0.05}^* = 0.1867$, for $\alpha = 0.025$ it is $E_{0.025}^* = 0.1064$ and for $\alpha = 0.01$ it is $E_{0.01}^* = 0.0491$.

4 Constrained second stage sample sizes

The maximum overall type 1 error rate described above is calculated without imposing any constraints on the sample size adaptation of the second stage, so that we allow early rejection in the interim analysis and infinite sample sizes at the second stage. Here the lower boundaries for the sample size inflation factors have been $r_{0,lo}^{(2)} = r_{1,lo}^{(2)} = 0$ and the upper

boundaries $r_{0,up}^{(2)} = r_{1,up}^{(2)} = \infty$ for the control and treatment group respectively. However, in reality the second stage sample size is constrained. [17] derived formulas for the maximum conditional error function applying constraints on the sample sizes balanced before and after the interim analysis. In the following we provide numerical calculations if the allocation rate may be adapted after the interim analysis. Figure 3 shows the maximum conditional type 1

error rate for a varying upper boundary $1 \leq r_{1,up}^{(2)} \leq 10$ (allowing up to a 10 times larger second stage sample size as compared to the first stage in the treatment group) and different restrictions on the second stage sample size of the control group. We calculated the numbers by numerical optimization and integration over the interim sample space. The figure shows results for $\alpha = 0.01, 0.025$ and 0.05 . Clearly, the maximum overall type 1 error rate increases

with increasing $r_{1,up}^{(2)}$. It is important to realize that the maximum overall type 1 error rate strikingly increases when the second stage sample size can be smaller than the first stage sample size, in particular when allowing for early rejection in the interim analysis, i.e. setting the second stage sample size to 0 (dashed as compared to solid lines). The dashed lines show the maximum overall type 1 error rate when no sample size reduction at the second stage is allowed as compared to the first stage for the treatment and control group respectively ($r_{0,lo}^{(2)} = r_{1,lo}^{(2)} = 1$) and $r_{0,up}^{(2)} = r_{1,up}^{(2)}$. Clearly, if $r_{0,up}^{(2)} = r_{1,up}^{(2)} = r_{0,lo}^{(2)} = r_{1,lo}^{(2)} = 1$ the maximum overall type 1 error rate is α since no adaptation is allowed for the second stage sample sizes (values at the dotted and dashed lines for $r_{1,up}^{(2)} = 1$ in Figure 3). Allowing for early rejection ($r_{0,lo}^{(2)} = r_{1,lo}^{(2)} = 0$) and not allowing sample size increase as compared to the first stage ($r_{0,up}^{(2)} = r_{1,up}^{(2)} = 1$) would increase the maximum overall type 1 error rate to 0.0224 for α

= 0.01 as well as to 0.0516 for $\alpha = 0.025$ and to 0.0959 for $\alpha = 0.05$ respectively. These are the values at the solid curves for $r_{1,\text{up}}^{(2)}=1$ in Figure 3.

4.1 The motivating example revisited

We now investigate the scenario described in Section 2: There is no reason from the beginning to change the sample size in the control group because sufficient information of the safety on the control therapy is known from previous experiments. A larger sample size would be helpful to get more information on the arising safety issue in the experimental treatment group. However, external financial constraints do not allow an increase of the planned second stage sample size to more than threefold. A reduction of the sample size in the experimental treatment group is not considered since a certain amount on treatment safety is required.

Since stopping for futility has not been pre-planned to be used for constructing decision boundaries the example corresponds to the situation shown in Figure 3 with a fixed $r_{0,\text{lo}}^{(2)}=r_{0,\text{up}}^{(2)}=1$ and $r_{1,\text{lo}}^{(2)}=1 \leq r_1^{(2)} \leq r_{1,\text{up}}^{(2)}=3$. Here the maximum type 1 error rate assuming knowledge of the common effect size in the interim analysis of the two equally effective treatments is 0.0304 (see Figure 3). If a formal stopping for futility is included into the probability calculations, e.g. proceed to the second stage only if the interim estimate of the treatment difference points in the correct direction ($Z_0^{(1)} < Z_1^{(1)}$) the worst case type 1 error rate is 0.0294. The maximum is not reduced markedly by this rule since the conditional error in case of $Z_0^{(1)} \leq Z_1^{(1)}$ will be small. Only if we sharpen the condition for futility stopping to $Z_1^{(1)} \leq Z_0^{(1)} + 1.064$ (with a standardized treatment difference not exceeding 0.735) would the above option of sample size adaptation lead to a maximal type 1 error rate not exceeding 0.025 in the conventional test.

These numbers may have different consequences. The strict experimenter may decide to use the above worst case conditional error rate approach to correct for the final decision boundary. Assuming that the constraint on the maximum second stage sample size does not depend on the observed outcome data, he would have to change the nominal error rate α in the final analysis to a value $\alpha' = 0.0204$ so that the worst case overall error rate for all possible second stage designs would not exceed the predefined 0.025. This price may be acceptable, since increasing the sample size gains power. Note however, that strictly this adjusted nominal error rate α' would have to be applied as an upper bound for this flexible type of sample size reassessment even when no adaptation is performed. We do not believe that using the nominal significance level of 0.025 in the situation when no adaptations are performed in any realistic scenario will inflate the actual type 1 error rate. However, in case of changing the allocation ratios a strict preservation of the conditional error rate is not possible in practice because the rate depends on the unknown nuisance parameter and therefore cannot be calculated. Methods like conditioning on the sufficient statistics for the nuisance parameter (see [10]) or taking weighted averages over the unknown nuisance parameters in the interim analysis must be applied in this case (see [11]).

Another experimenter may argue that this adaptation is not related to any efficacy data and therefore he may safely stay with the conventional analysis at the end. However, since safety and efficacy may be correlated, this argument is not stringent. For methods allowing for correlation between outcome variables see e.g. [18]. In order to demonstrate that such an adaptation may not lead to a dramatic increase in the overall type 1 error rate even if efficacy data and the value of the nuisance parameter μ were known the experimenter may include the above calculations when summarizing the study outcome.

An R-program to calculate the maximum conditional error rate with or without constraints on the second stage sample sizes is available under <http://statistics.msi.meduniwien.ac.at/index.php>

5 Historical control group

Let us now consider the limiting case where the first stage has no treatment group ($r^{(1)} = 0$). This case would apply, for example, if a recent sample under a standard treatment is available and will be included in the final analysis of the forthcoming experiment to compare a new treatment to the standard treatment (admittedly this would not be a sound and persuasive strategy in practice). We acknowledge that comparisons with historical controls requires some adjustment for potential confounders which for simplicity will not be considered here. The second stage sample sizes can again be chosen in an arbitrary way such that $0 \leq n_0^{(2)} \leq \infty$ and for the treatment group $0 < n_1^{(2)} \leq \infty$ (note that the second stage sample size of the treatment group has to be larger than 0 to allow a treatment-control comparison). Because of $n_1^{(1)} = 0$ we present the formulas in terms of $n_i^{(2)}$ instead of $r_i^{(2)}$, $i = 0, 1$. The conditional error can then be calculated as

$$CE_{\alpha}(n_0^{(2)}, n_1^{(2)}, Z_0^{(1)}) = 1 - \Phi \left(\frac{z_{1-\alpha} \sqrt{\frac{1}{n_1^{(2)}} + \frac{1}{n_0^{(1)} + n_0^{(2)}}} + Z_0^{(1)} \frac{\sqrt{n_0^{(1)}}}{n_0^{(1)} + n_0^{(2)}}}{\sqrt{\frac{1}{n_1^{(2)}} + \frac{n_0^{(2)}}{(n_0^{(1)} + n_0^{(2)})^2}}} \right) \quad (3)$$

(see Appendix A.4). The maximum overall type 1 error rate for this scenario can then be calculated by

$$E_{\alpha}^* = \int_{-\infty}^{\infty} \widetilde{CE}_{\alpha}(Z_0^{(1)}) \phi(Z_0^{(1)}) dZ_0^{(1)} \quad (4)$$

where $\widetilde{CE}_{\alpha}(Z_0^{(1)}) = \max_{0 \leq n_0^{(2)}, n_1^{(2)} \leq \infty} CE_{\alpha}(n_0^{(2)}, n_1^{(2)}, Z_0^{(1)})$. Again, we tried to maximize the type 1 error rate allowing second stage sample sizes tending to 0 or to ∞ to show the most extreme inflation of the type 1 error rate.

Splitting the integration (4) into 3 parts $\{Z_0^{(1)} \geq 0\}$, $\{-z_{1-\alpha} < Z_0^{(1)} < 0\}$ and $\{Z_0^{(1)} < -z_{1-\alpha}\}$ and using similar arguments as before the analytic solution for the maximum overall type 1 error rate is the same as for the z-test with constant allocation ratios derived by [1], i.e.

$E_{0.01}^* = 0.0267$ for $\alpha = 0.01$, $E_{0.025}^* = 0.0616$ for $\alpha = 0.025$ and $E_{0.05}^* = 0.1146$ for $\alpha = 0.05$.

6 Discussion

We have calculated the maximum overall type 1 error rate for the z-test (comparing two normal means, common standard deviation known) that may be reached if the sample sizes per treatment arm can be modified in an interim analysis utilizing the treatment effects observed. [1] solved this problem under the restriction of identical allocation rates between the treatment groups before and after the interim analysis. In this case the solution only depends on the first stage z-statistics. If the reallocation ratio changes in the interim analysis the maximum type 1 error rate depends on the deviation of the observed estimates from the nuisance parameter involved (the unknown common effect of all the treatments under the null hypothesis). Hence in calculating the type 1 error rate under the worst case scenario we assume that the experimenter knows the nuisance parameter. This assumption appears unrealistic although there may be situations in practice where precise estimates may exist for the effect in the control group from historical data (and thus also for the nuisance parameter under the null hypothesis). We tried to find explicit solutions for the worst case choice of the second stage sample sizes for different interim outcomes. We are aware that allowing second stage sample sizes tending to 0 or ∞ is not a realistic scenario. However, it was our intention to show the most extreme inflation of the type 1 error rate. As a consequence the actual type 1 error rates resulting from changes in sample sizes per treatment group in the interim analysis when the estimates of the treatment effects are available in practice may be well below these worst case solutions. By considering constraints on these sample size we have investigated more realistic situations, e.g. when the second stage allocation ratio of treatment to control will never be below 1. Not surprisingly the option of reducing the pre-planned sample size in case of observed large treatment differences, in the most extreme case to allow rejection of the null hypotheses already in the interim analysis, has a major impact on the magnitude of the yielded type 1 error rate. We also looked at the rather hypothetical situation that only the control group is investigated before the interim analysis ("historical control") which leads to the same solution as found by [1]. We did not look at situations with variance heterogeneity and other distributions than the normal distribution for the outcome variables because direct formulas for the conditional error are hardly accessible in these more general settings.

We have mentioned that response-adaptive randomization may be looked at as a special case of the types of adaptation considered in our paper. We have shown an example of how to use the calculations heuristically to account for such type of adaptations. This may be useful in applications, because in the presence of nuisance parameters, applying the principle of preserving the conditional type 1 error rate for the exact control of the overall type 1 error rate is not straightforward.

Acknowledgements

We thank Werner Brannath and Martin Posch for their helpful suggestions. We also thank an associate editor and two reviewers for their useful comments. This work was supported by the Austrian Science Fund (FWF) no. P23167.

A Appendix

In the following we assume that the null hypothesis applies ($\mu_0 = \mu_1 = \mu$).

A.1

In the final analysis after the second stage the test is based on the following global test statistic:

$$\begin{aligned}
 Z^{(2)} &= \frac{\frac{n_1^{(1)}\bar{x}_1^{(1)} + n_1^{(2)}\bar{x}_1^{(2)}}{n_1^{(1)} + n_1^{(2)}} - \frac{n_0^{(1)}\bar{x}_0^{(1)} + n_0^{(2)}\bar{x}_0^{(2)}}{n_0^{(1)} + n_0^{(2)}}}{\sqrt{\frac{1}{n_1^{(1)} + n_1^{(2)}} + \frac{1}{n_0^{(1)} + n_0^{(2)}}}} \\
 &= \frac{\left(\bar{x}_1^{(1)} - \mu\right)n_1^{(1)} + \left(\bar{x}_1^{(2)} - \mu\right)n_1^{(2)}}{n_1^{(1)} + n_1^{(2)}} - \frac{\left(\bar{x}_0^{(1)} - \mu\right)n_0^{(1)} + \left(\bar{x}_0^{(2)} - \mu\right)n_0^{(2)}}{n_0^{(1)} + n_0^{(2)}}}{\sqrt{\frac{1}{n_1^{(1)} + n_1^{(2)}} + \frac{1}{n_0^{(1)} + n_0^{(2)}}}} \\
 &= \frac{\frac{Z_1^{(1)} + \sqrt{r_1^{(2)}}Z_1^{(2)}}{\sqrt{r^{(1)}}(1+r_1^{(2)})} - \frac{Z_0^{(1)} + \sqrt{r_0^{(2)}}Z_0^{(2)}}{1+r_0^{(2)}}}{\sqrt{\frac{1}{r^{(1)}(1+r_1^{(2)})} + \frac{1}{1+r_0^{(2)}}}}
 \end{aligned}$$

where $Z_i^{(j)} = (\bar{x}_i^{(j)} - \mu) \sqrt{n_i^{(j)}}$, $i = 0, 1, j = 1, 2$, have independent standard normal distributions. If $Z^{(2)}$ is larger than $z_{1-\alpha}$ we get a false positive decision, which leads to the following inequality:

$$Y = \frac{\frac{Z_1^{(1)} + \sqrt{r_1^{(2)}}Z_1^{(2)}}{\sqrt{r^{(1)}}(1+r_1^{(2)})} - \frac{Z_0^{(1)} + \sqrt{r_0^{(2)}}Z_0^{(2)}}{1+r_0^{(2)}}}{\sqrt{\frac{1}{r^{(1)}(1+r_1^{(2)})} + \frac{1}{1+r_0^{(2)}}}} > z_{1-\alpha}$$

Since $Z_1^{(2)}$ and $Z_0^{(2)}$ have independent standard normal distributions for every set of values $r^{(1)}, r_0^{(2)}, r_1^{(2)}, Z_0^{(1)}$ and $Z_1^{(1)}$ (and hence are independent of these quantities)

$$Y \sim N \left[0, \frac{r_1^{(2)}}{r^{(1)}(1+r_1^{(2)})^2} + \frac{r_0^{(2)}}{(1+r_0^{(2)})^2} \right]$$

Hence the conditional error can be written as in formula (1).

A.2

The conditional error is maximized if the following expression is minimized (note that $r^{(1)} = 1$):

$$z_{1-\alpha} \sqrt{\frac{\frac{1}{1+r_1^{(2)}} + \frac{1}{1+r_0^{(2)}}}{\frac{r_1^{(2)}}{(1+r_1^{(2)})^2} + \frac{r_0^{(2)}}{(1+r_0^{(2)})^2}}} + \frac{\frac{-Z_1^{(1)}}{1+r_1^{(2)}} + \frac{Z_0^{(1)}}{1+r_0^{(2)}}}{\sqrt{\frac{r_1^{(2)}}{(1+r_1^{(2)})^2} + \frac{r_0^{(2)}}{(1+r_0^{(2)})^2}}} \quad (5)$$

In the rectangle $Z_1^{(1)} \leq 0$ and $Z_0^{(1)} \geq 0$ (area II in Figure 1) the second term is 0 or greater and tends to 0 if both $r_0^{(2)} = r_1^{(2)} \rightarrow \infty$. In the first term the denominator is at least as large as the numerator:

$$\frac{1}{1+r_1^{(2)}} + \frac{1}{1+r_0^{(2)}} - \frac{r_1^{(2)}}{(1+r_1^{(2)})^2} - \frac{r_0^{(2)}}{(1+r_0^{(2)})^2} = \frac{1}{r_1^{(2)} \left(1 + \frac{1}{r_1^{(2)}}\right)} \left(1 - \frac{1}{1 + \frac{1}{r_1^{(2)}}}\right) + \frac{1}{r_0^{(2)} \left(1 + \frac{1}{r_0^{(2)}}\right)} \left(1 - \frac{1}{1 + \frac{1}{r_0^{(2)}}}\right) \geq 0$$

Equality can be achieved by letting $r_0^{(2)} \rightarrow \infty$ and $r_1^{(2)} \rightarrow \infty$ so that the first term tends to $z_{1-\alpha}$. Thus, the maximal conditional error $\widetilde{CE}_\alpha$ for each $Z_1^{(1)}$ and $Z_0^{(1)}$ in this rectangle is $1 - \Phi(Z_{1-\alpha}) = \alpha$. Therefore the contribution to the maximum overall type 1 error rate of this quadrant $\{Z_0^{(1)} \geq 0, Z_1^{(1)} \leq 0\}$ is $\frac{\alpha}{4}$.

A.3

If $0 \leq Z_1^{(1)} < z_{1-\alpha}$ and $Z_0^{(1)} \geq 0$ (half of region III in Figure 1) then

$$\begin{aligned} & \max_{0 \leq r_0^{(2)}, r_1^{(2)} \leq \infty} CE_\alpha \left(r_0^{(2)}, r_1^{(2)}, Z_0^{(1)}, Z_1^{(1)} \right) \\ &= CE_\alpha \left(\tilde{r}_0^{(2)}, \tilde{r}_1^{(2)}, Z_0^{(1)}, Z_1^{(1)} \right) \leq CE_\alpha \left(\tilde{r}_0^{(2)}, \tilde{r}_1^{(2)}, 0, Z_1^{(1)} \right) \leq \max_{0 \leq r_0^{(2)}, r_1^{(2)} \leq \infty} CE_\alpha \left(r_0^{(2)}, r_1^{(2)}, 0, Z_1^{(1)} \right) \end{aligned}$$

(6)

Setting $Z_0^{(1)} = 0$ in (5) and using the result for the first term derived in the previous section it can be seen easily that for $0 \leq Z_1^{(1)}$ the conditional error is monotone in $r_0^{(2)}$. For all $Z_1^{(1)}$ and $r_1^{(2)}$ the conditional error has a maximum for $r_0^{(2)} \rightarrow \infty$. For $\tilde{r}_0^{(2)} = \infty$ we get for the conditional error

$$1 - \Phi \left[z_{1-\alpha} \sqrt{\frac{1+r_1^{(2)}}{r_1^{(2)}}} - Z_1^{(1)} \sqrt{\frac{1}{r_1^{(2)}}} \right]$$

which is independent of $Z_0^{(1)}$ and can be easily maximized in $r_1^{(2)}$, as in [1], leading to

$$\hat{r}_1^{(2)} = \left(\frac{z_{1-\alpha}}{Z_1^{(1)}} \right)^2 - 1$$

The corresponding worst case conditional error function is then the circular conditional error

function $1 - \Phi \left(\sqrt{z_{1-\alpha}^2 - (Z_1^{(1)})^2} \right)$. From (5) it follows that, given $Z_1^{(1)}$, the same solution

can be yielded for every $Z_0^{(1)}$ if $r_0^{(2)} \rightarrow \infty$ which for given value of $Z_1^{(1)}$ according to

Appendix 2 is the maximum achievable over the whole strip. As a consequence for $r_0^{(2)} \rightarrow \infty$ all the inequalities in (6) become equalities. Thus, the maximal conditional type 1 error rate over this area can be calculated as

$$\int_0^\infty \int_0^{z_{1-\alpha}} \left[1 - \Phi \left(\sqrt{z_{1-\alpha}^2 - (Z_1^{(1)})^2} \right) \right] \phi(Z_0^{(1)}) \phi(Z_1^{(1)}) dZ_1^{(1)} dZ_0^{(1)} = \frac{1}{2} \frac{1}{4} \left(\exp \left(-\frac{z_{1-\alpha}^2}{2} \right) - 2\alpha \right),$$

see [1]. For reasons of symmetry, the same result can be achieved for the strip

$-z_{1-\alpha} < Z_0^{(1)} \leq 0$ and $0 \geq Z_1^{(1)}$ where the optimal choice would be $\hat{r}_1^{(1)} = \infty$.

A.4

Assuming $r^{(1)} = 0$, after the second stage the test is based on the following test statistic:

$$Z^{(2)} = \frac{\frac{n_1^{(2)} x_1^{(2)}}{n_1^{(2)}} - \frac{n_0^{(1)} x_0^{(1)} + n_0^{(2)} x_0^{(2)}}{n_0^{(1)} + n_0^{(2)}}}{\sqrt{\frac{1}{n_1^{(2)}} + \frac{1}{n_0^{(1)} + n_0^{(2)}}}} = \frac{\frac{Z_1^{(2)}}{\sqrt{n_1^{(2)}}} - \frac{Z_0^{(1)} \sqrt{n_0^{(1)}} + Z_0^{(2)} \sqrt{n_0^{(2)}}}{n_0^{(1)} + n_0^{(2)}}}{\sqrt{\frac{1}{n_1^{(2)}} + \frac{1}{n_0^{(1)} + n_0^{(2)}}}}$$

where $Z_0^{(1)}$, $Z_0^{(2)}$ and $Z_1^{(2)}$ are standard normal distributed. If $Z^{(2)}$ is larger than $z_{1-\alpha}$ we get a false positive decision, which leads to the following inequality:

$$\frac{Z_1^{(2)}}{\sqrt{n_1^{(2)}}} - \frac{\sqrt{n_0^{(2)}} Z_0^{(2)}}{n_0^{(1)} + n_0^{(2)}} > z_{1-\alpha} \sqrt{\frac{1}{n_1^{(2)}} + \frac{1}{n_0^{(1)} + n_0^{(2)}}} + \frac{Z_0^{(1)} \sqrt{n_0^{(1)}}}{n_0^{(1)} + n_0^{(2)}}$$

Since $Z_1^{(2)}$ and $Z_0^{(2)}$ have independent standard normal distributions, independently of $Z_0^{(1)}$, $n_0^{(1)}$, $n_0^{(2)}$, $n_1^{(2)}$ we get (3) for the conditional type 1 error rate.

References

- [1]. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics*. 1995; 51:1315–1324. [PubMed: 8589224]
- [2]. Bauer P. Multistage testing with adaptive designs. *Biometrie und Informatik in Medizin und Biologie*. 1989; 20:130–148.
- [3]. Bauer P, Koehne K. Evaluations of experiments with adaptive interim analysis. *Biometrics*. 1994; 50:1029–1041. [PubMed: 7786985]
- [4]. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics*. 1999; 55:1286–1290. [PubMed: 11315085]
- [5]. Brannath W, Posch M, Bauer P. Recursive combination tests. *JASA*. 2002; 97:236–244.
- [6]. Mueller HH, Schaefer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*. 2001; 95:886–891.
- [7]. Mueller HH, Schaefer H. A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine*. 2004; 23:2497–2508. [PubMed: 15287080]
- [8]. Proschan MA. Sample size re-estimation in clinical trials. *Biometrical Journal*. 2009; 51:348–357. [PubMed: 19358221]
- [9]. Posch M, Timmesfeld N, Koenig F, Mueller HH. Conditional rejection probabilities of students t-test and design adaptations. *Biometrical Journal*. 2004; 46:389–403.
- [10]. Timmesfeld N, Schaefer H, Mueller HH. Increasing the sample size during clinical trials with t-distributed test statistics without inflating the type I error rate. *Statistics in Medicine*. 2006; 26:2449–2464. [PubMed: 17080491]
- [11]. Gtjahr G, Brannath W, Bauer P. A general approach to the conditional error rate principle with nuisance parameters. *Biometrics*. 2010 in press. DOI: 10.1111/j.1541.
- [12]. Committee for Proprietary Medical Products (CPMP). Statistical principles for clinical trials. 1998. CPMP/ICH/363/96
- [13]. Committee for Medicinal Products for Human Use (CHMP). Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. 2007. CHMP/EWP/2459/02
- [14]. Friede T, Henderson R. Exploring changes in treatment effects across design stages in adaptive trials. *Pharmaceutical Statistics*. 2009; 8:62–72. [PubMed: 18381595]
- [15]. Hu, F.; Rosenberger, WF. The theory of response-adaptive randomization in clinical trials. Wiley Interscience; New Jersey: 2006.
- [16]. Hung HM, Wang SJ, O'Neill RT. Challenges and regulatory experiences in non-inferiority trial design without placebo arm. *Biometrical Journal*. 2009; 51:324–334. [PubMed: 19358213]
- [17]. Brannath W, Koenig F, Bauer P. Estimation in flexible two stage designs. *Statistics in Medicine*. 2006; 25:3366–3381. [PubMed: 16220489]
- [18]. Stallard N. A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. *Statistics in Medicine*. 2009; 29:959–971. [PubMed: 20191605]

Figure 1.

Worst case scenarios in partitions of the $(Z_0^{(1)}, Z_1^{(1)})$ -plane.

Figure 2.

$E_{\alpha}^*(r^{(1)})/E_{\alpha}^*(1)$ for different values of $r^{(1)}$. α was set to 0.01 (solid line), 0.025 (dashed line) and 0.05 (dotted line).

Figure 3.

Maximum overall type 1 error rate (E_{α}^*) for varying upper boundary $r_{1,up}^{(2)}$ for the sample size inflation factor $r_1^{(2)}$ in the treatment group. Solid lines: $r_{0,lo}^{(2)}=r_{1,lo}^{(2)}=0$ and $r_{0,up}^{(2)}=r_{1,up}^{(2)}$ (sample size decrease and early rejection possible); dashed lines: $r_{0,lo}^{(2)}=r_{1,lo}^{(2)}=1$ and $r_{0,up}^{(2)}=r_{1,up}^{(2)}$ (no sample size reduction in the second stage); dotted lines: $r_{0,lo}^{(2)}=r_{1,lo}^{(2)}=1$ and $r_{0,up}^{(2)}=1$ (sample

size increase only in treatment group). $\alpha = 0.01$ (black lines with squares), 0.025 (dark grey lines with triangles) and 0.05 (light grey lines with dots).