Type I Error Inflation of Blinded Sample Size Re-Estimation in Equivalence Testing

Ekkehard Glimm, Lillian Yau, and Heike Woehling

Novartis Pharma AG, Basel, Switzerland; Sandoz Biopharmaceuticals, Hexal AG, Holzkirchen, Germany

1. Introduction

A medicine is considered biosimilar if it can be shown that its structure, function, and clinical effect matches that of an existing biologic medicine, or a reference, which has been approved by regulatory authorities (European Medicines Agency 2020; US Food & Drug Administration 2020b). In biosimilar product development, clinical studies are used to address residual uncertainties left from technical characterization, such as purity, chemical identity, and bioactivity, of a proposed biosimilar product as compared to its reference biologic. The clinical studies are aimed to demonstrate that there is no clinically meaningful difference between the proposed biosimilar and the reference product in terms of safety and effectiveness. This is usually demonstrated through “human pharmacokinetic (exposure) and pharmacodynamic (response) studies, an assessment of clinical immunogenicity, and, if needed, additional clinical studies” (US Food & Drug Administration 2020a).

Such biosimilar pharmacokinetic (PK) or pharmacodynamic (PD) studies are usually double-blinded and randomized with three treatment arms: the proposed biosimilar arm, US reference arm, and EU reference. The inclusion of both the US and the EU marketed reference products is due to the two regulatory bodies’ separate requirements. Equivalence must be established between the biosimilar product and the two reference products independently. To our knowledge, U.S. Food & Drug Administration (FDA) and European Medicine Agency (EMA) are the only two health authorities that require fully powered pivotal PK studies testing the product marketed in their respective countries or regions. Should this situation change, producers of biosimilar products will need to modify their approaches and strategies accordingly.

Biosimilar PK/PD studies are pivotal confirmatory studies that have to be adequately powered (European Medicines Agency 2020; US Food & Drug Administration 2020b). At the time of planning of these studies, information regarding the endpoints, for example, the variability, are often limited in the study population. Therefore, to avoid either under-powering or over-recruiting, an interim sample size re-estimation (SSR) during the study conduct can be very helpful.

Discussions on SSR have appeared in statistical literature as early as the 1940s (Stein 1945). Since then a growing body of literature has emerged, both on blinded (Wittes and Brittain 1990; Bristol and Shurzinske 2001; Friede and Kieser 2001; Golkowski, Friede, and Kieser 2014) and unblinded (Cui, Hung, and Wang 1999; Lehmacher and Wassmer 1999) SSR, and much of it in the context of adaptive designs (Bretz et al. 2009). The two types of SSR have also been referred to as noncomparative and comparative, respectively, by the FDA (US Food & Drug Administration 2019). For the two-sample t-test of superiority of a new treatment versus a standard control, methods based on estimators of the sample variance which can be calculated without unblinding the group affiliation of the observations available at an interim analysis have been introduced (Kieser and Friede 2003; Friede and Kieser 2006). The simplest of these is the “total variance estimate” which is the usual one-sample variance estimate calculated from treating all observations as if they were from only one treatment group. Although it has been shown that under this approach the Type I error is not strictly controlled (Proschan, Glimm, and Posch 2014; Lu 2016), in the case of superiority testing, the Type I error violations are extremely small and even then occur only in small samples. From a practical perspective, these investigations therefore
reinforce the recommendations made by Kieser and Friede (2003) and Friede and Kieser (2006). However, in the case of noninferiority (NI) testing, and thus also in equivalence (EQ) testing as used in biosimilar development, this inflation can be larger (Friede and Kieser 2003, 2011, 2013; Lu 2016).

In this article, we take a closer look at the nontrivial Type I error inflation in EQ (and NI testing when relevant), and the factors that impact it. In the community of bioequivalence trialists, there seems generally little awareness of this fact since the often cited references concern superiority trials. We also discuss practical implications for the planning of a bioequivalence study with a blinded SSR.

We base our discussion on a design with two stages and two treatment groups. Stage 1 consists of subjects whose data are used for estimating the total variance; stage 2 consists of subjects who are additionally recruited after the interim. In certain cases, no additional subjects are needed and stage 2 does not have any subjects. An important assumption is that beside the total sample size, other elements of the study design do not change after the interim.

In Section 2, we derive a decomposition of the total variance observed at the time of blinded SSR to demonstrate how Type I error inflation arises in NI and EQ tests. Section 3 provides a brief introduction to the two one-sided tests (TOST) approach to EQ testing. Section 4 outlines the simulation settings. Section 5 discusses simulation results and how different factors impact the Type I error. An example of a recently planned biosimilar PK study which implemented a blinded SSR is given in Section 6. In Section 7, we summarize our findings.

2. Type I Error Violation in Noninferiority Testing

A common strategy for EQ testing is the TOST approach (Schuirmann 1987). This uses two NI tests. For this reason, we start by looking at the latter.

Assume that we want to test $H_0 : \delta \geq \delta_0$ versus $H_A : \delta < \delta_0$ at level $\alpha$ on independent normal observations $y_{ij}$ from $N(\mu_j, \sigma^2)$ in two groups $j \in 1, 2$ and $i = 1, \ldots, n_j$. Let $\delta = \mu_1 - \mu_2$ be the true population difference, and $\bar{y}_j$ be the mean of the observations in group $j$. Without loss of generality, assume furthermore that $\delta_0 > 0$ is a prespecified NI margin. We proceed with the “total variance”

$$\hat{\sigma}_T^2 = \frac{Q}{n_s - 1} = \frac{1}{n_s - 1} \sum_{j=1}^{2} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2,$$

where $n_s = n_1 + n_2$ with $n_j$ being the stage 1 sample size in group $j$, and $\bar{y} = \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2}{n_s}$ is the mean across the two groups.

To calculate $E(\hat{\sigma}_T^2)$, we decompose $Q$ into $Q_1$ and $Q_2$, where

$$Q_1 = \sum_{j=1}^{2} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2,$$

$$Q_2 = \sum_{j=1}^{2} n_j (\bar{y}_j - \bar{y})^2.$$

It follows that $Q = Q_1 + Q_2$. By Cochran’s theorem, $Q_1$ and $Q_2$ are stochastically independent with distributions $Q_1 \sim \sigma^2 \chi^2(n_s - 2)$ and $Q_2 \sim \sigma^2 \chi^2(1; \frac{n_1 n_2 \delta^2}{n_s \sigma^2})$, which is a noncentral $\chi^2$ distribution. Hence,

$$E(Q) = E(Q_1) + E(Q_2) = (n_s - 2)\sigma^2 + \left(1 + \frac{n_1 n_2 \delta^2}{n_s (n_s - 1) \sigma^2}\right)\sigma^2,$$

and

$$E(\hat{\sigma}_T^2) = \frac{1}{n_s - 1} E(Q) = \sigma^2 \left(1 + \frac{n_1 n_2 \delta^2}{n_s (n_s - 1) \sigma^2}\right). \quad (2)$$

Note that $E(\hat{\sigma}_T^2) = \sigma^2$ if $\delta = 0$ such that under the null hypothesis of a superiority test, the total variance estimate $\hat{\sigma}_T^2$ is an unbiased estimate of the variance.

To illustrate how Type I error inflation may arise, consider the following SSR rule:

1. If $\hat{\sigma}_T^2$ is large, that is, $\hat{\sigma}_T^2 > c$ for some preselected threshold $c > 0$, we recruit $m_j, j \in 1, 2$, additional subjects.
2. If $\hat{\sigma}_T^2$ is small, that is, $\hat{\sigma}_T^2 \leq c$, we do not recruit additional subjects, but stop and test right away with the $n = n_1 + n_2$ subjects we have.

If the true $\delta > 0$, we can see from Equation (2) that a small value of $\hat{\sigma}_T^2$ provides evidence against $H_0$, because $E(\hat{\sigma}_T^2)$ does not only depend on the variance, but also on $\delta^2$, the squared true mean difference between the two groups. Hence, the rule above stops recruiting and uses the data obtained up to the interim analysis undiluted by additional observations if it happened to be in favor of $H_A$, but dilutes this evidence by recruiting more subjects if the data obtained up to the interim analysis shows evidence in favor of $H_0$. It is clear that this must inflate the Type I error of the entire strategy. The same is true for other blinded SSR rules where the final total sample size is an increasing function of $\hat{\sigma}_T^2$. On the other hand, if the SSR rule were a decreasing function of $\hat{\sigma}_T^2$, then we would have $\alpha$ deflation.

Two remarks are important. First, from Equation (2), we can see that the inflation is due to the fact that we test a shifted hypothesis $H_0 : \delta \geq \delta_0 > 0$. If $\delta_0 = 0$, then at the point of $\delta = \delta_0$, $\hat{\sigma}_T^2$ is an unbiased estimate of $\sigma^2$. Hence, the issue investigated here does not occur with superiority testing where $H_0 : \delta \geq \delta_0 = 0$. Intuitively, a large $\hat{\sigma}_T^2$ does not provide evidence in favor of, or against, the $H_0$ of the superiority test, because a large observed $\hat{\sigma}_T^2$ can equally likely arise from $y_1 - y_2$ or $y_2 - y_1$ being large. With the shifted $\delta = \delta_0 > 0$, however, a large value of $\hat{\sigma}_T^2$ will much more likely occur with a value of $y_1 - y_2 > \delta$ than with $y_1 - y_2 < -\delta$. We note in passing that simply subtracting a known constant, that is, a hypothetical true difference, from the total variance cannot solve this problem.

Second, the statements we have made here are qualitative: There will be a Type I error inflation, but its magnitude depends on many things: the standardized threshold $\delta_0/\sigma$, the stage 1 sample size, the range of possible values of the stage 2 sample size, the actual small sample properties of a test using a test statistic which is only asymptotically Normal, and the like. Equivalence testing with TOST is a case in point. If the equivalence margin is symmetric (i.e., $\delta_0 = \delta_{up} = -\delta_{low}$), then exactly the same mechanisms are at work at both the upper and the lower end of the equivalence range, and both one-sided level-$\alpha$ tests are subject to Type I error inflation with the sample
size review strategy described above. The actual probability of rejection in the TOST, however, is also subject to the inherent conservatism of the TOST strategy which arises from the fact that there is a positive probability of not rejecting $H_{01}: \delta \leq -\delta_0$ when in fact $H_{02}: \delta \geq \delta_0$ is true. This conservatism of course remains and may counteract against the liberality caused by the phenomenon we discussed.

An analytical calculation of the exact Type I error rate is often algebraically consuming. In Appendix A, we discuss the quantification of the Type I error inflation in a specific case. For general discussion, we use simulation studies.

### 3. Equivalence Testing and TOST

In the TOST approach to equivalence testing, two hypotheses are tested simultaneously:

$$H_{01}: \delta \leq \delta_{low} \text{ vs. } H_{A1}: \delta > \delta_{low},$$

and

$$H_{02}: \delta \geq \delta_{up} \text{ vs. } H_{A2}: \delta < \delta_{up},$$

where $\delta_{low}$ and $\delta_{up}$ are the lower and the upper *equivalence margins*. To claim equivalence, both null hypotheses must be rejected simultaneously at level $\alpha$. Using the same notation as in Section 2, and with the given *reference range* ($\delta_{low}$, $\delta_{up}$), we reject the null hypothesis of nonequivalence $H_0 = H_{01} \cup H_{02}$ if

$$\frac{\sqrt{n_1n_2}}{n_1 + n_2} \cdot \frac{d - \delta_{low}}{s} > t_{1-\alpha}(n_1 + n_2 - 2),$$

and

$$\frac{\sqrt{n_1n_2}}{n_1 + n_2} \cdot \frac{d - \delta_{up}}{s} < t_{\alpha}(n_1 + n_2 - 2),$$

where $d = \bar{y}_1 - \bar{y}_2$ and

$$s^2 = \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (y_{ij} - \bar{y}_i)^2.$$

This decision rule can also be described via a $(1 - 2\alpha)$ level confidence interval (CI) for $\delta$. If this CI is completely contained within $(\delta_{low}, \delta_{up})$, then equivalence between the two groups is declared at level $\alpha$.

We distinguish four possible cases with regard to the rejection of the two null hypotheses, and illustrate them in terms of the CI for $\delta$ in Figure 1. Of these cases, Case 1 is where equivalence is established.

Consider an EQ test using TOST where $H_{02}$ is true. Case 1 represents a false rejection and establishes equivalence incorrectly, that is, committing a Type I error. Although Case 2 is the least desirable outcome, it makes TOST conservative: it falsely rejects $H_{02}$, but also falsely does not reject $H_{01}$, thus comes to the correct conclusion that equivalence cannot be established, even though the reasoning is wrong. Case 3 is the most desired outcome since it correctly supports nonrejection of $H_{02}$ but rejection of $H_{01}$. Case 4 does not reject either null hypothesis, therefore, reaches the correct decision of not rejecting $H_{02}$.

In comparison, for NI testing where $H_{02}$ is true, Case 1 and Case 2 together represent false rejection of the null hypothesis, and both contribute to Type I error.

### 4. Simulation Studies

We investigate the impact of stage 1 sample size ($n_j$), stage 2 sample size ($m_j$), and standardized true difference $\delta/\sigma$ on Type I error inflation in a blinded SSR for EQ testing using simulation studies. Without loss of generality, we set $\sigma$ to 1, and assume symmetric equivalence margins where $\delta_{up} = \delta_0 = -\delta_{low}$; hence, in subsequent discussions $\delta_0$ and its standardized version $\delta_0/\sigma$ are numerically the same.

The sample sizes for the two groups are assumed the same at every stage, so the subscript $j$ is dropped in the discussion, and $\tilde{n}$ and $m$ are the per-group-sample size of stage 1 and 2, respectively. We follow this strategy: before the start of the study, define $\tilde{n}$, $n_{min}$, and $n_{max}$ where $\tilde{n} \leq n_{min} \leq \tilde{n} + m \leq n_{max}$. A blinded sample size reassessment is conducted when data from $\tilde{n}$ subjects per group are available.

#### 4.1. The SSR Rule

At the interim, the total sample size needed in each group, denoted $\tilde{N}$, to reject $H_{01}$ and $H_{02}$ for a desired power $1 - \beta$ is calculated based on $\tilde{\delta}_0^2$. See Section 4.3 for details of how this is implemented for this simulation. Subsequently, $\tilde{N}$ is compared to $n_{min}$ and $n_{max}$ such that

1. if $\tilde{N} \leq n_{min}$, then additional $m = n_{min} - \tilde{n}$ subjects are recruited per group in the second stage;
2. if $\tilde{N} > n_{min}$, then additional $m = \text{min}(\tilde{N}, n_{max}) - \tilde{n}$ subjects are recruited in the second stage.

Hypotheses $H_{01}$ and $H_{02}$ are tested with $n = \tilde{n} + m$ total subjects per group at the end of the study.

#### 4.2. Simulation Settings

All of the simulations use $\alpha = 5\%$ and $\beta = 10\%$. Table 1 summarizes the other conditions:

| Outcome   | $H_{01}: \delta \leq \delta_{low}$ | $H_{02}: \delta \geq \delta_{up}$ |
|-----------|------------------------------------|------------------------------------|
| Case 1    | Reject                             | Reject                             |
| Case 2    | Not reject                         | Reject                             |
| Case 3    | Reject                             | Not reject                         |
| Case 4    | Not reject                         | Not reject                         |

Figure 1. Four possible outcomes of a TOST.
Table 1. Simulation settings.

| Parameter | Values |
|-----------|--------|
| \( \hat{n} \) | 10, 15, 20, 25, 30, 40, 50, 60, 80 |
| \( n_{\text{min}}/\hat{n} \) | 1, 1.2, 1.4, 1.6, 1.8, 2 |
| \( n_{\text{max}}/\hat{n} \) | 2, 2.5, 3, 3.5, 4, \( \infty \) |
| \( \delta_{\alpha}/\sigma = \delta_0 \) | 0.05 to 1.5 in steps of 0.05 |

Take \( \hat{n} = 20, n_{\min}/\hat{n} = 1.4 \) (or \( n_{\min} = 20 \times 1.4 = 28 \)), and \( n_{\max}/\hat{n} = 2 \) (or \( n_{\max} = 20 \times 2 = 40 \)) as an example. The interim analysis is to be conducted when data from 20 subjects per group become available. Based on the total variance estimate \( \hat{\sigma}^2 \), if the total sample size \( \hat{N} \) is determined to be no more than 28 subjects per group, then add additional \( m = 8 \) subjects per group will be recruited in the second stage, making it total of \( n = n_{\min} = 28 \). If the reassessed sample size \( \hat{N} \) turned out to be larger than 28, then \( m = \min(\hat{N}, 40) - 20 \) subjects per group are to be recruited for the second stage.

There are total 9,720 combinations of the parameter values in Table 1. One million simulations are performed for each combination. Means and variances of normally distributed data per stage are generated. The algorithm used is given in Appendix C.

There are two horizontal dashed lines on each panel representing the 95% CI around the nominal 5% when one million simulations are performed. Any values within these two dashed lines can be attributed to random variation. In Panel A, percentages of Case 1 (i.e., false rejections of the equivalence hypothesis) are plotted against \( \delta = \delta_0 \). This corresponds to a true \( H_{02} \). In each simulated sample, an \( t \)-test is conducted between the two groups with \( n = \hat{n} + m \) subjects in each group. The percentages of each of the four cases from Figure 1 are summarized.

4.3. Sample Size Calculation in the Simulation

For each simulated sample, the total sample size \( \hat{N} \) is calculated from the stage 1 data by assuming a normal population variance \( \sigma^2 \) that is estimated by \( \hat{\sigma}^2 \). Ideally \( \hat{N} \) should be calculated using the noncentral \( t \)-distribution as outlined in Schuirmann (1987). However, this would require iterative approximation of degrees of freedom for each single simulation. Furthermore, we focus on Type I error inflation which ultimately depends on the sample size formula only via \( \hat{N} \). For this purpose, it is essentially irrelevant if the \( t \)- or the simpler \( z \)-formula is used. In both cases, Type I error inflations follow identical patterns. Therefore, we use the usual sample size formula based on the asymptotic normal distribution:

\[
\hat{N} = \frac{2(Z_{1-\beta}/2 + Z_{1-\alpha})^2}{((\delta_0 - D)/\hat{\sigma})^2},
\]

assuming 1:1 randomization ratio where \( Z_\gamma \) denotes the \( \gamma \)-quantile from the standard Normal distribution. Since we are blinded at interim analysis, we do not have an estimate for \( \delta \); a value \( D \) must be assumed. Without loss of generality, we assume \( D = 0 \). If there are reasons to believe \( D > 0 \), this will only increase the required sample size, and will not change the pattern of Type I error inflation. However, this is not to say that \( H_{02} : \delta \geq \delta_0 \) is correctly not rejected. Panel B of Figure 2 shows the percent of Case 2 against \( \delta \). Recall a Case 2 is where the lower end of the CI of the difference is below the lower margin. The probability of a false rejection of \( H_{02} \) starts out at 5% and remains close to 5% until starting to drop sharply at around \( \delta = 0.3 \). If \( H_{02} \) is falsely rejected, this is mostly due to Case 2 rather than Case 1 if \( \delta < 0.3 \).

For any fixed \( \sigma \), for example, \( \sigma = 1 \), as \( \delta = \delta_0 \) becomes larger the total variance \( \hat{\sigma}^2 \) becomes larger, which leads to a larger total final sample size \( n \) per group after the interim reassessment. As a result, the CI for the difference becomes narrower given the same SSR rule. This in turn shows in Panel A such that the percentage of Case 1 is increasing sharply when \( \delta \) is between 0.30 and 0.50; at the same time, percentage of Case 2 starts to decrease, and reaching 0% at \( \delta = 0.55 \) just when percent Case 1 reaches 5%. If we look at Panel C, we see the two cases add up to about 5% or exceeding that where there is inflation. In fact, Panel C (or Panel D for a close-up) are the simulated Type I errors in a NI test against \( H_{02} \) with the given SSR rule. For both NI and EQ testing, the \( \alpha \) inflation happens when \( \delta \) is between 0.70 and 1.25 with a peak at 0.95.

5. Results of the Simulations

5.1. An Illustrative Example of \( \alpha \) Inflation

We first illustrate the mechanisms leading to \( \alpha \) inflation by focusing on one of the simulation settings. For this purpose, we assume that \( \delta = \delta_0 \) and investigate what happens if this boundary of \( H_{02} \) is moved from small to large values when \( H_{02} \) is true. Figure 2 shows the observed Type I error rate for NI and EQ testing for the setting where \( \hat{n} = 15, n_{\min} = 18, \) and \( n_{\max} = 30 \).

In Figure 2, there are two horizontal dashed lines on each panel representing the 95% CI around the nominal 5% when one million simulations are performed. Any values within these two dashed lines can be attributed to random variation. In Panel A, percentages of Case 1 (i.e., false rejections of the equivalence hypothesis) are plotted against \( \delta = \delta_0 \). When \( \delta \) is at or below 0.30, the standard deviation \( \sigma \), which is set to be 1 for all the simulations, is much too large for \( \delta \). Therefore, even with the largest sample size allowed according to our SSR rule, nearly no CI for the difference between the two groups are narrow enough to be within (\( -\delta_0, \delta_0 \)). Hence, no Type I error is committed under the EQ hypothesis.

The patterns in Figure 2 are consistent in all of the settings we examined. The difference is in where the peak \( \alpha \) occurs, and how much it inflates to. Figure 3 summarizes the percent of Case 1 in form of heatmaps where a darker shade indicates a larger \( \alpha \) value. In addition, we can see from these plots how \( \hat{n}, n_{\min}, \) and \( n_{\max} \) influence \( \alpha \) inflation. Four stage 1 sample
Figure 2. Percent Case 1 or Case 2 when \( \tilde{n} = 15, n_{\text{min}} = 18, \) and \( n_{\text{max}} = 30. \) The two horizontal lines bracket values between 4.96% and 5.04%, where 95% rejection rates should be if the true rejection rate is 5% among one million simulations.

The graphs show that inflations of Type I errors occur at \( \delta = \delta_0 \) between 0.8 and 1.5, and that they are more severe with small stage 1 sample sizes and when \( n_{\text{min}} = \tilde{n}. \)

### 5.3. Stage 1 Sample Size \( \tilde{n} \) and \( \alpha \) Inflation

In Figure 3, we see a consistent decrease of \( \alpha \) inflation as \( \tilde{n} \) increases. This is related to the precision in the estimate of \( \text{var}(\hat{\sigma}_T^2). \) Given Equation (1), we derive

\[
\text{Var}(\hat{\sigma}_T^2) = \frac{2\sigma^2}{(2\tilde{n} - 1)^2} \left[ \tilde{n} (2 + (\delta/\sigma)^2) - 1 \right] = O \left( \frac{(\delta/\sigma)^2 \tilde{n}}{n} \right).
\]  

(4)

With fixed \( \delta, \text{var}(\hat{\sigma}_T^2) \) decreases as \( \tilde{n} \) increases. The sample size formula in Equation (3), which is investigated in the simulations, depends on the stage 1 sample size only via \( \hat{\sigma}_T^2. \) The more variation we see in \( \hat{\sigma}_T^2 \) the more variable second stage sample sizes will be. In turn this variability in stage 2 sample sizes induces \( \alpha \) inflation.

Figure 4 is a heatmap of the standard deviations of the total realized sample sizes \( n = \tilde{n} + m \) for the selected simulation settings that are shown in Figure 3, but without the last row where there is no upper limit of the second stage sample size. The plot clearly indicates a close correspondence between high inflation and large standard deviation of \( n = \tilde{n} + m. \) A corresponding plot of average total sample size per setting is available in the supplementary materials.

In addition, to have a closer look, we investigate how the percentage of rejections depends on \( \hat{\sigma}_T^2 \) for an exemplary case with \( \delta_0 = 1 \) between a fixed design and 2-stage design with SSR. Results (in Figure 5) are from \( 10^5 \) simulations assuming \( H_{02} : \delta = \delta_0 \) is true. First, the data from a total of 30 subjects, 15 per group, are randomly generated. After that, for the fixed design, ordinary \( t \)-tests are conducted from these 30 subjects. For the 2-stage design with SSR, the total variance \( \hat{\sigma}_T^2 \) is calculated from the 30 subjects, and a stage 2 sample size \( m, \) where \( 0 \leq m < \infty, \) is determined according to Equation (3) with \( \alpha = 5\% \) and
\( \beta = 10\% \). Data from the two stages are then combined to conduct the \( t \)-test across both stages.

Figure 5 shows that the percent of rejections of \( H_{02} \) is related to the values of \( \hat{\sigma}^2_T \) for both the fixed design (black bars) and the 2-stage design (gray bars). The height of the bars shows the percentage of rejections which occur in intervals of observed \( \hat{\sigma}^2_T \) values. The first (from the left) of these intervals contains all values of \( \hat{\sigma}^2_T \) which are so small that \( m = 0 \). In this example, this requires that \( \hat{\sigma}^2_T \leq 0.693 \), and includes 2.29% of all simulations. When this happens, the fixed and the 2-stage design are identical. 46% of the 2.29% where \( \hat{\sigma}^2_T \leq 0.693 \) (i.e., \( m = 0 \)) are binned into 10 equally sized intervals, that is, each interval contains 9771 simulation results.

The plot supports the results from Section 2 which show that small values of \( \sigma^2_T \) are providing information against \( H_{02} \). As \( \sigma^2_T \) gets larger, the percent rejections by the fixed design are decreasing toward 0: the weighted average across all intervals is of course \( \alpha = 5\% \). For the 2-stage design, the rejection rate is the same as the fixed design when \( \hat{\sigma}^2_T \leq 0.693 \) (because \( m = 0 \) there), but as \( \hat{\sigma}^2_T \) increases, they diverge from the fixed design. While the second (from the left) gray bar is lower than its black counterpart, from the 4th pair onward the 2-stage rejection rates are higher. This reflects the fact that additional stage 2 data helps to overcome the dampening effect of large \( \sigma^2_T \) on rejection probability. Overall, the weighted average of the gray bars becomes 5.83% demonstrating the Type I error inflation.

The graph also shows that while general tendencies are predictable—such that rejection rates must decrease; the two sets of rates diverge as we move to the right; the gray bars on average are higher than the black bars especially on the right of the plot, etc.—some peculiarities of the specific sample size formula in Equation (3) also play a role. In this case, for example, the largest total sample size requested by any of the simulation runs is 74. This shows that, although we have not put any “formal” upper limit on stage 2 sample size, the formula is such that stage 2 sample sizes which completely dominate the stage 1 data simply do not occur. If the sample size formula were such that huge stage 2 sample sizes would occur frequently enough, the gray bars would approach a height of 5% as we move to the right in the plot. The fact that the 2-stage design has a slightly
Figure 4. Standard deviations of total realized sample size \( n = \tilde{n} + m \) where \( \tilde{n} = 10, 15, 20, 30 \). \( s1ss \): Stage 1 sample size (\( \tilde{n} \)); \( n.\text{Min} \): minimal total sample size (\( n_{\text{min}} \)); \( n.\text{Max} \): maximum total sample size (\( n_{\text{max}} \)).

Figure 5. Percent false rejections in intervals of total variance \( \hat{\sigma}^2_T \) with \( \delta_0/\sigma = 1 \).

lower rejection rate in the second lowest interval of \( \sigma^2_T \) is also a peculiarity of this specific SSR rule.

In the case when stage 1 sample size \( \tilde{n} \) is large, the observed range for \( \hat{\sigma}^2_T \) is much smaller. The behavior of the observed \( \alpha \) values resembles a fixed design more closely. The inflation of \( \alpha \) is less severe.

As this reasoning shows, the variability of \( \hat{\sigma}^2_T \) in the simulations is a good indicator of the SSR method’s ability to react to different stage 1 data situations in favor of rejection of \( H_0 \). In contrast, when \( \hat{\sigma}^2_T \) is very precise, the total (and the second stage) sample size is then almost always the same from simulation to simulation, and we are led to always need the same overall sample size. This resembles a fixed stage design where the total sample size is predetermined and a mid-way look of the data does not change any element of the study design, hence, having no impact on \( \alpha \).

Therefore, one obvious remedy in reducing \( \alpha \) inflation is to increase the stage 1 sample size. However, in practice stage 1 sample size has to be reasonably small so that operationally there is enough time for an interim analysis to be carried out. Alternatively, we could reduce the variability of the stage 2 sample size by imposing an \( n_{\text{min}} \) or an \( n_{\text{max}} \) on the overall sample size. Of course, this is against the original motivation for the sample size review whose intent is to be able to modify the final actual sample size as freely as possible.
5.4. EQ Margin $\delta_0$ and $\alpha$ Inflation

From Equations (2) and (3), with $\sigma = 1$, $D = 0$, and under $H_{02}: \delta = \delta_0$, we derive the expected value of the total sample size for each group

$$E(\hat{N}) = 2(Z_{1-\beta/2} + Z_{1-\alpha})^2 \left( \frac{1}{\delta_0} + \frac{\bar{n}}{4n - 2} \right),$$

and its standard deviation

$$SD(\hat{N}) = 2(Z_{1-\beta/2} + Z_{1-\alpha})^2 \times \frac{SD(\hat{\gamma}^2)}{\delta_0^2} = O \left( \frac{1}{\delta_0} \cdot \frac{1}{\sqrt{n}} \right).$$

Although both $E(\hat{N})$ and $SD(\hat{N})$ approach infinity as $\delta_0$ goes to 0, the former approaches at a higher rate since

$$\lim_{\delta_0 \to 0} \frac{E(\hat{N})}{SD(\hat{N})} = \lim_{\delta_0 \to 0} \frac{1}{\delta_0} = \infty.$$

For small $\delta_0$, stage 2 sample size will always overwhelm stage 1 sample size if there is no restriction on the overall sample size. In our simulations, when $\delta_0 = 0.05$ the average total sample size required to reject the null hypothesis with 90% power at 5% nominal $\alpha$ level is close to 9,000 per group. Faced with such a large average $\hat{N}$, the information carried in stage 1 data are eclipsed. This again resembles a study with a single stage.

Figure 6 shows the EQ margin (x-axis) versus percent of Case 1, or Type I error (y-axis) for an EQ test for all the $\bar{n}$ we investigate, allowing $0 \leq m < \infty$. At the small end of $\delta_0$ regardless of the size of $\bar{n}$, $\alpha$ is around 5% for the reason just stated.

However, in reality the upper limit of the total sample size cannot be left un-capped. When $n_{\text{min}}$ and $n_{\text{max}}$ are predefined, since $\hat{N}$ is so large, $n_{\text{min}}$ does not play any role, and $n_{\text{max}}$ is always reached. What is more, since $n_{\text{max}} \ll \infty$, there is almost no power to reject $H_{01}$. This implies Case 2 is the predominating case where $H_{02}$ is falsely rejected. For example, Figure 7 gives results for $\bar{n} = 15$ and $n_{\text{min}}$ fixed at 30 whereas $n_{\text{max}}$ is infinity, 60, 38, or 30. As $n_{\text{max}}$ decreases from 60 to 38 and then 30, the power for rejecting the false $H_{01}$ decreases, and Case 1’s become Case 2’s. Visually, the values for $\delta$ where Case 1 rises to the range of 5% shifts from 0.45, to 0.60 and 0.70, respectively.

As the EQ margin $\delta_0$ moves away from 0, $\hat{N} < n_{\text{max}}$ more and more often, and the variability in stage 2 sample size increases. Therefore, $\alpha$ starts to become more than its nominal level for the same reason as given in Section 5.3.

At the other end of the spectrum where $\delta_0$ is large, Equation (5) gives

$$\lim_{\delta_0 \to \infty} \hat{N} = \frac{\bar{n}}{2\bar{n} - 1} (Z_{1-\beta/2} + Z_{1-\alpha})^2,$$

and

$$\lim_{\delta_0 \to \infty} SD(\hat{N}) = 0.$$

With $\beta = 0.10$ and $\alpha = 0.05$, when $\bar{n} = 1$, $\lim_{\delta_0 \to \infty} \hat{N} = 11$; when $\bar{n}$ gets larger it converges to 5.5. In any practical setting where $\bar{n}$ is at least 6, the stage 1 sample is already sufficient for the desired power. Therefore, any $n_{\text{min}} \geq \bar{n} \geq 6$ will have $n_{\text{min}}$ as the final total sample size, and $n_{\text{max}}$ will not play a role. This is why in Figure 6 where $0 \leq m < \infty$, we can see a convergence toward 5% for all $\bar{n}$ at the larger end of $\delta_0$.

Before $\delta_0$ becomes so large that $\hat{N}$ settles toward 5.5, $SD(\hat{N})$ is larger than 0, which can result in $\hat{N}$ that are above $n_{\text{min}}$ but not equal to or exceeding $n_{\text{max}}$. To reduce the variation in $\hat{N}$ and $\alpha$ inflation, we can set $n_{\text{min}}$ just high enough to include all foreseeable cases where this lower limit of $\hat{N}$ suffices to reject $H_0$ with desired power at the nominal $\alpha$ level. From Figure 8, we can see that even with a moderate increase in the mandatory minimum total sample size, the $\alpha$ inflation reduces sharply.

5.5. The Peak $\alpha$ Inflation

As discussed in Section 2, the precise Type I error inflation depends on many factors such as stage 1 sample size, the equivalence margin, the SSR rule which potentially includes upper or lower limits of the final sample size. It is difficult to give a general, analytical solution to the upper limit of the $\alpha$ inflation. As can be seen in all the figures, different versions of the SSR have led to different peaks both in terms of magnitude and the corresponding $\delta_0$ value even with the same stage 1 sample sizes.

Table 2 shows the standardized EQ margins $\delta_0/\sigma$ and the peak $\alpha$ for each $\bar{n}$ in our simulations where $0 \leq m < \infty$. If at the
For the NI test, the $\alpha$-level is 0.05. Although the table refers to the Type I error rate of an EQ test, it is usually calculated for the nominal $\alpha$.

Overpowered studies are larger than necessary. An example allows an approach where the reference range is scaled to the variability of the reference product (Davit et al. 2012).

Table 2: Peak $\alpha$ and its corresponding standardized margin value.

| $n$ | $\%$ Case 1 | $\delta_0/\sigma$ |
|-----|-------------|-------------------|
| 10  | 6.26        | 1.20              |
| 15  | 5.78        | 0.95              |
| 20  | 5.63        | 0.85              |
| 25  | 5.55        | 0.80              |
| 30  | 5.45        | 0.75              |
| 40  | 5.34        | 0.60              |
| 50  | 5.30        | 0.60              |
| 60  | 5.23        | 0.55              |
| 80  | 5.18        | 0.45              |

Table 2 can serve as a reference for planning future studies. Although the table refers to the Type I error rate of an EQ test, for the NI test, the $\alpha$ are negligibly larger.

### 6. An Example: A Double-Blinded, Randomized, Three-Arm Equivalence Study

In PK/PD studies, area under the drug concentration (AUC) curve is often the primary endpoint. As a random variable, AUC is regarded as having a log-normal distribution. Additionally, in pharmacology, data are conventionally summarized by geometric means and coefficient of variations (CV), where a CV, describing the variation, is the ratio between the standard deviation and the mean. Likewise, the hypotheses of the TOST are usually formulated in terms of the geometric mean ratio, that is

$$H_{01}: \frac{\tau_B}{\tau_j} \leq 0.80 \text{ vs. } H_{A1}: \frac{\tau_B}{\tau_j} > 0.80,$$

$$H_{02}: \frac{\tau_B}{\tau_j} \geq 1.25 \text{ vs. } H_{A2}: \frac{\tau_B}{\tau_j} < 1.25,$$

where $\tau_B$ and $\tau_j$ are the geometric means of AUC of the biosimilar treatment arm and Reference $j$, $j = 1, 2$, respectively, referring to U.S. and EU marketed reference products; and 0.80 and 1.25 are the lower and the upper equivalence margins, respectively, as required by many health authorities, for example, the FDA in the United States, and the EMA in the EU. Writing $\delta = \log(\tau_B/\tau_j)$, the hypotheses can be rewritten as

$$H_{01}: \delta \leq \log(0.80) = -0.223 \text{ vs. } H_{A1}: \delta > \log(0.80) = -0.223,$$

$$H_{02}: \delta \geq \log(1.25) = 0.223 \text{ vs. } H_{A2}: \delta < \log(1.25) = 0.223.$$

At the time when a PK study is being planned for a biosimilar product development, information on its reference product in the public domain is often sparse. Recently when we were preparing a PK study, although a few PK studies of the reference product existed, they were of different populations and with different objectives. The results from these studies invariably differed. The team decided to include a blinded interim sample size review into the study design. This was proposed to both FDA and EMA.

The study was 1:1:1 randomized. The blinded interim would take place when a total of 60 subjects (20 per group) was to have completed their scheduled assessments. The second stage sample size $m$ would then be chosen based on the desired power and the observed total (blinded) AUC CV from the 60 subjects as shown in Table 3. The concrete numbers $c_{v1} < c_{v2} < c_{v3}$ and $m_1 < m_2 < m_3 < m_4$ are not given here due to confidentiality.

Simulations were conducted to examine the extent of Type I error inflation, assuming the usual 5% significant level. We focused our investigations on where the true (unknown) CV is less than 40%. In Figure 9, each solid dot represents percent of rejection of the null hypothesis among 1 million simulations. The solid and dashed lines connecting the dots are for the two reference products.

As we can see, when the true CV is between 0.175 and 0.375, the false rejection rate clearly exceeds where it would expected to be when the true rate is 5%, that is, between the two horizontal
dashed lines in Figure 9. In fact, when CV is 0.21, the false rejection rate can be at 5.32%. In contrast, in superiority testing, when total interim sample size is as large as 20 per group, Type I error inflation is negligible (Prochan, Glimm, and Posch 2014). Hence, the regulatory guidelines on superiority testing consider $\alpha$ adjustment for blinded SSR unnecessary (US Food & Drug Administration 2019).

Given these findings, the team decided that the primary endpoints should be tested at a reduced 4.5% significance level instead of the usual accepted 5%. This is an approach that other authors have also adopted (Friede and Stammer 2010). The simulations were repeated, and Figure 10 presents the false rejection rates. The two horizontal lines bracket values between 4.46% and 4.54%, within which, among one million simulations, 95% rejection rates should be if the true rejection rate is 4.5%.

The strategy and the tightened significance level for the primary endpoints were proposed to FDA and EMA, and have been accepted by both health authorities.

7. Summary

We have investigated causes of Type I error inflation arising when blinded SSR is performed in equivalence trials. In Section 2, we have given an explanation for why nontrivial Type I error inflation can arise in this situation. The reasoning behind this explanation allows the construction of blinded sample size estimation rules which emphasize the inflation (one such rule is investigated in the Appendix). However, from an applied perspective, it is more important to understand how this phenomenon impacts SSR rules which are used in practice.
To investigate this, we have performed extensive simulations using SSR outlined in Section 4.1. We have set up the simulations to focus on scenarios where inflation must be suspected, but that are also practically relevant.

We find that the variation of the stage 2 sample size is a good, but not the only, indicator of (the magnitude of) Type I error inflation. The stage 2 sample size is an increasing function of the total variance estimated at the interim. Hence, it is more variable when stage 1 is small and when there are no limits on the minimum and the maximum sample size allowed in stage 2. When the standardized EQ margin (δ0/σ) is very strict, stage 2 sample size is invariably estimated to be large in general and no inflation can arise. Conversely, when it is very generous, the stage 2 sample size tends to be always small and converges to a constant number as the margins tend toward infinity. Hence, there can also be no inflation. For this and practical reasons, the focus of our attention is on cases where the equivalence margin is roughly the same as the standard deviation.

We observe that if stage 2 sample size is not limited by additional restrictions, Type I error inflations to 6.3% can arise for stage 1 sample sizes of 10 subjects per group, a small, but not entirely unrealistic number, and persist (5.2%) for as many as 80 subjects per group in stage 1.

These inflations are reduced when limits are placed on the allowed stage 2 sample sizes. In the range of practically relevant EQ margins, the lower limit (n_{\text{min}}) is much more important than the upper limit (n_{\text{max}}) here. The upper limit has an substantial impact on Type I error probability only for values of the EQ margin that are a lot smaller than the standard deviation of the data.

Our investigation shows that the more similar to a fixed design a study is, the less α inflation can arise. This points to designing a study with a narrower range of n_{\text{min}} and n_{\text{max}}. Our simulations indicate that α inflation remains 5.3% or less if one can restrict the SSR rule in the following way: (a) choose an interim sample size \( \hat{n} \geq 15 \); (b) pick \( m \geq \hat{n} \); and (c) cap \( n_{\text{max}} \leq 3\hat{n} \). We note that based on simulation studies, Birkett and Day (1994) and Sandvik et al. (1996) suggested to have at least 10 subjects per treatment group, or an overall 20 degrees of freedom, as the stage 1 sample size for a superiority test. Our recommendation of at least 15 here is in line with this suggestion considering our finding that in NI and EQ settings, to reduce α inflation, a slightly larger stage 1 sample size is required compared to when conducting a superiority test.

When an interim is deemed necessary at the planning stage of a study, this will be due to substantial uncertainties in key assumptions. Not having a lot of flexibility in choosing a second stage sample size limits the opportunities to rectify these uncertainties, hence, defeats the purpose of having an interim analysis and reduces the attractiveness of a blinded SSR.

As alternatives, one can consider unblinded SSR for equivalence testing (see, e.g., Maurer, Jones, and Chen 2018). Otherwise, one can reduce nominal significance level (e.g., from 5% to 4.5% as in our example) in such a way that the maximum Type I error inflation is confined within the actual significance level (i.e., 5%). This requires investigation of the α inflation through simulations.

### Appendix A. Numerical Evaluation of Type I Error Inflation for NI and EQ Testing

Here we aim to demonstrate two points. First, we give a numerical example of a SSR rule where an increased stage 1 sample size does not imply asymptotic decreasing of α inflation (see Appendix B for some more discussion on asymptotics). Second, through this example we illustrate the feasibility of analytical computation of the exact α once the components of an SSR rule are defined.

In this example, we assume that no additional subjects are recruited if the total variance estimate \( \hat{\sigma}_T^2 \), defined in Equation (1), is less than or equal to a threshold \( \tilde{c} \). When this happens, the study is unblinded, and a final hypothesis testing is conducted based on \( d = \hat{y}_1 - \hat{y}_2 \), the observed difference between the mean responses in group 1 and group 2 after stage 1—the only stage, of the trial.

Since \( \hat{\sigma}_T^2 = \frac{Q_2 + Q_1}{n_1} \), we can express this as a boundary \( c = (n - 1) \cdot \tilde{c} \) on \( Q_1 + Q_2 \).

For ease of notation, in this section, let stage 1 sample sizes for the two groups be \( n_1 \) and \( n_2 \) such that \( n_1 = n_2 \), and total stage 1 sample size \( n = n_1 + n_2 = 2n_1 \). Therefore, \( \sqrt{\frac{n}{2}} d \sim N(\sqrt{\frac{\hat{\sigma}^2}{n}}, \sigma^2) \), and \( Q_2 = \frac{\hat{\sigma}^2}{n} \cdot d^2 \).

Consider the test

\[ H_{02} : \delta \geq \delta_{\text{up}} \]

which will be rejected if

\[ t_{\text{up}} = \sqrt{\frac{n_1}{2}} \cdot \frac{d - \delta_{\text{up}}}{\sqrt{\frac{Q_2}{n}}} \leq t_{\alpha}(n - 2). \]

It follows that \( P(\text{rejecting } H_{02}) = P(t_{\text{up}} \leq t_{\alpha}(n - 2)) \) can be calculated in two steps: (1) the probability when \( Q_1 + Q_2 \leq c \), and (2) the probability when \( Q_1 + Q_2 > c \).

Without loss of generality, we assume that \( \alpha < 0.5 \), \( \delta = \delta_0 \), \( \sigma^2 = 1 \) and \( c > 0 \). We first calculate \( P(t_{\text{up}} \leq t_{\alpha}(n - 2) \mid Q_1 + Q_2 \leq c) \).

In general, for a given \( Q_2 = x \), we have

\[ P(t_{\text{up}} \leq t_{\alpha}(n - 2) \mid Q_1 = x) = \]

\[ P \left( \frac{n_1}{2} \cdot \frac{(d - \delta_{\text{up}})}{\sqrt{\frac{Q_2}{n}}} \leq q \cdot \sqrt{\frac{x}{n - 2}} \right) = \]

\[ = \Phi \left( q \cdot \sqrt{\frac{x}{n - 2}} - \frac{n_1}{2} \cdot \frac{(d - \delta_{\text{up}})}{\sqrt{\frac{Q_2}{n}}} \right), \]

where \( \Phi() \) denotes the cdf of \( N(0, 1) \), and \( q = t_{\alpha}(n - 2) \).

Conditional on a given \( x < c \), the joint distribution of rejecting \( H_{02} \) and \( Q_1 + Q_2 \leq c \) is given by

\[ P(t_{\text{up}} \leq t_{\alpha}(n - 2), \frac{n_1}{2} \cdot |d| \leq \sqrt{c - x} \mid Q_1 = x) = \]

\[ P \left( \frac{n_1}{2} \cdot (d - \delta_{\text{up}}) \leq q \cdot \sqrt{\frac{x}{n - 2}} \right) \cdot \frac{n_1}{2} \cdot (d - \delta_{\text{up}}) \]

\[ \leq \sqrt{c - x} - \sqrt{\frac{n_1}{2} \cdot \delta_{\text{up}}} \cdot \frac{n_1}{2} \cdot (d - \delta_{\text{up}}) \leq -\sqrt{c - x} - \sqrt{\frac{n_1}{2} \cdot \delta_{\text{up}}} = \]

\[ \Phi \left( \frac{n_1}{2} \cdot \frac{q \cdot \sqrt{x}}{n - 2} \right) - \Phi \left( -\frac{n_1}{2} \cdot \frac{q \cdot \sqrt{x}}{n - 2} \right) \]

if \( q \cdot \sqrt{x} > -\sqrt{c - x} - \frac{n_1}{2} \cdot \delta_{\text{up}} \) and 0 otherwise.

Hence, the joint distribution of rejection of \( H_{02} \) and \( Q_1 + Q_2 \leq c \) is given by

\[ P(t_{\text{up}} \leq t_{\alpha}(n - 2), Q_1 + Q_2 \leq c) = \]

\[ \int_{x=c}^{\Phi} \left[ \Phi \left( \frac{n_1}{2} \cdot \frac{q \cdot \sqrt{x}}{n - 2} \right) - \Phi \left( -\frac{n_1}{2} \cdot \frac{q \cdot \sqrt{x}}{n - 2} \right) \right] f(x) \, dx, \]

where \( f(\cdot) \) denotes the pdf of \( \chi^2(n - 2) \) and \( 0 \leq \alpha^* < c \) limit the range of integration in such a way that the integrand is larger than 0. The limits \( \alpha^* \)
and $c^*$ are those solutions of the two equations
\[
q \cdot \sqrt{x} - \sqrt{\frac{n_1}{2} \delta_{up}} = -c - \sqrt{\frac{n_2}{2} \delta_{up}} \quad \text{(A.2)}
\]
and
\[
q \cdot \sqrt{x} - \sqrt{\frac{n_1}{2} \delta_{up}} = \sqrt{c - x} - \sqrt{\frac{n_1}{2} \delta_{up}} \quad \text{(A.3)}
\]
which fall inside the interval $[0, c]$ (if there are such solutions). If no solution falls inside the interval $[0, c]$, then $r^* = 0$ and $c^* = c$, respectively. Equations (A.2) and (A.3) can both be converted into quadratic equations and thus have closed-form solutions. To find a solution, we first have to check whether the solutions of the quadratic equations corresponding to (A.2) or (A.3) are actually solutions to (A.2) or (A.3), and then check which of these solutions fall into $[0, c]$. This is a mathematically straightforward, but somewhat tedious process.

The conditional distribution of rejection of $H_{02}$ given $Q_1 + Q_2 \leq c$ is thus
\[
P(t_{up} \leq t_u(n - 2), Q_1 + Q_2 \leq c) = P(Q_1 + Q_2 \leq c) = P_{(Q_1 + Q_2 \leq c)}^{-1}.
\]

For example, let $\delta_{low} = -\delta_{up}$. Then we obtain (calculations for $n_1 = 12$, numbers in brackets for $n_1 = 24, 40$, respectively):
\[
P \{t_{low} \geq t_{1-\alpha} (n - 2), t_{up} \leq t_\alpha (n - 2), Q_1 + Q_2 \leq c\} = 0.0009 (0.0193, 0.0379),
\]
and
\[
P \{t_{low} \geq t_{1-\alpha} (n - 2), t_{up} \leq t_\alpha (n - 2) \mid Q_1 + Q_2 \leq c\}
\]
is thus
\[
0.0009 = 0.0015 (0.0340, 0.0687).
\]
Finally, the unconditional rejection probability of both null hypotheses is
\[
P \{ reject \ H_01 \ and \ H_02 \} = 0.0015 + 0.05 \cdot (1 - 0.5946) = 0.0212 (0.0410, 0.0603).
\]

**Appendix B. Some Notes on the Asymptotics of Blinded SSR in Equivalence Testing**

Assume that we start a 2-stage-clinical trial by recruiting $n_1$ patients. After this stage 1, a blinded sample size review is performed. For this purpose, we assume that there is some statistical test for the true value of the parameter $\theta$, say, (e.g., the mean $\mu$ of a normal distribution) and that there is a "nuisance parameter" $\tau$ (e.g., the variance $\sigma^2$ of the normal distribution). Blinded SSR uses the estimate $\hat{\tau}$ of $\tau$ to determine an additional sample size $n_2 = r \cdot n_1$ to be obtained in stage 2 of the trial where $r = r(\hat{\tau}) \in (0, \infty)$ is a function of the estimate. At the end of stage 2, the data from all $(1 + r) \cdot n_1$ patients is used to test $H_0$ (the hypothesis about $\theta$) using a test statistic $t$. The word "blinded" refers to the fact that the value of $\hat{\tau}$ is the only bit of data from stage 1 which is used in the sample size calculation for stage 2.

Assume now that $n_1 \to \infty$, but $r$ remains finite. Assume further that $\hat{\tau}$ is a function of $n_1$ which converges to a single constant value $w.p.1$. If $r$ is a continuous function of $\hat{\tau}$, then the $\tau$ "inherits" the convergence to a single value from $\hat{\tau}$ and the sample size review asymptotically leads to one unique value of $n_2$. Consequently, the test performed after all $(1 + r) \cdot n_1$ patients has a fixed number of patients (this number is unknown to us, but that does not matter). If the "nonadaptive" test used at the final analysis asymptotically keeps the Type I error, then this blinded SSR test does too. Asymptotically, with and without blinded SSR, we do the same test.

Note that $r = 0$ is allowed and that $\hat{\tau}$ does not have to be consistent, for example, the variance estimate used in the blinded SSR does not need to fulfill $\hat{\tau} \to \tau$ for $n_1 \to \infty$. It is sufficient for $\hat{\tau}$ to converge to a unique value $w.p.1$.

However, the SSR function must not have a discontinuity in the place that the nuisance parameter $\tau$ converges to under $H_0$. This additional subtlety has implications for testing noninferiority or equivalence, as will now be demonstrated.

Let us assume that we want to test $H_0 : \delta = \mu_1 - \mu_2 \in [-\delta_0, \delta_0]$ with independent samples $y_i \sim N(\mu_i, \sigma^2)$, $i = 1, 2$. Let $y_j$ denote the average of the sample from group $j$, $\hat{\sigma}^2$ the usual pooled variance estimate and $\hat{\sigma}_{FF}^2$ the "total" variance estimate which is obtained from calculating the usual empirical variance ignoring the group allocation. Let us furthermore assume that we are exactly on the boundary of the null space, wlog $\delta = 0$. Then $y_2 - y_1 \sim N(\delta, \sigma^2/j)$ where $j$ is an increasing function of the sample size. Finally, consider the following sample size rule:

If $\hat{\sigma}_{FF}^2 \leq q$ then $r = 0$. Otherwise $r = r^* > 0$. Thus, there are only two possible sample sizes $n_2 = 0$ or $n_2 = r^* \cdot n_1$.

As is well known from standard theory (cf. formula (2) in the main text),
\[
E(\hat{\sigma}_{FF}^2) = \sigma^2 \left(1 + \frac{n_1 n_1 n_2}{n_1 (n_1 - 1) \sigma^2} \right),
\]
where $n_2$ is the stage-1-sample size of group $j$ after stage 1, $n_1 = n_{11} + n_{12}$. Let us assume for simplicity that $n_{11} = n_{12}$. If $n_1 \to \infty$, $\hat{\sigma}_{FF}^2 \to E(\hat{\sigma}_{FF}^2)$.

Consider now the case where (possibly unknowingly) the SSR threshold was selected to be $q = E(\hat{\sigma}_{FF}^2)$. By the central limit theorem $\hat{\sigma}_{FF}^2$ converges in distribution to a Normal distribution with mean $q$. Hence, even asymptotically, $P(\hat{\sigma}_{FF}^2 < q) = P(\hat{\sigma}_{FF}^2 > q) = 0.5$. As a consequence, $r$ does not converge to a fixed constant as $n_1 \to \infty$ (even though $\hat{\sigma}_{FF}^2$ does), and an
asymptotic violation of the Type I error \( \alpha \) can (and, in actual fact, will) arise. (It will arise since \( E(\hat{\delta}) \) is influenced by the true value of \( \delta \) which in turn is related to \( \bar{y}_1 - \bar{y}_2 \) such that a small value of \( \hat{\delta} \) is pointing to a small value of \( \bar{y}_1 - \bar{y}_2 \) and a large value to a large value of \( \bar{y}_1 - \bar{y}_2 \) in stochastic tendency.)

In the case of a superiority hypothesis \( H_0 : \delta = 0 \), the situation is different. In that case, \( \hat{\delta} \) converges to the true \( \delta^* \). Here too, a small value of \( \hat{\delta} \) implies that a below-expected standard deviation enters the test statistic in the denominator, but a small value of \( \hat{\delta} \) also indicates a small \( \bar{y}_1 - \bar{y}_2 \) in stochastic tendency.

For the superiority hypothesis \( H_0 : \delta = 0 \), a small \( \bar{y}_1 - \bar{y}_2 \), however, is associated with nonrejection of \( H_0 \). Hence, under-estimation of \( E(\hat{\delta}) \) by \( \hat{\delta} \) and under-estimation of \( \delta \) by \( \bar{y}_1 - \bar{y}_2 \) are related and point toward the alternative in the NI-case, but do not point toward the alternative or toward the null in the superiority case. If \( \hat{\delta} \) is small, it may be due to small \( \sigma^* \) (this would make it easier to reject \( H_0 : \delta = 0 \)) or due to small \( \bar{y}_1 - \bar{y}_2 \) (would make it harder to reject \( H_0 : \delta = 0 \)).

### Appendix C. Simulation Algorithm

Let \( i = 1, 2 \) be the indicator of treatment groups, and \( j = 1, 2 \) be the indicator of the stages of the study, so that \( \bar{y}_i \) and \( \bar{n}_j \) are the sample mean and sample size, respectively, of treatment group \( i \), stage \( j \). In our investigation, the randomization ratio between the two treatment groups is not changed after the interim reassessment of the sample size. The following are the steps taken to generate one sample.

**Quantities from stage 1:**

1. Generate interim (first stage) sample means for the two groups
   \[
   \bar{y}_{11} \sim N \left( \delta_0, \frac{\sigma^2}{n_{11}} \right), \quad \bar{y}_{21} \sim N \left( 0, \frac{\sigma^2}{n_{21}} \right)
   \]
   and the independent sum of squares from the first stage
   \[Q_1 \sim \sigma^2 X^2(n_{11} + n_{21} - 2).\]
   In our simulation, \( n_{11} = n_{21} = \bar{n} \) and \( \sigma^2 = 1 \).

2. Derive the interim total variance
   \[
   \hat{\delta}_1^2 = \frac{Q_1 + Q_2}{n_{11} + n_{21} - 1},
   \]
   where
   \[Q_2 = \frac{n_{11}n_{21}(\bar{y}_{11} - \bar{y}_{21})^2}{n_{11} + n_{21}}.\]

3. Calculate \( \hat{N} \), the total sample size needed, using Equation (3).
4. Determine second stage sample sizes \( n_{12} \) and \( n_{22} \) as outlined in Section 4.1.
5. Calculate \( r = \frac{n_{11} + n_{21}}{n_{11} + n_{22}} \), the ratio between the second stage and the first stage sample sizes. In our simulation, \( n_{12} = n_{22} = m \).

**Quantities from stage 2:**

6. If \( m > 0 \), generate second stage sample means for the two groups
   \[
   \bar{y}_{12} \sim N \left( 0, \frac{\sigma^2}{n_{12}} \right), \quad \bar{y}_{22} \sim N \left( 0, \frac{\sigma^2}{n_{22}} \right)
   \]
   and if \( m > 1 \), the independent sum of squares from the second stage
   \[R_1 \sim \sigma^2 X^2(n_{12} + n_{22} - 2)\]
   If \( m = 1 \), \( R_1 := 0 \).

6a. If \( m = 0 \), \( \bar{y}_{12}, \bar{y}_{22}, \) and \( R_1 \) do not exist. For numerical convenience they can be set to 0.

**Quantities for final decision making:**

7. Derive the pooled sample variance:
   (a) Derive the total sum of squares
   \[
   K = v - w,
   \]
   where
   \[v = Q_1 + R_1 + \sum_{i=1}^{2} \sum_{j=1}^{2} n_{ij}\bar{y}_{ij}^2, \]
   and
   \[w = \sum_{i=1}^{2} n_{1i}\bar{y}_{1i}^2 + 2r \sum_{i=1}^{2} n_{1i}\bar{y}_{1i}\bar{y}_{2i} + r^2 \sum_{i=1}^{2} n_{ii}\bar{y}_{2i}^2.\]
   If \( m = 0 \), this results in \( K = Q_1 \).
   (b) The pooled sample variance is
   \[
   s^2_{pooled} = \frac{K}{n - 2},
   \]
   where \( n = n_{11} + n_{12} + n_{21} + n_{22} \).

8. Construct the 1 – 2\( \alpha \) level confidence interval:
   (a) Calculate the difference in the sample means of the two treatment groups
   \[
   y_0 = \bar{y}_1 - \bar{y}_2
   \]
   where
   \[
   \bar{y}_1 = \frac{1}{n_{11} + n_{12}} \sum_{i=1}^{2} n_{1i}\bar{y}_{1i}, \quad \bar{y}_2 = \frac{1}{n_{21} + n_{22}} \sum_{i=1}^{2} n_{2i}\bar{y}_{2i}.
   \]
   (b) Calculate the standard error of the mean
   \[
   se = \sqrt{\frac{s^2_{pooled}}{n_{11} + n_{12}} + \frac{1}{n_{21} + n_{22}}}.
   \]
   (c) The 1 – 2\( \alpha \) level confidence interval is
   \[
   y_0 \pm se \cdot t_{1-\alpha} (n - 2).
   \]

**Decision:**
9. Reject \( H_0 \) if the confidence interval is completely within the reference range; otherwise do not reject \( H_0 \).

### Supplementary Materials

1. Heatmap of average sample size for selected simulated settings as mentioned in Section 5.3.
2. R code used to simulate the settings as given in Table 1.

### Acknowledgments

We would like to thank the anonymous reviewers and the editors for their thoughtful comments. They have helped to improve the article considerably.

### References

Birkett, M. A., and Day, S. J. (1994), "Internal Pilot Studies for Estimating Sample Size," *Statistics in Medicine*, 13, 2455–2463. [166]
Bretz, F., König, F., Brannath, W., Glimm, E., and Posch, M. (2009), "Adaptive Designs for Confirmatory Clinical Trials," *Statistics in Medicine*, 28, 1181–1217. [156]
Bristol, D. R., and Shurzinske, L. (2001), "Blinded Sample Size Adjustment," *Drug Information Journal*, 35, 1123–1130. [156]
Cui, L., Hung, H. M. J., and Wang, S. J. (1999), "Modification of Sample Size in Group Sequential Clinical Trials," *Biometrics*, 55, 853–857. [156]
Davitt, B., Chen, M., Conner, D., Haidar, S. H., Kim, S., Lee, C. H., Lionberger, R. A., Makhlouf, F. T., Nwakama, P. E., Patel, D. T., and Schuurmann, D. J. (2012), "Implementation of a Reference-Scaled Aver- age Bioequivalence Approach for Highly Variable Generic Drug Products by the US Food and Drug Administration," *The AAPS Journal*, 14, 915–924. [164]
European Medicines Agency (2020), "Biosimilar Medicines: Overview" available at https://www.ema.europa.eu/en/human-regulatory/overview/biosimilar-medicines-overview. [156]
Friede, T., and Kieser, M. (2001), "Sample Size Adjustment in Clinical Trials for Proving Equivalence," *Drug Information Journal*, 35, 1401–1408. [156]
Friede, T., and Kieser, M. (2003), "Blinded Sample Size Reassessment in Non-Inferiority and Equivalence Trials," *Statistics in Medicine*, 22, 995–1007. [157]
Sample Size Recalculation in Internal Pilot Study Designs: A Review, *Biometrical Journal*, 48, 537–555. [156,157]

Sample Size Reassessment in Non-Inferiority Trials: Internal Pilot Study Designs With ANCOVA, *Methods of Information in Medicine*, 50, 237–243. [157]

Blinded Sample Size Re-Estimation in Superiority and Noninferiority Trials: Bias Versus Variance in Variance Estimation, *Pharmaceutical Statistics*, 12, 141–146. [157]

Blinded Sample Size Recalculation in Noninferiority Trials: A Case Study in Dermatology, *Drug Information Journal*, 44, 598–608. [165]

Blinded Sample Size Re-Estimation in Crossover Bioequivalence Trials, *Pharmaceutical Statistics*, 13, 157–162. [156]

PASS 11, Kaysville, UT: NCSS, LLC., available at www.ncss.com. [159]

Simple Procedures for Blinded Sample Size Adjustment That Do Not Affect the Type I Error Rate, *Statistics in Medicine*, 22, 3571–3581. [156,157]

PowerTOST: Power and Sample Size Based on Two One-Sided t-Tests (TOST) for (Bio)equivalence Studies, R Package Version 1.4-7, available at https://CRAN.R-project.org/package=PowerTOST. [159]

Adaptive Sample Size Calculations in Group Sequential Trials, *Biometrics*, 55, 1286–1290. [156]

Distribution of the Two-Sample t-Test Statistic Following Blinded Sample Size Re-Estimation, *Pharmaceutical Statistics*, 15, 208–215. [156,157]

Controlling the Type I Error Rate in Two-Stage Sequential Adaptive Designs When Testing for Average Bioequivalence, *Statistics in Medicine*, 37, 1587–1607. [166]

Connections Between Permutation and t-Tests: Relevance to Adaptive Methods, *Statistics in Medicine*, 33, 4734–4742. [156,165]

A Method for Determining the Size of Internal Pilot Studies, *Statistics in Medicine*, 15, 1587–1590. [166]

A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability, *Journal of Pharmacokinetics and Pharmacodynamics*, 15, 657–680. [157,159]

A Two-Sample Test for a Linear Hypothesis Whose Power Is Independent of the Variance, *The Annals of Mathematical Statistics*, 16, 243–258. [156]

Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry, [156,165]

Biosimilar and Interchangeable Products, available at https://www.fda.gov/drugs/biosimilars/biosimilar-and-interchangeable-products. [156]

Biosimilar Development, Review, and Approval, available at https://www.fda.gov/drugs/biosimilars/biosimilar-development-review-and-approval. [156]

The Role of Internal Pilot Studies in Increasing the Efficiency of Clinical Trials, *Statistics in Medicine*, 9, 65–72. [156]