Statistics
in Medicine

Research Article

Received 29 September 2014, Accepted 24 November 2015 Published online 23 December 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6848

Maximum type I error rate inflation from sample size reassessment when investigators are blind to treatment labels

Magdalena Żebrowska,*† Martin Posch and Dominic Magirr

Consider a parallel group trial for the comparison of an experimental treatment to a control, where the second-stage sample size may depend on the blinded primary endpoint data as well as on additional blinded data from a secondary endpoint. For the setting of normally distributed endpoints, we demonstrate that this may lead to an inflation of the type I error rate if the null hypothesis holds for the primary but not the secondary endpoint. We derive upper bounds for the inflation of the type I error rate, both for trials that employ random allocation and for those that use block randomization. We illustrate the worst-case sample size reassessment rule in a case study. For both randomization strategies, the maximum type I error rate increases with the effect size in the secondary endpoint and the correlation between endpoints. The maximum inflation increases with smaller block sizes if information on the block size is used in the reassessment rule. Based on our findings, we do not question the well-established use of blinded sample size reassessment methods with nuisance parameter estimates computed from the blinded interim data of the primary endpoint. However, we demonstrate that the type I error rate control of these methods relies on the application of specific, binding, pre-planned and fully algorithmic sample size reassessment rules and does not extend to general or unplanned sample size adjustments based on blinded data. © 2015 The Authors. *Statistics in Medicine* Published by John Wiley & Sons Ltd.

Keywords: sample size reassessment; type I error rate control; adaptive clinical trials; random allocation; block randomization; blinded interim analysis

1. Introduction

Blinding is a generally accepted tool to address bias in randomized clinical trials. It ensures that up to the investigated intervention all subjects are handled equally across treatment groups and outcomes are assessed in the same way. Furthermore, blinding of study subjects allows one to distinguish specific treatment effects from potential placebo effects. Blinding is also essential to avert statistical bias in hypotheses testing procedures if data dependent changes to the analysis strategy are made. The ICH E9 guideline [1], for example, recommends to review (and possibly update) the statistical analysis plan based on a blinded data review and notes that "Decisions made at this time should be described in the report, and should be distinguished from those made after the statistician has had access to the treatment codes, as blind decisions will generally introduce less potential for bias". Similarly, in adaptive clinical trials where adaptations of the trial designs such as a reassessment of sample size can be performed after an interim analysis, blinding is important: it is well known that sample size reassessment based on unblinded interim data may lead to inflation of the type I error by more than 100% [2, 3] if the adaptation is not accounted for by using appropriate adaptive testing procedures [4–6]. To address the various sources of bias in adaptive trials, regulatory guidelines [7–9] recommend to avoid breaking the blind and to perform adaptations based on blinded interim analyses instead. An assumption underlying these guidance documents is that

Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, 1090 Vienna , Austria
Correspondence to: Magdalena Żebrowska, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria.
†*E-mail: magdalena.zebrowska@meduniwien.ac.at*

adaptations based on blinded interim analyses are less prone to bias. Indeed, it has been demonstrated in several settings that adaptations based on blinded interim analysis do not require an adjusted analysis to control the type I error:

- for superiority studies comparing normally distributed endpoints in a parallel group design where the sample size is reassessed based on the "lumped variance" (the variance of the total sample pooling the observations from both groups), the type I error rate is essentially unaffected [10–12]),
- also for superiority studies comparing event rates, where the sample size is reassessed based on the overall number of events across treatment groups, no relevant inflation of the type I error rate was observed [13]. Analogous results were obtained also for count data [14],
- if permutation tests are applied, Posch & Proschan [15] and Proschan *et al*. [16] showed that adaptations based on blinded interim data will indeed control the type I error rate if the clinical trial is restricted to a univariate testing problem where a single endpoint is observed. If adaptations are restricted to the choice between endpoints, the result extends to trials where two endpoints are considered simultaneously. Asymptotically, these results are also valid for *t*-tests.

However, blinding is not a panacea to prevent bias. If sample sizes are low, a minor increase of the type I error rate is observed for non-inferiority trials with sample size reassessment based on the lumped variance [17]. Also in superiority tests, Proschan *et al*. [16] showed that for general sample size reassessment rules based on the lumped variance the type I error rate may be inflated for small sample sizes. Furthermore, if the sample size reassessment rule may depend on more than one endpoint, type I error rate control is no longer guaranteed: if the null hypothesis holds for the primary endpoint but not for a secondary endpoint such as, for example, the level of drug in the blood, the secondary endpoint may completely unblind the investigator. However, the bias can also occur in less extreme settings, where the secondary endpoint unblinds the investigator only partially, as may be the case for a safety endpoint. In such settings, the potential type I error rate inflation is similar to that of a clinical trial where adaptations are performed in an unblinded interim analysis without being accounted for in the testing strategy [15].

In this paper, we investigate the potential consequences of blinded sample size reassessment approaches that deviate from the accepted statistical practice of applying a binding, algorithmic, and blinded sample size reassessment procedure for which type I error rate control has been demonstrated. In particular, we consider settings where no blinded sample size reassessment has been pre-specified in the protocol, settings where an option for blinded sample size reassessment (but no binding rule) are pre-specified, and settings where a binding rule have been pre-specified but the data monitoring committee decided not to follow the rule. Sponsors may argue for a more flexible approach for several reasons: for example, the deviation of nuisance parameter estimates from planning assumptions may not have been anticipated in the planning phase; the maximum number of available patients is unknown in advance such that no binding rule can be pre-specified; recruitment is lower than anticipated or safety concerns arise such that it is argued that the pre-planned sample size algorithm cannot be followed; or information from other trials may arise that serves as an argument for a change in pre-specified strategies. Recent regulatory guidance documents appear to acknowledge such unplanned adaptations. For example, the FDA adaptive designs draft guidances state, *"Certain blinded-analysis-based changes, such as sample size revisions based on aggregate event rates or variance of the endpoint, are advisable procedures that can be considered and planned at the protocol design stage, but can also be applied when not planned from the study outset if the study has remained unequivocally blinded."* [8] and *"While it is strongly preferred that such adaptations be preplanned at the start of the study, it may be possible to make changes during the studys conduct as well. In such instances, the FDA will expect sponsors to be able to both justify the scientific rationale why such an approach is appropriate and preferable, and demonstrate that they have not had access to any unblinded data (either by coded treatment groups or completely unblinded) and that the data has been scrupulously safeguarded."* [9]. Unplanned sample size adjustment is also accepted by European regulators in specific settings, see, for example, Case Study 3 in [18].

We consider the setting of a superiority test of a new experimental treatment over control, with a parallel group design and both blocked and unblocked randomization, where the sample size is reassessed after a blinded interim analysis. We assume that blinded data of the primary and a secondary endpoint is observed. This secondary endpoint – which may or may not be correlated with the primary endpoint – could be a surrogate endpoint, a clinical outcome, or a biomarker. For simplicity, the joint distribution of the two endpoints is assumed to be bivariate normal. If the null hypothesis of no treatment effect holds for the primary but not for the secondary endpoint, then the blinded secondary endpoint data provide the investigator with some information about the likely treatment assignment. We quantify the extent to which this can lead to biased analysis results.

In Section 2, we introduce the notation and statistical model. In Section 3, we derive an upper bound on the type I error rate for a trial using a random allocation strategy. The case of blocked randomization is considered in Section 4.

The results are applied to a case study in Section 5, and the impact of our investigation on the conduct of blinded interim analyses in clinical trials is discussed in Section 6.

2. The model

Consider a parallel group comparison of an experimental treatment to a control with n subjects in total. Denote the primary endpoint measurement of subject $i = 1, \dots, n$ by X_i, and let $G_i \in \{0, 1\}$ denote the random treatment allocation. We assume that outcomes are normally distributed with means μ_{G_i} and common variance σ^2. A hypothesis test of

$$H_0 : \mu_1 \leqslant \mu_0 \text{ against } H_1 : \mu_1 > \mu_0$$

is performed at level α. After $n_1 = n/2$ observations, an interim analysis is performed, and the sample size is reassessed. The new second-stage sample size is denoted by n_2, and the new total sample size is $N := n_1 + n_2$. Besides the primary endpoint X_i, we assume that the experimenter also observes a secondary endpoint Y_i for each subject i. Assume that the response of patient i, conditional on G_i, is distributed as

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \mid G_i = g_i \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ g_i\nu_1 + (1 - g_i)\nu_0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) ,$$

where ν_1, ν_0 are the means of the secondary endpoint in treatment and control groups, respectively. At the end of the trial, H_0 will be rejected if $Z_N > z_{1-\alpha}$ where, assuming balanced group sizes,

$$rCl\, Z_N = \frac{1}{\sigma\sqrt{N}} \sum_{i=1}^{N} \left\{ \mathbf{1}(G_i = 1) - \mathbf{1}(G_i = 0) \right\} X_i = \frac{1}{\sigma\sqrt{N}} \sum_{i=1}^{N} (2G_i - 1) X_i.$$

The maximum conditional type I error rate given the first stage data from both endpoints is therefore

$$\max_{n_2 \in (0,\infty)} P\left\{ Z_N > z_{1-\alpha} \mid (X_i, Y_i)_{i=1}^{n_1} \right\} . \tag{1}$$

3. Random allocation

Our aim is to quantify the extent of potential type I error rate inflation when the outcomes of a secondary endpoint provide partial information about the treatment assignment. We first wish to quantify this inflation for a “random allocation” strategy, where exactly $n_1/2$ patients received the experimental treatment, with each of the

$$\frac{n_1!}{(n_1/2)!(n_1/2)!} \tag{2}$$

possible sequences of $\boldsymbol{G} = (G_1, G_2, \dots, G_{n_1})$ equally likely. After n_1 responses have been observed, a blinded interim analysis is performed, and a second stage sample size n_2 is chosen. We further assume random allocation in the second stage such that exactly $n_2/2$ patients receive the experimental treatment. The maximum conditional type I error rate given the first stage data from both endpoints is given by (1). Exact evaluation of (1) is difficult for all, but the smallest n_1 because the number of possible assignments (2) grows exponentially. We approach this problem in two ways. Firstly, we use an MCMC algorithm to simulate from the conditional distribution of $\boldsymbol{G}$ given the blinded interim data. Secondly, we derive asymptotic results for a “simple randomization” strategy – that is, assuming each patient is allocated to the experimental treatment independently with probability $1/2$ – and use a combination of heuristic arguments and simulation results to show that the same asymptotic results are applicable to random allocation.

3.1. Computational approach

The type I error rate conditional on the unblinded data $(X_i, Y_i, G_i)_{i=1}^{n_1}$ is easy to compute:

$$\begin{aligned} P\left\{Z_N > z_{1-\alpha} \mid (X_i, Y_i, G_i)_{i=1}^{n_1}\right\} &= P\left\{\frac{1}{\sigma\sqrt{n_2}}\sum_{i=n_1+1}^{N}(2G_i-1)X_i > \sqrt{\frac{N}{n_2}}z_{1-\alpha} - \sqrt{\frac{n_1}{n_2}}Z_1\right\} \\ &= 1-\Phi\left(\sqrt{\frac{N}{n_2}}z_{1-\alpha} - \sqrt{\frac{n_1}{n_2}}Z_1\right), \end{aligned} \tag{3}$$

where $Z_1 = \sum_{i=1}^{n_1}(2G_i-1)X_i/(\sigma\sqrt{n_1})$. It is therefore straightforward to find

$$rCI\,P\left\{Z_N > z_{1-\alpha} \mid (X_i, Y_i)_{i=1}^{n_1}\right\} = \int P\left\{Z_N > z_{1-\alpha} \mid (X_i, Y_i, G_i)_{i=1}^{n_1}\right\} \mathrm{d}P\left\{(G)_{i=1}^{n_1} \mid (X_i, Y_i)_{i=1}^{n_1}\right\} \tag{4}$$

provided that we can integrate over the conditional distribution of $\boldsymbol{G}$ given the blinded data. Although this distribution is over a large space of possible permutations, it can be sampled from using standard MCMC techniques [19]. To maximize this conditional type I error rate, we select the N that maximizes (4). R code is provided in the Supporting Information.

3.2. Asymptotic considerations and an upper bound for the type I error rate

While the aforementioned computational approach can tell us the maximum conditional type I error rate given a specific blinded data set $(x_i, y_i)_{i=1}^{n_1}$, it cannot tell us the overall properties of the sample size reassessment procedure without considerable computational effort. Therefore, we study the asymptotic conditional distribution of Z_1. We first derive the conditional distribution of Z_1 under simple randomization (instead of random allocation with fixed per group sample sizes) and then, based on heuristic arguments and supported by simulation, we argue that the same asymptotic distribution applies also for random allocation. If each patient is allocated to the experimental treatment independently with probability $1/2$ then by Bayes' theorem

$$\begin{aligned} P\left\{G_j = 1 \mid (X_i, Y_i)_{i=1}^{n_1} = (x_i, y_i)_{i=1}^{n_1}\right\} &= P\left\{G_j = 1 \mid (X_j, Y_j) = (x_j, y_j)\right\} \\ &= \frac{\varphi_{\nu_1,\sigma,\rho}(x_j, y_j)}{\varphi_{\nu_0,\sigma,\rho}(x_j, y_j) + \varphi_{\nu_1,\sigma,\rho}(x_j, y_j)} =: q_j \end{aligned} \tag{5}$$

for $j = 1, \ldots, n_1$, where $\varphi_{\nu_1,\sigma,\rho}(\cdot,\cdot)$ and $\varphi_{\nu_0,\sigma,\rho}(\cdot,\cdot)$ denote the density functions of the two dimensional normal distribution of (X_j, Y_j) under experimental treatment and control, respectively. By the central limit theorem for the sum of independent but non-identically distributed random variables (e.g., Theorem 2.7.1 in [20]),

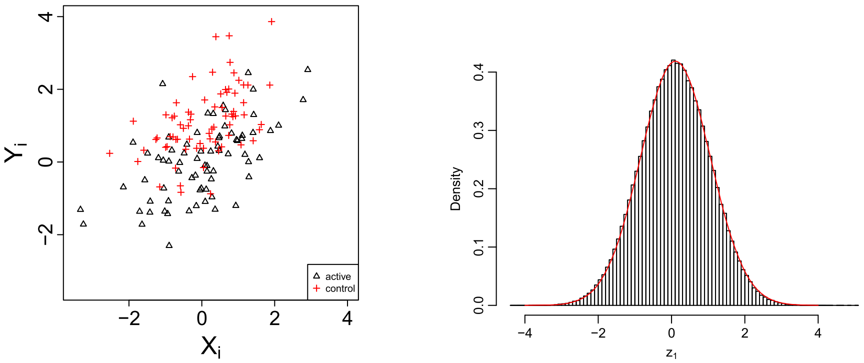
$$Z_1 \mid (X_i, Y_i)_{i=1}^{n_1} = (x_i, y_i)_{i=1}^{n_1}$$

is asymptotically normal with mean $m_1 = \sum_{i=1}^{n_1}(2q_i-1)x_i/(\sigma\sqrt{n_1})$ and variance $V_1 = 4(\sigma^2 n_1)^{-1}\sum_{i=1}^{n_1} x_i^2 q_i(1-q_i)$. We argue that this approximation is valid also under random allocation as, for n_1 large enough, the information provided by the known allocation ratio becomes negligible. In particular,

$$E\left\{G_j \mid (X_i, Y_i)_{i=1}^{n_1} = (x_i, y_i)_{i=1}^{n_1}\right\} \approx q_j, \quad \operatorname{cov}\left\{G_j, G_k \mid (X_i, Y_i)_{i=1}^{n_1} = (x_i, y_i)_{i=1}^{n_1}\right\} \approx 0 \text{ for } j \neq k.$$

To add support to our claim, we simulated multiple data sets under various choices of n_1, ν_1 and ρ, and compared the normal approximation with the output of the MCMC algorithm. An example is shown
F1 in Figure 1, where the normal curve agrees well. As ν_1 and ρ increase, it becomes easier to identify the likely treatment assignment, making the conditional distribution of Z_1 more discrete and the speed of convergence to a normal distribution slower. This can be seen in Figure 9.1–9.3 of the supporting information. For a sample size of $n_1 = 144$, the approximation appears to be adequate provided that $\nu_1 \leqslant 2$ and $\rho \leqslant 0.8$.

Equation (5) is also useful to illustrate the impact of the secondary endpoint effect size $\nu_1 - \nu_0$ on the potential to unblind the data: If $\nu_0 = \nu_1$, then $q_j = \frac{1}{2}$ and the secondary endpoint gives no information on the treatment allocation. If, in contrast, $|\nu_1 - \nu_0|$ increases then $q_j(X, Y)$ converges in distribution either to



(a) Simulated data set with $n_1 = 144$, $\nu_0 = 0$, $\nu_1 = 1$, $\sigma = 1$ and $\rho = 0.5$

(b) Histogram of draws from density of $Z_1 \mid (X_i, Y_i)_{i=1}^{n_1}$ based on 100,000 simulation runs

Figure 1. Comparing the asymptotic results with MCMC output for an example data set.

0 (for observations in the control) or to 1 (for observations from the experimental treatment group) even if the correlation ρ is zero: Indeed, for $\rho = 0$ and if Y is drawn, for example, from the control group then for all $\epsilon > 0$, we have $P(|Y - \nu_0| > c) \leqslant \epsilon$ for c large enough. However, for y, such that $|y - \nu_0| \leqslant c$, we have $q(x, y) = \varphi_{\nu_1,\sigma}(y)/\left[\varphi_{\nu_0,\sigma}(y) + \varphi_{\nu_1,\sigma}(y)\right] = 1/\{1+\exp\left[(\nu_1 - \nu_0)(\nu_1 + \nu_0 - 2y)/2\right]\} \to 0$ as $|\nu_1 - \nu_0| \to \infty$.

To maximize the overall conditional error rate, note that for any given blinded first-stage data set the maximum conditional type I error rate is

$$\begin{aligned} &\max_{n_2\in(0,\infty)} P\left\{Z_N > z_{1-\alpha} \mid (X_i, Y_i)_{i=1}^{n_1} = (x_i, y_i)_{i=1}^{n_1}\right\} \\ &= \max_{n_2\in(0,\infty)} P\left\{\frac{1}{\sigma\sqrt{N}}\sum_{i=n_1+1}^{N}(2G_i - 1)X_i + \sqrt{\frac{n_1}{N}}Z_1 > z_{1-\alpha} \mid (X_i, Y_i)_{i=1}^{n_1} = (x_i, y_i)_{i=1}^{n_1}\right\} \\ &\approx \max_{n_2\in(0,\infty)} 1 - \Phi\left(\frac{z_{1-\alpha} - \sqrt{\frac{n_1}{N}}m_1}{\sqrt{\frac{n_1V_1+n_2}{N}}}\right). \end{aligned} \tag{6}$$

Here, we approximated the conditional distribution of Z_1 by a $N(m_1, V_1)$ distribution. Assume there are minimum and maximum sample sizes $n_2^{\min}, n_2^{\max}$ for the second stage sample size such that $n_2 \in [n_2^{\min}, n_2^{\max}]$. Then, the value of n_2 maximizing (6) is (Appendix A)

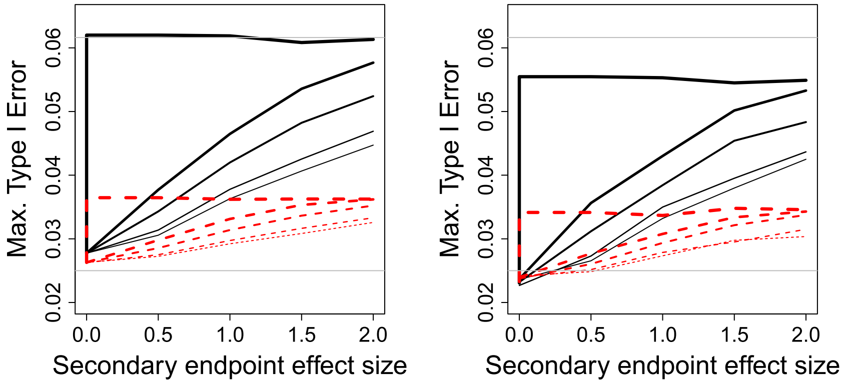
$$\tilde{n}_2(m_1, V_1) = \begin{cases} n_2^{\max} & \text{if } m_1 < z_* \\ \left[\left(\frac{z_{1-\alpha}(1-V_1)}{m_1}\right)^2 - 1\right] n_1 & \text{if } m_1 \in \left[z_*, z^*\right] \\ n_2^{\min} & \text{if } m_1 > z^* \end{cases} \tag{7}$$

if $V_1 \leqslant 1$, and

$$\tilde{n}_2(m_1, V_1) = \begin{cases} n_2^{\min} & \text{if } m_1 > z^* \\ \left[\left(\frac{z_{1-\alpha}(1-V_1)}{m_1}\right)^2 - 1\right] n_1 & \text{if } m_1 \leqslant z^* \end{cases} \tag{8}$$

if $V_1 > 1$, where $z_* = \frac{z_{1-\alpha}(1-V_1)}{\sqrt{1+n_2^{\max}/n_1}}$, $z^* = \frac{z_{1-\alpha}(1-V_1)}{\sqrt{1+n_2^{\min}/n_1}}$.

Figure 2 shows the maximum type I error rate as function of the secondary endpoint effect size for different correlations between the primary and the secondary endpoint $\rho \in \{0, 0.5, 0.8, 0.9, 1\}$. The worst case conditional error rate was determined by simulation (200,000 simulation runs if not indicated otherwise) setting $\sigma = 1$, a nominal one-sided significance level of 2.5%, and $n_1 = 144$ (which is half the total sample size required for a z-test with power 80% to detect an absolute treatment effect of 1/3 in the primary endpoint). We consider effect sizes in the secondary endpoint ranging from 0 to 2. On first sight, the latter may appear large for trials with the chosen sample size; however, effects in secondary or



(a) Z-test

(b) T-test

Figure 2. Maximum type I error rate as a function of the secondary endpoint effect size with and without restrictions for the second stage sample size. Here, the first stage sample size $n_1 = 144$ and $\sigma = 1$. (Black) solid lines denote unrestricted results, and (red) dashed lines results for restricted case with $n_2^{min} = n_1/2$ and $n_2^{max} = 4n_1$. $\rho \in \{0, 0.5, 0.8, 0.9, 1\}$ and the larger the ρ the thicker the line.

safety endpoints (such as, for example, laboratory parameters) are typically not relevant for the power calculation and may substantially differ from the treatment effect in the primary endpoint the study is powered for.

In each simulation run, primary and secondary endpoint data were simulated from a bivariate normal distribution, and the maximum conditional error rate was computed based on the approximation (6) with the second stage sample size as defined by (7) and (8). The final maximum type I error rate was then calculated based on the total Z-test statistics Z_N for the unrestricted case with $n_2^{\min} = 0$ and $n_2^{\max} = \infty$ and for the restricted case with $n_2^{\min} = n_1/2$ and $n_2^{\max} = 4n_1$ Figure 2(a). For both cases, the maximum type I error rate increases with the correlation ρ between the primary and the secondary endpoint. If this correlation is $\rho = 1$, and providing that a secondary endpoint effect is present, the maximum type I error rate under unrestricted case is $\alpha_{max} = 0.062$, which equals the maximum type I error rate inflation for an unblinded analysis reported by [2].

Careful examination of Figure 2(a) reveals that the type I error rate is already inflated when the secondary endpoint effect size is zero. This is an artifact of assuming the variance σ^2 to be known. In this case, $m_1 = 0$ and V_1 is proportional to $\sum x_i^2$. Rules (7) and (8) reduce to choosing n_2 as small as possible when $V_1 > 1$, that is, when there is excess variation in the blinded data, and as large as possible when $V_1 < 1$. Intuitively, this is because a larger variance increases the chance of obtaining a significant result. In an attempt to remove this artifact, we re-ran the simulations, but this time performing a t-test at the final analysis Figure 2(b). In this case, the procedure becomes slightly conservative at a secondary endpoint effect size of zero. Again, this makes sense for a reassessment procedure that tends to produce a high variance estimate (since this term appears in the denominator of the t statistic).

4. Block randomization

Often, randomization is performed in blocks to guarantee that the treatment allocation frequencies in the earlier and later phases of a trial are balanced (e.g., Miller *et al.* (2009) [21]). Consider a trial with block randomization with blocks of length τ, where τ is even and the treatment allocation is balanced ($P(G_i = 0) = P(G_i = 1) = 1/2$). In this section, we investigate the extent to which the additional information on the treatment allocation provided by the blocking allows one to introduce additional bias by sample size reassessment.

For example, for a block size of $\tau = 2$, for each block, there are only two possible allocation sequences, AB and BA. Both have probability 1/2. Obviously, the conditional probability, given the blinded data, that the first subject has been assigned to group A is equal to the conditional probability of the allocation sequence AB, conditional on the data of the primary and secondary endpoints of both subjects in the block. As we now use data from two patients to estimate probability of the allocation sequence AB (which is the allocation probability of the first subject) and there are only two possible sequences, we obtain a more informative estimate than in the random allocation scenario, where the allocation probability of each patient was estimated based on its own data only. However, the additional information on allocation probabilities provided by the consideration of the allocation sequences decreases with the block size. For a block size of four, for example, there are $\binom{4}{2} = 6$ possible allocation sequences: $AABB, ABAB, ABBA, BABA, BBAA, BAAB$. Each has (unconditional) probability $1/6$. To compute the conditional probability that the first patient is in group A, given the blinded data of the four patients in the block, we need to sum the conditional allocation probabilities of the first three allocation sequences. While for block size two, we used data from two patients to estimate the probabilities of two possible allocation sequences; for a block size of four, we used the data of four patients to estimate the probabilities of six possible allocations. In general, for block length τ, there are $K = \binom{\tau}{\tau/2}$ possible allocation sequences, each with unconditional probability $1/K$, and we need to estimate K allocation probabilities based on the blinded data of τ patients. Because $K >> \tau$ for larger τ, it is intuitively clear that for larger block length the additional information provided by blocking decreases (see also [21]).

To compute the worst case sample size reassessment rule in case of blocked randomization, we need to introduce some notation. Let $\mathcal{T} = \{1, 1+\tau, 1+2\tau, \ldots, n-\tau+1\}$ denote the set of indices where a new block starts. For $i \in \mathcal{T}$ let $\mathbf{b}_i = (x_j, y_j)_{j=i}^{i+\tau-1}$, denote the observations in the block starting with the i^{th} patient. Let $\boldsymbol{\omega}_k^{(i)} = (\omega_{k,j}^{(i)})_{j=1}^{\tau}$, $k = 1\ldots, K$ denote the indicator vectors of the K possible treatment allocations for block $\mathbf{b}_i, i \in \mathcal{T}$, where $\omega_{k,j}^{(i)} \in \{0,1\}$ and $\sum_{j=1}^{\tau}\omega_{k,j}^{(i)} = \tau/2$ for all $i \in \mathcal{T}$. Here, $\omega_{k,j}^{(i)} = 0$ denotes that in the k^{th} treatment allocation for i^{th} block the j^{th} patient in the block was allocated to group A (control), and $\omega_{k,j}^{(i)} = 1$ denotes that this patient was allocated to group B (treatment). Under block randomization, each allocation is equally likely, such that $P\left(\boldsymbol{\omega}_k^{(i)}\right) = 1/K$ for $k = 1, 2, \ldots, K$ and $i \in \mathcal{T}$ and the joint density for the observations b_i in block i is given by

$$f(b_i) = \frac{1}{K}\sum_{k=1}^{K} f\left(b_i|\boldsymbol{\omega}_k^{(i)}\right) ,$$

where $f(b_i|\boldsymbol{\omega}_k^{(i)}) = \prod_{l=0}^{\tau-1} f(x_{i+l}, y_{i+l}|g_{i+l} = \omega_{k,l+1}^{(i)})$, and $f(x_{i+l}, y_{i+l}|g_{i+l} = \omega_{k,l+1}^{(i)})$ denotes a bivariate normal density with mean vector $(\nu_0, \omega_{k,l+1}^{(i)}\nu_1 + (1-\omega_{k,l+1}^{(i)})\nu_0)$, variances σ^2, and correlation ρ. Then, the conditional probability of each treatment allocation, given the data of block b_i, is given by

$$P\left(\boldsymbol{\omega}_k^{(i)}|b_i\right) = \frac{f\left(b_i|\boldsymbol{\omega}_k^{(i)}\right)\cdot P(\boldsymbol{\omega}_k^{(i)})}{f(b_i)} = \frac{f\left(b_i|\boldsymbol{\omega}_k^{(i)}\right)}{\sum_{k=1}^{K} f\left(b_i|\boldsymbol{\omega}_k^{(i)}\right)} , \quad k = 1, 2, \ldots, K.$$

To derive the sample size reassessment rule that maximizes the type I error rate, we compute the conditional expectation and conditional variance of the first stage test statistics Z_1, conditional on the blinded first stage observations $(X_i, Y_i)_{i=1}^{n_1}$:

$$m_{Z_1} = E\left(Z_1|(X_i, Y_i)_{i=1}^{n_1} = (x_i, y_i)_{i=1}^{n_1}\right) = \frac{1}{\sigma\sqrt{n_1}}\sum_{i\in\mathcal{T}} E\left(m_{\tau,i}|b_i\right) = \frac{\sum_{i\in\mathcal{T}}\sum_{k=1}^{K} P\left(\boldsymbol{\omega}_k^{(i)}|b_i\right) m_{\tau,i}^{(k)}}{\sqrt{n_1}} ,$$

$$v_{Z_1} = Var\left(Z_1|(X_i, Y_i)_{i=1}^{n_1} = (x_i, y_i)_{i=1}^{n_1}\right) = \frac{1}{\sigma^2 n_1}\sum_{i\in\mathcal{T}}\left(\sum_{k=1}^{K} P\left(\boldsymbol{\omega}_k^{(i)}|b_i\right)\left(m_{\tau,i}^{(k)}\right)^2 - \left(\sum_{k=1}^{K} P\left(\boldsymbol{\omega}_k^{(i)}|b_i\right) m_{\tau,i}^{(k)}\right)^2\right) ,$$

where $m_{\tau,i} = \sum_{l=0}^{\tau-1}(2G_{i+l} - 1)X_{i+l}$, and $m_{\tau,i}^{(k)}$ is a realization of $m_{\tau,i}$ at k^{th} treatment allocation of i^{th} block at $(X_i, Y_i) = (x_i, y_i)$. As in the random allocation case, the conditional distribution of Z_2 given the blinded first stage observations $(X_i, Y_i)_{i=1}^{n_1}$ is standard normal, and we approximate the conditional distribution of Z_1 by a normal distribution with mean m_{Z_1} and variance v_{Z_1}. As in the unblocked case, we can express the overall test statistic Z_N as a weighted sum of the stage wise test statistics such that the conditional error rate is given by

$$P_{H_0}\left(Z_N > z_{1-\alpha}|(X_i, Y_i)_{i=1}^{n_1} = \left(x_i, y_i\right)_{i=1}^{n_1}\right) = 1 - \Phi\left(\frac{z_{1-\alpha} - \sqrt{\frac{n_1}{N}}m_{Z_1}}{\sqrt{\frac{n_1}{N}v_{Z_1} + \frac{n_2}{N}}}\right) . \qquad (9)$$

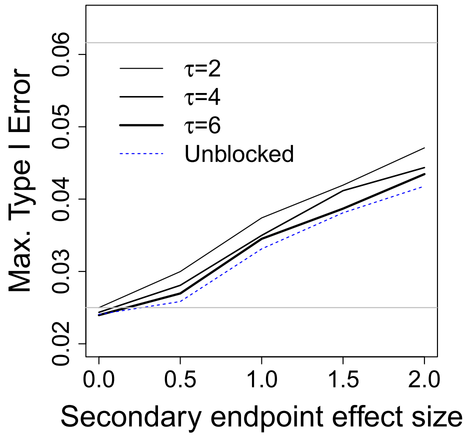


Figure 3. Maximum type I error rate without restrictions for the second stage sample size and for blocked randomization with block sizes 2, 4, 6 and the unblocked design ($n_1 = 144$, $\rho = 0$, $\sigma = 1$, and $2.5 \cdot 10^5$ ($2 \cdot 10^5$) simulation runs for block size 2 (4, 6, unblocked design)).

If there are restrictions for the second stage sample size that is $n_2 \in [n_2^{\min}, n_2^{\max}]$ for some $n_2^{\min}$ and $n_2^{\max}$, then (Appendix A) the value of n_2 maximizing (9) can be calculated as in the unblocked case by (7) or (8) with m_1 and V_1 replaced by m_{Z_1} and v_{Z_1}, respectively.

Figure 3 shows the maximum type I error rate of the trial with block randomization, with block sizes $\{2, 4, 6\}$ (and per group sample size 72) for $\rho = 0$ and unrestricted second stage sample size (i.e., $n_2 \in [0, +\infty)$). Results for other correlations for both unrestricted and restricted second stage sample size are given in the Supporting Information Figure 9.4. As expected, using the additional information on the blocking of observations increases the maximum type I error rate. The smaller the block size the better the data can be unblinded, and the larger is the maximal type I error rate.

To implement the aforementioned worst case sample size adaptation rule, one must know the block size. However, also if the block sizes are not known, the type I error rate may be inflated. Consider a clinical trial where block randomization is used, but the worst case sample size reassessment rule for random allocation (7, 8) is applied (which does not require knowledge of the block sizes). To estimate the type I error rate for such a setting, simulation studies for different block sizes were performed as in the preceding text. In all considered scenarios, the simulated maximum type I error rate is very close to the maximum error rate observed in the setting of Section 3, where the same sample size reassessment rule for random allocation (7, 8) is applied, but random allocation is used to allocate patients (Figure 9.7–9.10 in the Supporting Information).

5. A clinical trial example

As an illustrative example, consider a Phase III clinical trial to asses efficacy and safety of Fingolimod in patients with relapsing-remitting multiple sclerosis along the lines of the FREEDOMS trial [22, 23]. While in the original trial, 1,272 patients were randomized to receive oral Fingolimod doses of 0.5 or 1.25 mg or placebo daily; for simplicity, we consider a trial with two parallel groups, comparing only the higher dose with placebo with $N = 800$ patients in total, randomly allocated to groups (such that the per-group sample size is similar to the original trial). The annualized aggregate relapse rate (ARR) during months 0 to end of study was set as a primary endpoint and is defined as the number of confirmed relapses in a year. Among the additionally measured parameters is the mean lymphocyte count. It is known from earlier studies that Fingolimod lowers the lymphocyte count compared wtih placebo. Based on the results in [24], we assume a mean $\nu_0 = 1.8$ [$\times 10^9$ cells/L] for the placebo group and $\nu_1 = 0.55$[$\times 10^9$ cells/L] for the 1.25 mg Fingolimod group with a common standard deviation of $\sigma_x = \sigma_y = 0.31$[$\times 10^9$ cells/L] for the mean lymphocyte counts at day 7 (values approximated based on Figure 6 in [24]). Furthermore, it is known that the lymphocyte counts are causally related to the primary endpoint through the mechanism of action of Fingolimod. Lee JY *et al.* (2013)[25] investigated the relation between the predicted lymphocyte count to ARR. Because no correlation coefficient is given in [25], and we do not have access to the raw data; we computed the maximum inflation of the type I error rate and the correlation between the blinded and unblinded effect size estimates for a grid of ρ in $[0, 0.9]$.

Assume that in the Phase III trial an interim analysis is performed after $n_1 = 400$ subjects have been randomized. By (6), the maximum type I error rate for $\rho \in [0, 0.9]$ ranges from $\alpha_{max} = 0.054$ to $\alpha_{max} = 0.059$ and if we restrict the second stage sample size such that $\tilde{n}_2 \in [200, 1600]$ the maximum type I error rate ranges from $\alpha_{max} = 0.035$ to $\alpha_{max} = 0.036$ (see Supporting Information Figure 9.6 (a)).

As another example, assume that instead of the lymphocyte counts the mean total white blood cell counts (WBC) are used. The mean WBC count for the placebo group at 24 months is $\nu_0 = 6.5 \times 10^9$ cells/L with a standard deviation of 1.8, the mean for Fingolimod 1.25 mg group $\nu_1 = 3.8 \times 10^9$ cells/L with a standard deviation of 1.3 (estimated from the Supporting Information Figure 1 C in [26]). As we are not aware of published data on the correlation of WBC and ARR, we computed the maximum inflation of the type I error rate and the correlation between the blinded and unblinded effect size estimates for a grid of ρ in $[0, 0.9]$. Then, pooling the group wise standard deviations to $\sigma = 1.57$, the upper bound for the type I error rate for $\rho \in [0, 0.9]$ ranges from $\alpha_{max} = 0.041$ to $\alpha_{max} = 0.054$ for unrestricted second stage sample size and from $\alpha_{max} = 0.031$ to $\alpha_{max} = 0.035$ if the second stage sample size is restricted to the interval [200, 1600] (see Supporting Information Figure 9.6 (b)).

6. Discussion

In this work, we demonstrated that even blinded sample size reassessment may lead to an inflation of the type I error rate if there are secondary endpoints for which the alternative holds. This implies that unscheduled sample size reassessment, even in a blinded setting, may damage the integrity of the trial. The numerical results give an upper bound for the inflation of the type I error that may occur due to blinded sample size reassessment in a setting where the distribution of a secondary endpoint is known to the experimenter, for example from historical data. While this is a simplifying assumption, the impact of a treatment on surrogate endpoints is often known from Phase II trials before a Phase III trial is started.

However, the approach can be extended to settings where no prior information on the distribution of the secondary endpoint is available. In this case, the distribution of the secondary endpoint can be estimated from the blinded data based on a mixture model with an expectation–maximization (EM) algorithm as in [27] or [28]. It has been shown that such estimators, when applied to the data of the primary endpoint, are only reliable for very large effect or sample sizes [29] and perform poorly for effect sizes usually occurring in clinical trials. However, while large treatment effects in the primary endpoint do occur rarely, this does not necessarily apply to effect sizes for secondary or safety endpoints (see, for example, the clinical trial example in Section 5), which are relevant for the setting considered in this manuscript. Overall, depending on the effect size in the secondary endpoint, the type I error rate resulting from sample size reassessment based on expectation–maximization algorithms will still be affected, albeit on a lower scale.

In the computation of the worst case sample size reassessment rule, we used only the information from a single secondary endpoint to estimate the treatment allocation. Instead, one could use the data from several endpoints: to derive the resulting maximum type I error rate, one needs to replace the bivariate normal densities in (5) by the respective multivariate densities. To extend the setting of a single interim analysis to multiple blinded interim analyses, one can derive worst case adaptation rules and the resulting maximum type I error rate with a backwards induction approach.

We investigated the impact of different randomization procedures on the maximum type I error rate and found that block randomization, especially with small block sizes, increases the type I error rate inflation, if the information on the block size is used in the sample size adjustment. If the latter information is not used, blocking leads to essentially the same inflation as under random allocation. These findings support current recommendations against too small block sizes and inclusion of information on block sizes in study protocols [30].

Wang *et al.* [31] consider a related problem and derive the maximum type I error rate for sample size reassessment rules based on *unblinded* interim effect size estimates of a secondary endpoint that is correlated with the primary endpoint, but assuming that the primary endpoint is not observed in the interim analysis. The maximum type I error rate in this setting depends only on the correlation ρ of the primary and the secondary endpoints, and there is no inflation of the type I error rate if $\rho = 0$. In contrast, in the blinded setting considered in this paper, even if the correlation between the primary and the secondary endpoint is zero, the type I error rate may be inflated. This holds because we assume that the primary endpoint is observed and the blinding is partially lost due to a treatment effect in the secondary endpoint that gives information on the treatment allocation. The potential inflation of the type I rate is related to the fact that this partial loss of blinding allows one to estimate the unblinded first stage effect size estimate in the primary endpoint $\bar{X} = [\sum_{i=1}^{n_1} 2(2G_i - 1)X_i]/n_1$: if the unknown G_i are replaced by q_i as defined in (5), a blinded estimate is given by $\bar{X}_b = [\sum_{i=1}^{n_1} 2(2q_i - 1)X_i]/n_1$. The correlation r between $\bar{X}$ and $\bar{X}_b$ (not to be confused with the correlation ρ between primary and secondary endpoint) can be interpreted as a measure of unblinding and increases with the effect size in the secondary endpoint and ρ. In the clinical trial of Section 5, for example, r ranges from 0.97 up to nearly 1 in the first and from 0.68 to 0.96 in the second example for $\rho \in [0, 0.9]$ (see Figure 9.5 in the Supporting Information Figures and Section 8.1 in the Supporting Information for computational details).

Our findings do not contradict the well established use of blinded sample size reassessment based on aggregate event rates or variance estimates computed from blinded primary endpoint interim data. However, they demonstrate that the type I error rate control of these methods relies on the application of specific, binding, pre-planned, and fully algorithmic sample size reassessment rules (as recommended for data monitoring committee charters, see for example [32]) for which type I error control has been demonstrated. The type I error rate control does not extend to general sample size adjustments based on blinded data. Therefore, including only a non-binding option for blinded sample size reassessment in clinical trial protocols is not sufficient to guarantee type I error rate control. In particular, we quantify the maximum type I error rate inflation when a worst case adaptation rule is applied that also uses information from a secondary endpoint.

Our work also implies that post hoc adjustments of the sample size may lead to type I error rate inflations, even if justified by post hoc scientific arguments (as required in the guideline quoted in the Introduction). Consider, for example, a scenario where blinded outcome data is available and adaptations following the rule in Section 3.3 are applied whenever a post hoc selected sample size reassessment rule (or scientific arguments external to the trial) can be found that justifies that choice. Otherwise, the pre-specified sample size is used. Because the conditional error rate is increased in all instances where the sample size is adapted but is unchanged otherwise, the overall type I error rate will be inflated by such a strategy. Furthermore, note that even aggregate statistics (as referred to in the quoted guidelines) may contain information on the unblinded treatment effect estimate and therefore may lead to type I error rate inflation. Examples are the correlation coefficient of the primary endpoint and a secondary or safety endpoint (if there is a treatment effect in the latter), or per group means of subgroups whose definition is based on such secondary or safety endpoints. While the assumption that a worst case sample size rule is applied in an actual clinical trial may not be realistic, it is a means to derive an upper bound for the type I error rate in settings where no binding sample size reassessment procedure is pre-specified, or post hoc adaptations are performed, and secondary endpoint data has been available. While the actual type I error may be substantially lower than this upper bound, it can not be computed because it depends not only on the realized sample sizes but also on the sample sizes that would have been applied had other interim data been observed.

In settings where no sample size adjustment algorithm has been pre-specified, alternatives to fixed sample hypothesis tests are tests based on combination functions or the conditional error rate principle [2, 4, 33, 34] that control the type I error rate even without pre-specified adaptation rules. The conditional error rate based procedures even control the type I error rate if no adaptations were pre-planned but are introduced during the conduct of the study.

Appendix A: Computation of the sample size reassessment rule maximizing the type I error rate.

To maximize the type I error rate in the second stage sample size $\tilde{n}_2$, one needs to maximize the corresponding conditional error rate, which is (6) for the random allocation case and (9) for the blocked case. For computational convenience, let us express both conditional error rates as functions of $R = \frac{\tilde{n}_2}{n_1}$. Then, in both considered cases, the conditional error rate can be expressed as $1 - \Phi(f(R))$, where

$$f(R) = \frac{z_{1-\alpha}\sqrt{1+R} - m}{\sqrt{v+R}}$$

with $m = m_1$ and $v = V_1$ for (6) and with $m = m_{Z_1}$ and $v = v_{Z_1}$ for (9). Finding the $\tilde{n}_2$ that maximizes the conditional error rate is then equivalent to determining the R that minimizes f. The latter can be found by taking the derivative of f in R. We consider the general case where the second stage sample size may be restricted that is $\tilde{n}_2 \in [n_2^{\min}, n_2^{\max}]$ for some $n_2^{\min} \geq 0$ and $n_2^{\max} \leqslant \infty$. This translates to boundaries

R_{min}, R_{max} for R where $R_{min} = n_2^{\min}/n_1$ and $R_{max} = n_2^{\max}/n_1$. In the unrestricted case, we set $n_2^{\min} = 0$ and $n_2^{\max} = \infty$ such that $R_{min} = 0$ and $R_{max} = \infty$. The first derivative of f is given by

$$\frac{\partial f(R)}{\partial R} = \frac{z_{1-\alpha}(v-1) + m\sqrt{1+R}}{2\sqrt{1+R}(v+R)^{3/2}} .$$

Assume first that $v < 1$. To determine extrema of f, we consider the following cases:

(1) $m \leqslant 0$. Then, $\frac{\partial f(R)}{\partial R} < 0$ and $f(R)$ is minimized at $\tilde{R} = R_{max}$;

(2) $m \in \left(0, \frac{z_{1-\alpha}(1-v)}{\sqrt{1+R_{max}}}\right)$. Then,

$$\frac{\partial f(R)}{\partial R} < \frac{z_{1-\alpha}(1-v)\left[\sqrt{\frac{1+R}{1+R_{max}}} - 1\right]}{2\sqrt{1+R}(v+R)^{3/2}} ,$$

and because $\sqrt{\frac{1+R}{1+R_{max}}} \leqslant 1$, $\frac{\partial f(R)}{\partial R} < 0$ for $R \in [R_{min}, R_{max}]$ and $f(R)$ is minimized at $\tilde{R} = R_{max}$;

(3) $m \in \left[\frac{z_{1-\alpha}(1-v)}{\sqrt{1+R_{max}}}, \frac{z_{1-\alpha}(1-v)}{\sqrt{1+R_{min}}}\right]$.

Then, $\frac{\partial f(R)}{\partial R} = 0$ for

$$\tilde{R} = \left(\frac{z_{1-\alpha}(1-v)}{m}\right)^2 - 1 ;$$

$\tilde{R}$ is indeed a local minimum because the second derivative of f

$$\frac{z_{1-\alpha}(1-v)(3+v+4R) - 3m(1+R)^{3/2}}{4(1+R)^{3/2}(v+R)^{5/2}} ,$$

evaluated at $\tilde{R}$ is equal to

$$-\frac{m^6(v-1)}{4z_{1-\alpha}^2\left((v-1)^3\left(m^2+(v-1)z_{1-\alpha}^2\right)\right)^{3/2}} > 0 .$$

Furthermore,

$$f(\tilde{R}) = \sqrt{\frac{z_{1-\alpha}^2(v-1) + m^2}{v-1}} .$$

(4) $m > \frac{z_{1-\alpha}(1-v)}{\sqrt{1+R_{min}}}$. Then,

$$\frac{\partial f(R)}{\partial R} > \frac{z_{1-\alpha}(1-v)\left[\sqrt{\frac{1+R}{1+R_{min}}} - 1\right]}{2\sqrt{1+R}(v+R)^{3/2}} ,$$

and because $\sqrt{\frac{1+R}{1+R_{min}}} > 1$, $\frac{\partial f(R)}{\partial R} > 0$ for $R \in [R_{min}, R_{max}]$ and $f(R)$ is minimized at $\tilde{R} = R_{min}$.

Taking all the four cases together, we obtain for $v < 1$ that the $\tilde{R}$ minimizing $f(R)$ is

$$\tilde{R}(m,v) = \begin{cases} R_{max} & \text{if } m < z_* \\ \left[\left(\frac{z_{1-\alpha}(1-v)}{m}\right)^2 - 1\right] & \text{if } m \in \left[z_*, z^*\right] \\ R_{min} & \text{if } m > z^* \end{cases} , \tag{A.1}$$

where $z_* = \frac{z_{1-\alpha}(1-v)}{\sqrt{1+n_2^{\max}/n_1}}$, $z^* = \frac{z_{1-\alpha}(1-v)}{\sqrt{1+n_2^{\min}/n_1}}$.

Now, taking $\tilde{n}_2(m, v) = n_1\tilde{R}$, we obtain the sample size reassessment rule (for $v < 1$)

$$\tilde{n}_2(m, v) = \begin{cases} n_2^{\max} & \text{if } m < z_* \\ \left[\left(\frac{z_{1-\alpha}(1-v)}{m}\right)^2 - 1\right] n_1 & \text{if } m \in \left[z_*, z^*\right] \\ n_2^{\min} & \text{if } m > z^* \end{cases}, \tag{A.2}$$

If now $v > 1$, then consider the following cases:

- $m \geq 0$: then, $\frac{\partial f(R)}{\partial R} > 0$ and $f(R)$ is minimized at $\tilde{R} = R_{min}$,
- $m \in (z^*, z_*)$: then, $\frac{\partial f(R)}{\partial R} > 0$ and $f(R)$ is minimized at $\tilde{R} = R_{min}$,
- $m \leqslant z^*$: then, $f(R)$ is minimized at $\tilde{R} = \left(\frac{z_{1-\alpha}(1-v)}{m}\right)^2 - 1$,
- $m > z_*$: then, $\frac{\partial f(R)}{\partial R} > 0$ and $f(R)$ is minimized at $\tilde{R} = R_{min}$.

Summarizing for $v > 1$ the $\tilde{R}$ minimizing $f(R)$ is

$$\tilde{R}(m, v) = \begin{cases} R_{min} & \text{if } m > z^* \\ \left[\left(\frac{z_{1-\alpha}(1-v)}{m}\right)^2 - 1\right] & \text{if } m \leqslant z^* \end{cases}, \tag{A.3}$$

Taking $\tilde{n}_2(m, v) = n_1\tilde{R}$, we obtain the sample size reassessment rule (for $v > 1$)

$$\tilde{n}_2(m, v) = \begin{cases} n_2^{\min} & \text{if } m > z^* \\ \left[\left(\frac{z_{1-\alpha}(1-v)}{m}\right)^2 - 1\right] n_1 & \text{if } m \leqslant z^* \end{cases}. \tag{A.4}$$

Acknowledgements

This study was supported by the Austrian Science Fund (FWF):P23167.

References

1. *ICH Topic E9: Notes for guidance on statistical principles for clinical trials*, European Agency for the Evaluation of Medical Products: London, UK, 1998.
2. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics* 1995; **51**:1315–1324.
3. Graf AC, Bauer P. Maximum inflation of the type 1 error rate when sample size and allocation rate are adapted in a pre-planned interim look. *Statistics in Medicine* 2011; **30**:1637–1647.
4. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 1994; **50**:1029–1041.
5. Posch M, Bauer P, Brannath W. Issues in designing flexible trials. *Statistics in Medicine* 2003; **23**:953–969.
6. Bretz F, König F, Brannath W, Glimm E, Posch M. Adaptive designs for confirmatory clinical trials. *Statistics in Medicine* 2003; **28**:1181–1217.
7. EMEA. Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. *EMEA Doc. Ref. CHMP/EWP/2459/02*, 2007. Available at http://www.emea.europa.eu/pdfs/human/ewp/245902enadopted.pdf.
8. *Guidance for industry: adaptive design clinical trials for drugs and biologics (Draft guidance)*, CDER, FDA: Rockville, MD, USA, 2010.
9. *Guidance for industry: adaptive designs for medical device clinical studies (Draft guidance)*, CBER, FDA: Rockville, MD, USA, 2015.
10. Kieser M, Friede T. Simple procedures for blinded sample size adjustment that do not affect the type I error rate. *Statistics in Medicine* 2003; **22**:3571–3581.
11. Proschan M. Two-stage sample size re-estimation based on a nuisance parameter: a review. *Journal of Biopharmaceutical Statistics* 2005; **15**:559–574.
12. Proschan MA, Lan KKG, Wittes JT. *Statistical Monitoring of Clinical Trials: A Unified Approach*. Springer: New York, 2006.
13. Friede T, Kieser M. Sample size recalculation for binary data in internal pilot study designs. *Pharmaceutical Statistics* 2004; **3**:269–279.
14. Friede T, Schmidli H. Blinded sample size reestimation with count data: methods and applications in multiple sclerosis. *Statistics in Medicine* 2010; **29**:1145–1156.
15. Posch M, Proschan MA. Unplanned adaptations before breaking the blind. *Statistics in Medicine* 2012; **31**:4146–4153.

16. Proschan M, Glimm E, Posch M. Connections between permutation and t-tests: relevance to adaptive methods. *Statistics in Medicine* 2014; **33**(27):4734–4742.
17. Friede T, Kieser M. Blinded sample size reassessment in non-inferiority and equivalence trials. *Statistics in Medicine* 2003; **22**:995–1007.
18. Elsäßer A, Regnstrom J, Vetter T, Koenig F, Hemmings RJ, Greco M, Papaluca-Amati M, Posch M. Adaptive clinical trial designs for european marketing authorization: a survey of scientific advice letters from the european medicines agency. *Trials* 2014; **15**:383.
19. Hastings WK. Monte carlo sampling methods using markov chains and their applications. *Biometrika* 1970; **57**(1):97–109.
20. Lehmann EL. *Elements of Large-Sample Theory*, Springer Texts in Statistics. Springer: New York, 1998.
21. Miller F, Friede T, Kieser M. Blinded assessment of treatment effects utilizing information about the randomization block length. *Statistics in Medicine* 2009; **28**:1690–1706.
22. A service of the U.S. National Institutes of Health. Efficacy and safety of Fingolimod in patients with relapsing-remitting multiple sclerosis (FREEDOMS). *ClinicalTrials.gov, Bethesda, MD, USA* 2006. Identifier:NCT00289978.
23. Kappos L, Radue EW, O'Connor P, Polman C, Hohlfeld R, Calabresi P, Selmaj K, Agoropoulou C, Leyk M, Zhang-Auberson L, Burtin P. A placebo-controlled trial of oral fingolimod in relapsing multiple sclerosis. *New England Journal of Medicine* 2010; **362**(5):387–401.
24. Boulton C, Meiser K, David OJ, Schmouder R. Pharmacodynamic effects of steady-state fingolimod on antibody response in healthy volunteers: a 4-week, randomized, placebo-controlled, parallel-group, multiple-dose study. *The Journal of Clinical Pharmacology* 2012; **52**:1879–1890.
25. Lee JY, Wang Y. Use of a biomarker in exposure-response analysis to support dose selection for fingolimod. *Pharmacometrics and Systems Pharmacology* 2013; **2**(8):e67.
26. Francis G, Kappos L, OConnor P, Collins W, Tang D, Mercier F, Cohen JA. Temporal profile of lymphocyte counts and relationship with infections with fingolimod therapy. *Multiple Sclerosis Journal* 2014; **20**(4):471–480.
27. Gould AL, Shih WJ. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics, Part A. Theory and Methods* 1992; **21**:2833–2853.
28. Teel C, Park T, Sampson AR. Em estimation for finite mixture models with known mixture component size. *Communications in Statistics - Simulation and Computation* 2015; **44**(6).
29. Friede T, Kieser M. On the inappropriateness of an EM algorithm based procedure for blinded sample size re-estimation. *Statistics in Medicine* 2002; **21**:165–176.
30. CPMP Q6 working party on efficacy of medicinal products note for guidance iii/3630/92-en. Biostatistical methodology in clinical trials in applications for marketing authorizations for medicinal products. *Statistics in Medicine* 1995; **14**(15): 1659–1682.
31. Wang SJ, Brannath W, Brückner M, James Hung HM, Koch A. Unblinded adaptive statistical information design based on clinical endpoint or biomarker. *Statistics in Biopharmaceutical Research* 2013; **5**:293–310.
32. Ellenberg SS, Fleming TR, DeMets DL. *Appendix A: The Data Monitoring Committee Charter*. John Wiley & Sons, Ltd: The Atrium, Southern Gate, Chichester, West Sussex PO198SQ, England, 2003,175–183.
33. Brannath W, Posch M, Bauer P. Recursive combination tests. *Journal of the American Statistical Association* 2002; **97**: 236–244.
34. Müller HH, Schäfer H. A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine* 2004; **23**:2497–2508.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.