Received: 31 May 2017 | Revised: 6 June 2018 | Accepted: 7 June 2018

DOI: 10.1002/bimj.201700092

RESEARCH PAPER

Biometrical Journal

Blinded and unblinded sample size reestimation in crossover trials balanced for period

Michael J. Grayling | Adrian P. Mander | James M. S. Wason

MRC Biostatistics Unit, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge, UK

Correspondence
Michael J. Grayling, MRC Biostatistics Unit, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge CB2 0SR, UK.
Email: mjg211@cam.ac.uk

Funding information
Medical Research Council, Grant/Award Numbers: MC_UU_00002/3, MC_UU_00002/6

Abstract

The determination of the sample size required by a crossover trial typically depends on the specification of one or more variance components. Uncertainty about the value of these parameters at the design stage means that there is often a risk a trial may be under- or overpowered. For many study designs, this problem has been addressed by considering adaptive design methodology that allows for the re-estimation of the required sample size during a trial. Here, we propose and compare several approaches for this in multitreatment crossover trials. Specifically, regulators favor reestimation procedures to maintain the blinding of the treatment allocations. We therefore develop blinded estimators for the within and between person variances, following simple or block randomization. We demonstrate that, provided an equal number of patients are allocated to sequences that are balanced for period, the proposed estimators following block randomization are unbiased. We further provide a formula for the bias of the estimators following simple randomization. The performance of these procedures, along with that of an unblinded approach, is then examined utilizing three motivating examples, including one based on a recently completed four-treatment four-period crossover trial. Simulation results show that the performance of the proposed blinded procedures is in many cases similar to that of the unblinded approach, and thus they are an attractive alternative.

KEYWORDS

blinded, crossover trial, internal pilot study, sample size reestimation

1 | INTRODUCTION

Crossover trials, in which participants are randomly allocated to receive a sequence of treatments across a series of time periods, are an extremely useful tool in clinical research. Their nature permits each patient to act as their own control, exploiting the fact that in most instances the variability of measurements on different subjects in a study will be far greater than that on the same subject. In this way, crossover trials are often more efficient than parallel group trials. Like most experimental designs, the determination of the sample size required by a crossover trial, to achieve a certain power for a particular treatment effect, depends on the significance level, and at least one factor that accounts for the participant's variance in response to treatment. While the former are designated quantities, the variance factors will usually be subject to substantial uncertainty at the design stage. Their value will often be greatly affected by components of the current trial, such as inclusion/exclusion criteria for example, that renders estimates obtained from previous trials biased. This is troubling since sample size calculation is of paramount importance in study design. Planning a trial that is too large results in an unnecessary number of patients being made susceptible to interventions that may be harmful. It also needlessly wastes valuable resources in terms of time, money, and available trial participants. In contrast, too small a sample size confers little chance of success for a trial. The consequences of this could be far reaching: a wrong decision may lead to the halting of the development of a therapy, which could deprive future patients of a valuable treatment option.

To address this problem in a parallel group setting with normally distributed outcome variables, Wittes and Brittain (1990), building upon previous work by Stein (1945), proposed the internal pilot study design. In their approach, at an interim time period the accrued data is unblinded, the within-group variance computed, and the trial's required sample size adjusted if necessary. However, unblinding an ongoing trial can reduce its integrity and introduce bias (ICH, 1998). Consequently, Gould and Shih (1992) explored several approaches for reestimating the required sample size in a blinded manner. Since then, a number of papers have advocated for reestimation in a parallel group setting to be based upon a crude one-sample estimate of the variance, and methodology has also been proposed that allows the type-I error-rate to be more accurately controlled (Kieser & Friede, 2003). More recently, much work has been conducted on similar methods for an array of possible trial designs and types of outcome variable (see, e.g. Jensen & Kieser, 2010; and Togo & Iwasaki, 2011), with these methods also gaining regulatory acceptance (CHMP, 2007; FDA, 2010).

Thus, today, sample size reestimation procedures have established themselves for parallel group trials as an advantageous method to employ when there is pre-trial uncertainty over the appropriate sample size. In contrast, there has been little exploration of such methodology within the context of multitreatment crossover trials. Golkowski, Friede, and Kieser (2014) recently explored a blinded sample size reestimation procedure for establishing bioequivalence in a trial utilizing an AB/BA crossover design. Jones and Kenward (2014) discussed how the results of Kieser and Friede (2003) could be rephrased for an AB/BA crossover trial testing for superiority. In addition, several unblinded reestimation procedures for AB/BA bioequivalence trials have been proposed (Potvin et al., 2007; Montague et al., 2012; Xu et al., 2016), the performance of which has recently been extensively compared (Kieser & Rauch, 2015). The work of Lake, Kammann, Klar, and Betensky (2002) and van Schie and Moerbeek (2014) on sample size reestimation in cluster randomized trials has some parallels with the methodology required for crossover trials, because of the necessitated mixed model for data analysis. Likewise, this is true of the methodology presented by Zucker and Denne (2002) on reestimation procedures for longitudinal trials. However, we are unaware of any article that explicitly discusses reestimation in crossover trials with more than two-treatments. There are many examples of such trials in the literature, while they also remain the focus of much research (see, e.g. Bailey & Druilhet, 2014; and Lui & Chang, 2016).

In this article, we consider several possible approaches to the interim reassessment of the sample size required by a multitreatment crossover trial. We assume a normally distributed outcome variable, and that a commonly utilized linear-mixed model will be employed for data analysis. We focus primarily on a setting in which the final analysis is based on many-to-one comparisons for one-sided null hypotheses, but provide additional guidance for other possibilities in the Supplementary Material. Blinded procedures for estimating the between and within person variance in response are proposed, following either simple or block randomisation to sequences that are balanced for period. The performance of these estimators is contrasted to that of an unblinded procedure via a simulation study motivated by a real four-treatment four-period crossover trial. Additionally, in the Supplementary Material we provide results for two additional examples. We now proceed by specifying the notation used in the re-estimation procedures. Our findings are then summarized in Section 3, before we conclude in Section 4 with a discussion.

2 | METHODS

2.1 | Hypotheses, notation, and analysis

We consider a crossover trial with D treatments, indexed $d = 0, \dots, D-1$. Treatments $d = 1, \dots, D-1$ are considered experimental, and are to be compared to the common control $d = 0$. We suppose that K sequences, indexed $k = 1, \dots, K$, are utilised for treatment allocation, and denote by n_k the number of patients allocated to sequence k. The number of periods in the trial, which is equal to the length of each of the sequences, is denoted by P.

We restrict our focus to trials with normally distributed outcome data, to be analysed using the following linear-mixed model

$$y_{ijk} = \mu_0 + \pi_j + \tau_{\mathrm{d}(j,k)} + s_{ik} + \epsilon_{ijk}, \qquad i = 1, \dots, n_k,\ j = 1, \dots, P,\ k = 1, \dots, K. \tag{1}$$

Here

- **(i)** y_{ijk} is the response for individual i, in period j, on sequence k;
- **(ii)** μ_0 is an intercept term; the mean response on treatment 0 in period 1;

(iii) π_j is a fixed effect for period j, with the identifiability constraint $\pi_1 = 0$;

(iv) $\tau_{\mathrm{d}(j,k)}$ is a fixed direct treatment effect for the treatment administered to an individual in period j, on sequence k, with the identifiability constraint $\tau_0 = 0$. Thus $\mathrm{d}(j,k) = 0, \dots, D-1$;

(v) $s_{ik} \sim N(0, \sigma_b^2)$ is a random effect for individual i on sequence k;

(vi) $\epsilon_{ijk} \sim N(0, \sigma_e^2)$ is the residual for the response from individual i, in period j, on sequence k.

This model, and its implied covariance structure, is the standard for a crossover trial that ignores the possible effects of carryover. Thus we are implicitly heeding the advice of Senn (1992), and others, that a crossover trial should not be conducted when carryover is likely to be an issue. Furthermore, note that by the above, two observations $y_{i_1 j_1 k_1}$ and $y_{i_2 j_2 k_2}$ are independent unless $i_1 = i_2$ and $k_1 = k_2$.

We assume that the following hypotheses are to be tested, to attempt to establish the superiority of each experimental intervention versus the control

$$H_{0d} : \tau_d \le 0, \qquad H_{1d} : \tau_d > 0, \qquad d = 1, \dots, D-1.$$

Note though that for Examples 1 and 3, slightly different hypotheses are assessed, as negative effects imply efficacy. Additionally, in the Supplementary Material we detail how one can handle alternate hypotheses of interest.

We suppose that it is desired to strongly control the FWER, the maximal probability of one or more incorrect rejections among the family of null hypotheses for all possible treatment effects, to some specified level $\alpha \in (0,1)$. There are several possible ways to define power in a multitreatment setting. Throughout, we assume that pairwise power of at least $1-\beta \in (0,1)$ to reject, without loss of generality, H_{01} is required when $\tau_1 = \delta > 0$ for designated type-II error-rate β and clinically relevant difference δ. Thus, from here, when referring to power we mean the probability that H_{01} is rejected. However, in the Supplementary Material we describe how a desired familywise power could be achieved.

To test the hypotheses, we assume that N patients in total will be recruited to the trial, with each randomized to one of the K sequences, and that the the linear-mixed model (1) will be fitted to the accumulated data. Note that in fitting this model, a choice must be made over whether to utilize maximum likelihood, or restricted error maximum likelihood (REML), estimation. Given the bias of the maximum likelihood estimator of the variance components of a linear-mixed model in finite samples, and that crossover trials are often conducted with relatively small sample sizes, here we always take the latter approach. Note though that this would have little effect for larger sample sizes. For further details on these considerations, we refer the reader to, for example, Fitzmaurice, Laird, & Ware (2011). In brief, the REML estimation procedure, for a linear-mixed model of the form $\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{b} + \boldsymbol{\epsilon}$ with $\boldsymbol{b} \sim N(\boldsymbol{0}, G)$ and $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, R)$, iteratively optimizes the parameter estimates for the effects in the model. The following modified log-likelihood is maximized to provide an estimate, $\hat{\Sigma}$, for $\Sigma = ZGZ^\top + R$, using an estimate, $\hat{\boldsymbol{\beta}}$, for $\boldsymbol{\beta}$

$$-\frac{1}{2}\log|\Sigma| - \frac{1}{2}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^\top \Sigma^{-1}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) - \frac{1}{2}\log|X^\top \Sigma^{-1} X|.$$

Then, $\hat{\boldsymbol{\beta}}$ is updated to

$$\hat{\boldsymbol{\beta}} = (X^\top \hat{\Sigma}^{-1} X)^{-1} X^\top \hat{\Sigma}^{-1} \boldsymbol{y},$$

and the process repeated. Given the final solutions $\hat{\boldsymbol{\beta}}$ and $\hat{\Sigma}$, we take $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (X^\top \hat{\Sigma}^{-1} X)^{-1}$.

In our case, $\boldsymbol{\beta} = (\mu_0, \pi_2, \dots, \pi_P, \tau_1, \dots, \tau_{D-1})^\top$, and the following $D-1$ Wald test statistics are formed

$$T_d = \frac{\hat{\tau}_d}{\sqrt{\mathrm{Var}(\hat{\tau}_d)}}, \qquad d = 1, \dots, D-1,$$

where $\hat{\tau}_d$ and $\mathrm{Var}(\hat{\tau}_d)$ are extracted from $\hat{\boldsymbol{\beta}}$ and $\mathrm{Var}(\hat{\boldsymbol{\beta}})$, respectively.

Next, we reject H_{0d} if $T_d > e$, with e chosen to control the FWER. Explicitly, using a Dunnett test (Dunnett, 1955), we take e as the solution to

$$1 - \alpha = \Psi_{D-1}\{(e, \dots, e)^T, \mathrm{Var}(\boldsymbol{T}), \nu_N\}, \tag{2}$$

where $\Psi_M\{\mathbf{x}, \Lambda, \nu\}$ is the M-dimensional cumulative distribution function of a central multivariate t-distribution with covariance matrix Λ and ν degrees of freedom. We take the degrees of freedom here, for sample size N, to be

$\nu_N = (N-1)(P-1)-(D-1)$, which arises from that associated with an analogous multilevel ANOVA design. Moreover, $\text{Var}(\boldsymbol{T})$ is the covariance matrix of $\boldsymbol{T} = (T_1, \dots, T_{D-1})^\top$, which can be calculated using $\text{Var}(\hat{\boldsymbol{\beta}})$.

Now, in this case, if σ_e^2 and σ_b^2 were known, and we assumed that $n_1 = \cdots = n_K$, we could derive a simple formula for the total number of patients, N, required to achieve the desired power for the trial. Here, we denote this formula using the function $\text{N}(\sigma_e^2, \sigma_b^2)$, explicitly stating its dependence upon the within and between person variances. In the Supplementary Material, we elaborate on how this formula can be derived.

Our problem, as discussed, is that in practice σ_e^2 and σ_b^2 are rarely known accurately at the design stage. Therefore, we propose to reestimate the required sample size at an interim analysis timed after $n_{\text{int}} \in \mathbb{N}$ patients. That is, we consider several methods to construct estimates, $\hat{\sigma}_e^2$ and $\hat{\sigma}_b^2$, for σ_e^2 and σ_b^2 , respectively, based on the data accrued up to the interim analysis. Then, the final sample size for the trial is taken as

$$\hat{N} = \begin{cases} n_{\text{int}} & \text{if } \text{N}(\hat{\sigma}_e^2, \hat{\sigma}_b^2) \le n_{\text{int}}, \\ \lceil \text{N}(\hat{\sigma}_e^2, \hat{\sigma}_b^2) \rceil & \text{if } n_{\text{int}} < \text{N}(\hat{\sigma}_e^2, \hat{\sigma}_b^2) < n_{\text{max}}, \\ n_{\text{max}} & \text{if } \text{N}(\hat{\sigma}_e^2, \hat{\sigma}_b^2) \ge n_{\text{max}}, \end{cases}$$

where $\lceil x \rceil$ denotes the nearest integer greater than or equal to x and $n_{\text{max}} \in \mathbb{N}$ is a specified maximal allowed sample size. It could be based, for example, on the cost restrictions or feasible recruitment rate of a trial. Of course, if $\hat{N} = n_{\text{max}}$ then the trial will be expected to be underpowered. Thus, if necessary, additional patients are recruited and a final analysis conducted as above based on the calculated values of the test statistics T_d, and the critical value e as defined in Equation (2).

Throughout, to give our function $\text{N}(\cdot)$ a simple form, we consider values of n_{int} that imply an equal number of patients could be allocated to each of the K sequences, and assume randomisation schemes that ensure this is the case. Moreover, for reasons to be elucidated shortly, we consider from here only settings where the K sequences are balanced for period. That is, across the chosen sequences, each treatment appears an equal number of times in each period. We now proceed by detailing each of our explored methods for estimating σ_e^2 and σ_b^2 based on the internal pilot data.

2.2 | Unblinded estimator

The first of the methods we consider is an unblinded procedure. As noted, such an approach is typically less well favored by regulatory agencies. However, though this may not always actually prove to be the case (see, e.g. Friede & Kieser, 2013), one may anticipate its performance in terms of estimating the key variance components and provided desired operating characteristics to be preferable to that of the blinded procedures. This method therefore serves as a standard against which to assess the blinded estimators. Explicitly, this approach breaks the randomization code and fits the linear-mixed model (1) to the accrued data using REML estimation. With the REML estimates of σ_e^2 and σ_b^2 obtained, they are utilized in the reestimation procedure as described above.

2.3 | Adjusted blinded estimator

Zucker, Wittes, Schabenberger, and Brittan (1999) considered a blinded estimator for two-arm parallel trial designs based on an adjustment to the one-sample variance. Golkowski et al. (2014) considered a similar unadjusted procedure for two-arm bioequivalence trials. Here, we consider a similar approach for multi-treatment crossover trials. Specifically, the following blinded estimators of the within and between person variances are used

$$\hat{\sigma}_e^2 = \frac{1}{2(P-1)(n_{\text{int}}-1)} \sum_{j=2}^{P} \sum_{k=1}^{K} \sum_{i=1}^{n_{\text{int}}/K} (p_{ijk} - \bar{p}_j)^2 +$$

$$- \frac{n_{\text{int}}}{2K(P-1)(n_{\text{int}}-1)} \sum_{j=2}^{P} \sum_{k=1}^{K} \left(\tau^*_{\text{d}(j,k)} - \tau^*_{\text{d}(j-1,k)}\right)^2,$$

$$\hat{\sigma}_b^2 = \frac{1}{2} \left\{ \frac{1}{2(P-1)(n_{\text{int}}-1)} \sum_{j=2}^{P} \sum_{k=1}^{K} \sum_{i=1}^{n_{\text{int}}/K} (q_{ijk} - \bar{q}_j)^2 - \hat{\sigma}_e^2 + \right.$$

$$-\frac{n_{\text{int}}}{2K(P-1)(n_{\text{int}}-1)}\sum_{j=2}^{P}\sum_{k=1}^{K}\left(\tau^*_{\text{d}(j,k)}+\tau^*_{\text{d}(j-1,k)}\right)^2+$$

$$\left.+\frac{2n_{\text{int}}}{D^2(n_{\text{int}}-1)}\left(\sum_{k=1}^{K}\tau_{\text{d}(1,k)}\right)^2\right\},$$

for specified τ^*_d, $d=0,\ldots,D-1$, with $\tau^*_0=0$, where

$$p_{ijk}=y_{ijk}-y_{ij-1k},$$

$$q_{ijk}=y_{ijk}+y_{ij-1k},$$

$$\bar{p}_j=\frac{1}{n_{\text{int}}}\sum_{k=1}^{K}\sum_{i=1}^{n_{\text{int}}/K}p_{ijk},$$

$$\bar{q}_j=\frac{1}{n_{\text{int}}}\sum_{k=1}^{K}\sum_{i=1}^{n_{\text{int}}/K}q_{ijk}.$$

In the Supplementary Material, we show that if $\tau^*_d=\tau_d$ for $d=1,\ldots,D-1$ then $\mathrm{E}(\hat{\sigma}^2_e)=\sigma^2_e$ and $\mathrm{E}(\hat{\sigma}^2_b)=\sigma^2_b$, and thus $\hat{\sigma}^2_e$ and $\hat{\sigma}^2_b$ are unbiased estimators for σ^2_e and σ^2_b , respectively. This is the reason for our restrictions on the employed randomization scheme (which assumes $n_1=\cdots=n_K=n_{\text{int}}/K$ at the interim reassessment), and the employed sequences (which are assumed to be balanced for period). The above estimator could be used when there is imbalance in the number of patients allocated to each sequence, or without making this restriction on the sequences, but results on the expected values of the variance components would have a more complex form. It is therefore advantageous to ensure an equal number of patients are allocated to each sequence, and also logical to utilize period-balanced sequences. We also view it as sensible therefore to explore the performance of the estimators in this case.

It is also important to assess the sensitivity of the performance of these estimators to the choice of the τ^*_d, hoping for it to have negligible impact as in analogous procedures for other trial settings (Kieser & Friede, 2002). Adapting previous works (see, e.g. Kieser & Friede, 2003; Zucker et al., 1999; Gould & Shih, 1992), we assess this procedure for $\tau^*_d=0$, and $\tau^*_d=\delta$, $d=1,\ldots,D-1$, and refer to these henceforth as the null adjusted and alternative adjusted reestimation procedures, respectively.

Note that one limitation of this approach in practice is that there is no guarantee that the above value for $\hat{\sigma}^2_b$ will be positive. Therefore, we actually reevaluate the required sample size as $\mathrm{N}\{\hat{\sigma}^2_e,\max(0,\hat{\sigma}^2_b)\}$. For the examples provided in the Supplementary Material, we demonstrate that the above procedure still performs well despite this inconvenience. Moreover, in certain routinely faced scenarios, as will be discussed shortly, the value of σ^2_b is inconsequential and this issue therefore no longer exists. However, in general this must be kept in mind when considering using this procedure for sample size reestimation.

2.4 | Blinded estimator following block randomization

The above reestimation procedures are explored within the context of a simple randomisation scheme that only ensures an equal number of patients are allocated to each sequence prior to the interim reassessment. In contrast, the final blinded estimator we consider exploits the advantages block randomization can bring, extending the methodology presented in Xing and Ganju (2005) for parallel arm trials to crossover studies.

We suppose that patients are allocated to sequences in B blocks, each of length n_B (with these values chosen such that $Bn_B=n_{\text{int}}$). We recategorize our data as y_{ijb}, the response from patient $i=1,\ldots,n_B$, in period j, in block b. Then, the following blinded estimators are used to recalculate the required sample size

$$\hat{\sigma}^2_e=\frac{1}{2(P-1)(n_{\text{int}}-B)}\sum_{j=2}^{P}\sum_{b=1}^{B}\sum_{i=1}^{n_B}(p_{ijb}-\bar{p}_{jb})^2,$$

$$\hat{\sigma}^2_b=\frac{1}{2}\left\{\frac{1}{2(P-1)(n_{\text{int}}-B)}\sum_{j=2}^{P}\sum_{b=1}^{B}\sum_{i=1}^{n_B}(q_{ijb}-\bar{q}_{jb})^2-\hat{\sigma}^2_e\right\},$$

where

$$p_{ijb} = y_{ijb} - y_{ij-1b},$$

$$q_{ijb} = y_{ijb} + y_{ij-1b},$$

$$\bar{p}_{jb} = \frac{1}{n_B}\sum_{i=1}^{n_B} p_{ijb},$$

$$\bar{q}_{jb} = \sum_{i=1}^{n_B}\sum_{i=1}^{n_B} q_{ijb}.$$

In the Supplementary Material, provided that an equal number of patients are allocated to each of a set of period balanced sequences, these are also shown to be unbiased estimators for σ_e^2 and σ_b^2. Note though that as above, we must actually reestimate N using $\mathrm{N}\{\hat{\sigma}_e^2, \max(0, \hat{\sigma}_b^2)\}$. Additionally, when using block randomization, the actual sample size used by a trial may differ from $\hat{N}$, if it is not divisible by the block length n_B.

3 | SIMULATION STUDY

3.1 | Motivating examples

We present results for three motivating examples based on real crossover trials. Example 1 is described in Section 3.2, with Examples 2 and 3 discussed in the Supplementary Material, where their associated results are also presented. Among the three examples we consider settings with a range of required sample sizes, utilising complete block, incomplete block, and extra-period designs. This allows us to provide a thorough depiction of the performance of the various estimators in a wide range of realistic trial design settings.

R (R Core Team, 2016) source code to reproduce our results is available as Supporting Information on the journal's web page (`http://onlinelibrary.wiley.com/doi/10.1002/bimj.201700092/suppinfo`).

3.2 | Example 1: TOMADO

First, we assess the performance of the various reestimation procedures using the TOMADO trial as motivation. TOMADO compared the clinical effectiveness of a range of mandibular devices for the treatment of obstructive sleep-apnea hypopnea. Precise details can be found in Quinnell et al. (2014). Briefly, TOMADO was a four-treatment four-period crossover trial, with patients allocated treatment sequences using two Williams squares. The data for the outcome Epworth Sleepiness Scale was to be analyzed using linear-mixed model (1), with the following hypotheses tested

$$H_{0d} : \tau_d \geq 0, \qquad H_{1d} : \tau_d < 0, \qquad d = 1, \dots, D-1,$$

since a reduction in the Epworth Sleepiness Scale score is indicative of an efficacious treatment. Consequently, the null hypotheses were to be rejected if $T_d < -e$, using the value of e determined as above.

Following the methodology described in the Supplementary Material, we can demonstrate that when complete-block period-balanced sequences are used for treatment allocation, that the required sample size has no dependence upon the between person variance σ_b^2. Explicitly, we have

$$\mathrm{N}(\sigma_e^2, \sigma_b^2) \equiv \mathrm{N}(\sigma_e^2) = \frac{2\sigma_e^2(z_{1-\alpha_*} + z_{1-\beta})^2}{\delta^2},$$

where α_* is defined in the Supplementary Material. See Jones and Kenward (2014), for an alternative derivation of this formula. This substantially simplifies the reestimation procedure, as we only need to provide a value for σ_e^2, and do not require use of the estimators for σ_b^2.

TOMADOs complete case analysis estimated the following values for the various components of the linear-mixed model (1)

$$\hat{\mu}_0 = 10.65, \quad \hat{\pi}_2 = -0.77, \quad \hat{\pi}_3 = -0.96, \quad \hat{\pi}_4 = -0.55,$$

$$\hat{\tau}_1 = -1.51, \quad \hat{\tau}_2 = -2.15, \quad \hat{\tau}_3 = -2.37, \quad \hat{\sigma}_e^2 = 6.51, \quad \hat{\sigma}_b^2 = 10.12.$$

Therefore, for $\sigma_e^2 = \hat{\sigma}_e^2$, the trials planned recruitment of 72 patients would have conferred power of 0.8 at a significance level of 0.05 for $\delta = -1.24$. Consequently, we set $\beta = 0.2$ and $\alpha = 0.05$ throughout. In the main manuscript, we additionally take $\delta = -1.24$ and $\sigma_b^2 = 10.12$ always. The effect of other underlying values for δ and σ_b^2 is considered in the Supplementary Material. In contrast, whilst we focus here on the case with $\sigma_e^2 = 6.51$, we also consider the influence of alternative values for this parameter. When simulating data we take $\mu_0 = 10.65$, $\pi_2 = -0.77$, $\pi_3 = -0.96$, and $\pi_4 = -0.55$. However, the effect of other period effects is discussed in Section 4 and in the Supplementary Material.

We explore the performance of the procedures under the global null hypothesis ($\tau_1 = \tau_2 = \tau_3 = 0$), when only treatment one is effective ($\tau_1 = \delta, \tau_2 = \tau_3 = 0$), when treatments one and two are effective ($\tau_1 = \tau_2 = \delta, \tau_3 = 0$), under the global alternative hypothesis ($\tau_1 = \tau_2 = \tau_3 = \delta$), and under what we refer to henceforth as the observed treatment effects ($\tau_1 = -1.51, \tau_2 = -2.15, \tau_3 = -2.37$). For simplicity, we assume a single Latin square was used for treatment allocation, and set $n_{\text{max}} = 1{,}000$ so that there is no practical upper limit on the allowed sample size. In all cases, the average result for a particular design and analysis scenario was determined using 100,000 trial simulations.

3.3 | Distributions of $\hat{\sigma}_e^2$ and $\hat{N}$

First, the performance of the reestimation procedures was explored for the parameters listed in Section 3.2, with $\sigma_e^2 = 6.51$, and $n_{\text{int}} \in \{8, 16, 24, 32, 40\}$. The resulting distributions of $\hat{\sigma}_e^2$, the interim estimate of σ_e^2, are shown in Figure 1 via the median, lower and upper quartiles in each instance. Additionally, Figure 2 depicts the equivalent results for the distribution of $\hat{N}$, the interim reestimated value for N. The results are grouped according to the timing of the reestimation and by the true value of the treatment effects. Note that $n_B = 4$ is only considered for values of n_{int} that allows an equal number of patients to be allocated to each sequence by the interim analysis.

The median value of $\hat{\sigma}_e^2$ for the unblinded procedure is always close to, but typically slightly less than, the true value σ_e^2. The same statement holds for the block randomization procedure with $n_B = 2$ or 4. However, while this is true for the adjusted procedures under the global null hypothesis, it is not otherwise always the case. In particular, both perform poorly for the observed treatment effects.

As would be anticipated, the alternative adjusted procedure has lower median values for $\hat{\sigma}_e^2$ than the null adjusted procedure. Moreover, using the block randomised reestimation procedure with $n_B = 4$ seems to improve performance over $n_B = 2$, both in terms of the median value of $\hat{\sigma}_e^2$, and by imparting a smaller interquartile range for $\hat{\sigma}_e^2$.

The results for $\hat{N}$ mirror those for $\hat{\sigma}_e^2$. Thus $\hat{N}$ is larger for the adjusted estimators under the observed treatment effects, but otherwise the distributions are comparable.

Increasing the value of n_{int} reduces the interquartile range for $\hat{\sigma}_e^2$ and $\hat{N}$ for each procedure, and results in median values closer to the truth, as would be expected. Finally, we observe that the interquartile range for the unblinded procedure is often smaller than that of its adjusted or block randomisation counterparts.

3.4 | Familywise error-rate and power

For the scenarios from Section 3.3 that were not conducted under the observed treatment effects, the estimated FWER and power were also recorded. The results are displayed in Table 1.

The FWER for each of the procedures is usually close to the nominal level, with a maximal value of 0.052 for the unblinded procedure with $n_{\text{int}} = 32$. The adjusted procedures arguably have the smallest inflation across the considered values of n_{int}.

In most cases the reestimation procedures attain a power close to the desired level. Of the adjusted procedures, the null adjusted has a larger power, as would be anticipated given our observations on $\hat{\sigma}_e^2$ and $\hat{N}$ above. In fact, the null adjusted method conveys the highest power for each value of n_{int}. The power of the block randomized procedures is typically similar to that of the alternative adjusted method. In addition, whether only treatment one, treatments one and two, or all three treatments are effective has little effect on the power.

There is no clear to trend as to the effect of increasing n_{int} on the FWER, however it leads in almost all instances to an improvement in power. Finally, increasing the value for n_B in the block randomization procedure increases power as would be predicted.

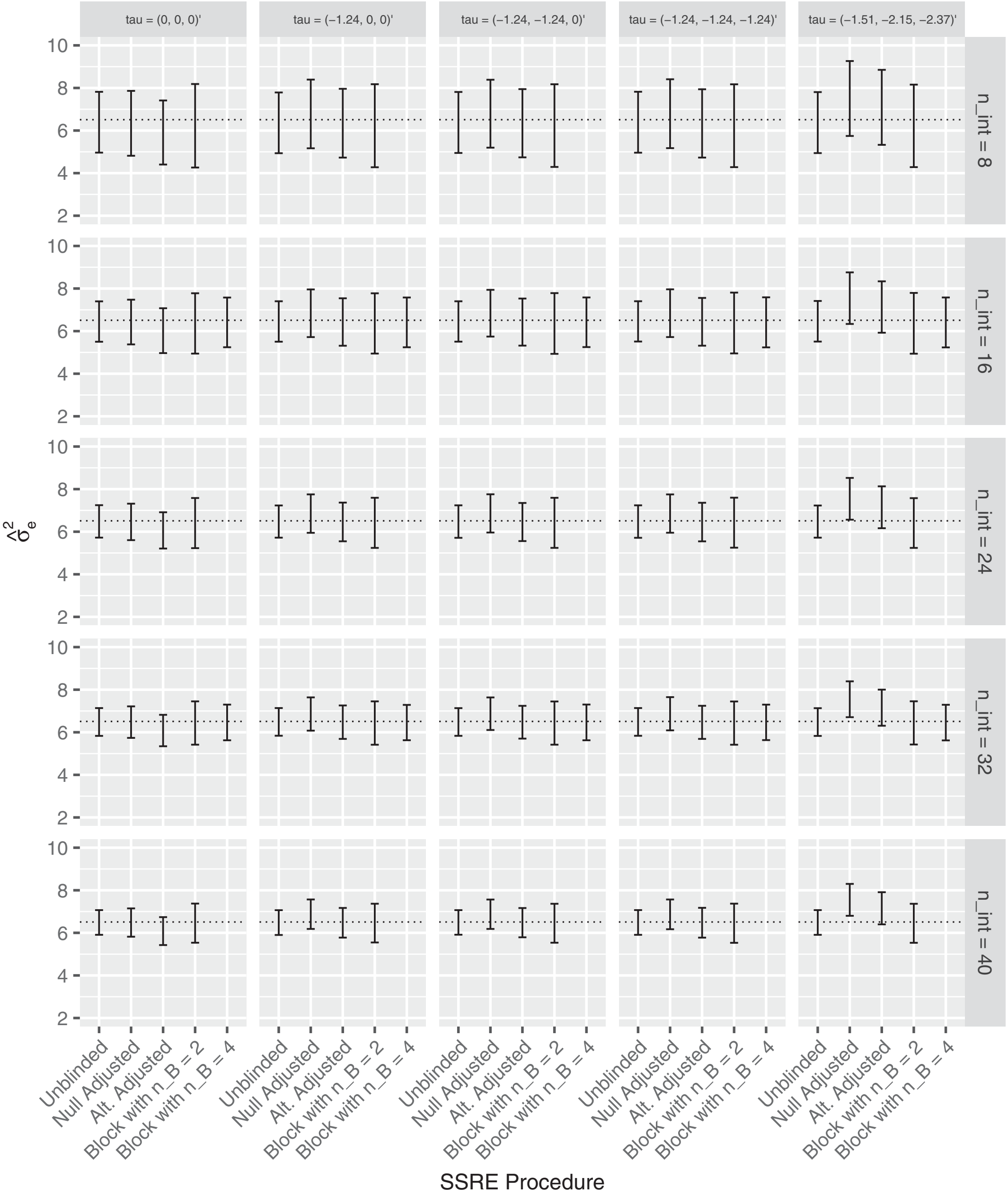

FIGURE 1 The distribution of $\hat{\sigma}_e^2$ is shown for each of the reestimation procedures for several values of $\boldsymbol{\tau}$, and several values of n_{int}, for Example 1. Precisely, for each scenario, the median, lower, and upper quartile values of $\hat{\sigma}_e^2$ across the simulations are given. The dashed line indicates the true value of σ_e^2

3.5 | Influence of σ_e^2

In this section, we consider the influence of the value of σ_e^2 on the performance of our reestimation procedures. Specifically, while we know that increasing σ_e^2 will increase the required sample size, we would like to assess the effect this has upon the ability of the methods to control the FWER and attain the desired power.

Figures 3 and 4 respectively present our results on the FWER and power of the various reestimation procedures when $n_{\text{int}} \in \{16, 32\}$ for several values of $\sigma_e^2 \in [0.25(6.51), 4(6.51)]$, under the global null and alternative hypotheses, respectively. Corresponding findings for $\hat{N}$ are provided in the Supplementary Material.

Arguably, we observe that the FWER is more variable for smaller values of σ_e^2, with it changing little for several of the procedures when $\sigma_e^2 > 10$. There is additionally some evidence to suggest that increasing the value of n_{int} reduces the overall effect σ_e^2 has on the FWER.

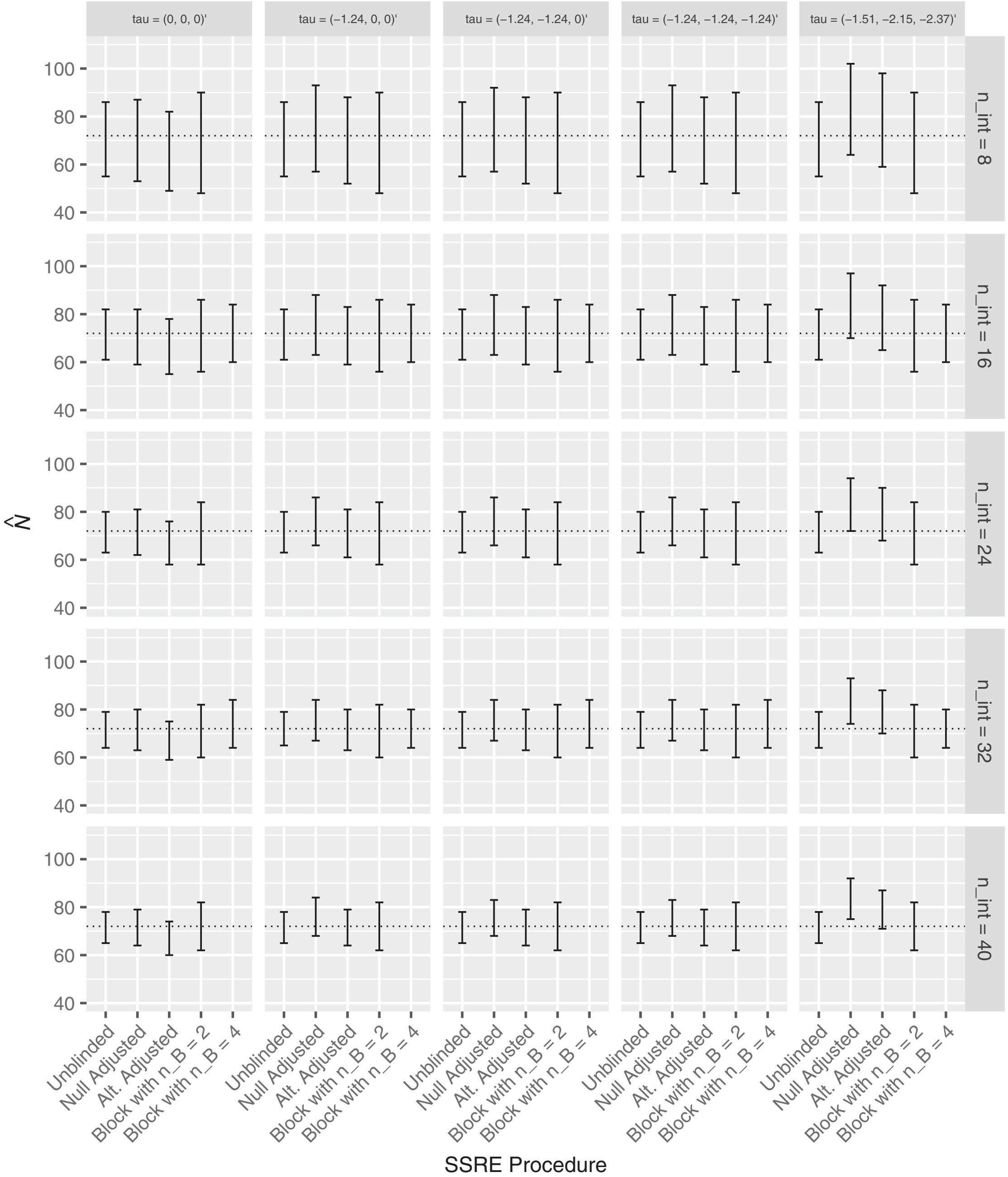

FIGURE 2 The distribution of $\hat{N}$ is shown for each of the reestimation procedures for several values of $\boldsymbol{\tau}$, and several values of n_{int}, for Example 1. Precisely, for each scenario, the median, lower, and upper quartile values of $\hat{N}$ across the simulations are given. The dashed line indicates the true required value of N

For the power, as would be anticipated, the reestimation procedures are over-powered when $n_{\text{int}} = 32$ and σ_e^2 is small. Moreover, increasing the value of n_{int} universally increases the power. Finally, as σ_e^2 increases beyond approximately $\sigma_e^2 = 5$, for both considered values of n_{int}, there is little change in power.

3.6 | Influence of δ

Here, we consider the case where $\pi_2 = -0.77$, $\pi_3 = -0.96$, $\pi_4 = -0.55$, and $\sigma_b^2 = 10.12$, focusing on the influence δ has upon the procedures FWER and power. Precisely, Figures 5 and 6 respectively present our findings for the FWER and power of the various reestimation procedures when $n_{\text{int}} \in \{16, 32\}$ for several values of $\delta \in [2(-1.24), 0.5(-1.24)]$, under the global null and alternative hypotheses, respectively. Complimentary findings for $\hat{N}$ are provided in the Supplementary Material.

TABLE 1 The estimated familywise error-rate (FWER) is shown for each of the considered reestimation procedures and several values of n_{int} under the global null hypothesis, for Example 1

			Power		
Reestimation procedure	n_{int}	**FWER**	$\tau = (\delta, 0, 0)$	$\tau = (\delta, \delta, 0)$	$\tau = (\delta, \delta, \delta)$
Unblinded	8	0.0513	0.7704	0.7694	0.7687
Null Adjusted	8	0.0496	0.7743	0.7809	0.7753
Alt. Adjusted	8	0.0500	0.7440	0.7512	0.7432
Block rand. with $n_B = 2$	8	0.0509	0.7443	0.7455	0.7428
Unblinded	16	0.0506	0.7906	0.7893	0.7867
Null Adjusted	16	0.0512	0.7956	0.8010	0.7942
Alt. Adjusted	16	0.0495	0.7702	0.7731	0.7691
Block rand. with $n_B = 2$	16	0.0512	0.7720	0.7723	0.7747
Block rand. with $n_B = 4$	16	0.0525	0.7858	0.7887	0.7868
Unblinded	24	0.0509	0.7963	0.7934	0.7950
Null Adjusted	24	0.0496	0.8019	0.8071	0.7990
Alt. Adjusted	24	0.0508	0.7776	0.7793	0.7770
Block rand. with $n_B = 2$	24	0.0504	0.7821	0.7838	0.7835
Unblinded	32	0.0520	0.7977	0.7962	0.7988
Null Adjusted	32	0.0509	0.8055	0.8109	0.8072
Alt. Adjusted	32	0.0498	0.7772	0.7857	0.7812
Block rand. with $n_B = 2$	32	0.0514	0.7907	0.7879	0.7887
Block rand. with $n_B = 4$	32	0.0511	0.8014	0.8002	0.8035
Unblinded	40	0.0516	0.7967	0.8010	0.8000
Null Adjusted	40	0.0504	0.8081	0.8115	0.8062
Alt. Adjusted	40	0.0498	0.7828	0.7858	0.7842
Block rand. with $n_B = 2$	40	0.0518	0.7914	0.7926	0.7942

Corresponding values of the power when only treatment one is effective, treatments one and two are effective, or under the global alternative hypothesis when all three experimental treatments are effective, are also shown. The Monte Carlo error of the FWER and power values is approximately 0.0007 and 0.0013, respectively in each instance. All figures are given to four decimal places

In Figure 5 we can see that there is no clear pattern to the effect on the FWER of changing δ, with the fluctuations for several of the estimators relatively small. However, there is some evidence to suggest that increasing the value of δ (i.e., making it closer to zero) reduces the FWER, as may be expected as this implies a larger requisite sample size.

Similar statements are true for the power when examining Figure 6. Analogous to our discussions around Figure 4, the reestimation procedures are over-powered when $n_{\text{int}} = 32$ and δ is large in magnitude. Furthermore, increasing the value of n_{int} once more universally increases the power, while there appears to be a point beyond which the power remains relatively constant.

3.7 | Sample size inflation factor

While the above suggests the overall performance of the reestimation is good, there are several simple refinements that can be implemented to improve the observed results.

One such refinement, to help ensure the power provided by the reestimation procedures is at least the desired $1 - \beta$, is to utilise a sample size inflation factor as originally proposed by Zucker et al. (1999). With it, the value of $\hat{N}$ as determined using the arguments above, is enlarged by the following factor

$$\left(\frac{t_{1-\alpha, \nu_{n_{\text{int}}}} + t_{1-\beta, \nu_{n_{\text{int}}}}}{z_{1-\alpha} + z_{1-\beta}} \right)^2 .$$

Of course, one must be careful that the new implied sample size does not exceed any specified value of $n_{\max}$. However, this factor has then been shown to improve the performance of reestimation procedures in both superiority (Zucker et al., 1999), noninferiority (Friede & Kieser, 2013), and two-treatment bioequivalence trials (Golkowski et al., 2014).

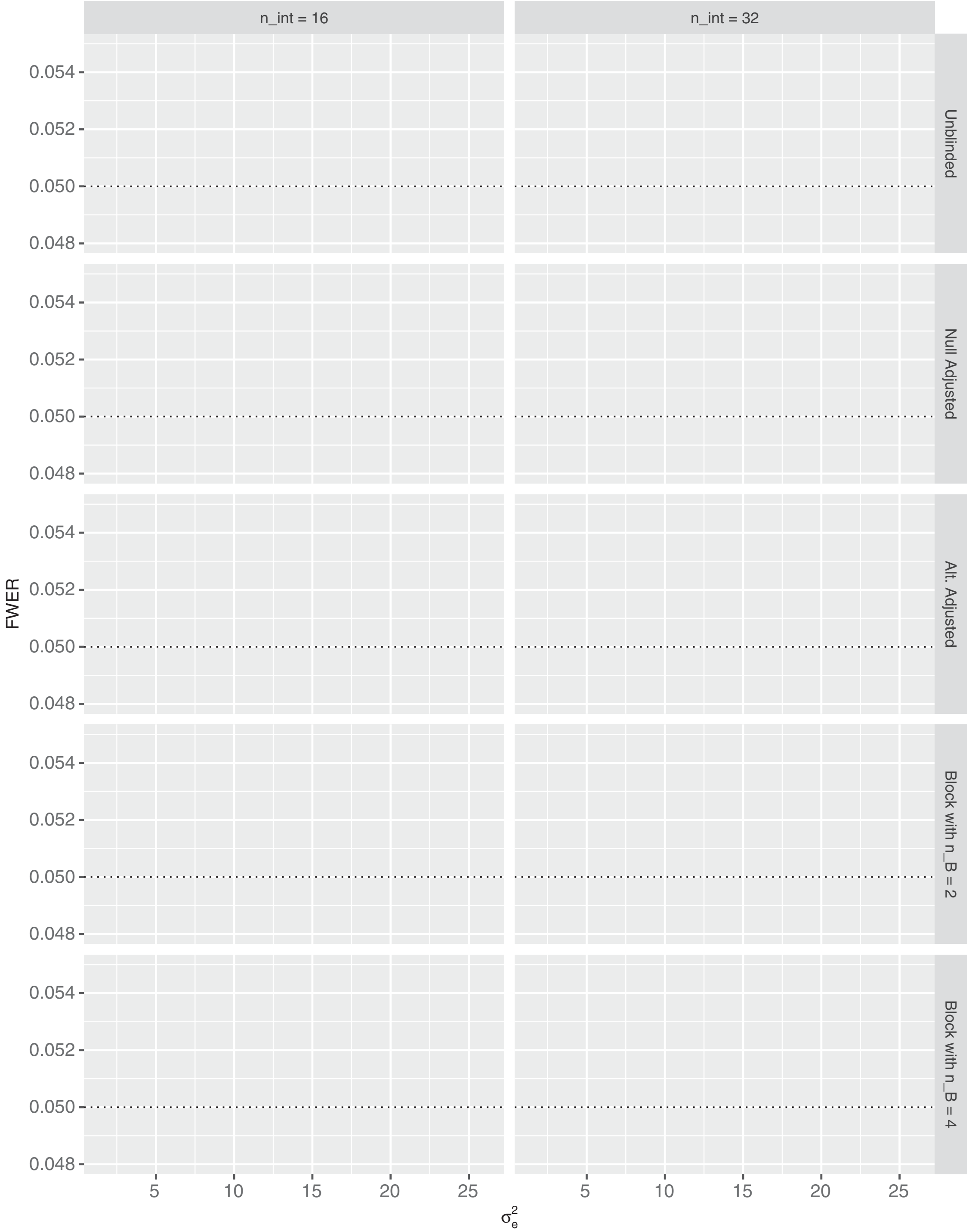

FIGURE 3 The simulated familywise error-rate (FWER) is shown under the global null hypothesis for each of the reestimation procedures when $n_{\text{int}} \in \{16, 32\}$, as a function of the within person variance σ_e^2, for Example 1. The Monte Carlo error is approximately 0.0007 in each instance. The dashed line indicates the desired value of the FWER

Figure 7 displays its effect in the context of our multitreatment crossover trials. Explicitly, the power of the various reestimation procedures under the global alternative hypothesis, for $n_{\text{int}} \in \{8, 16, 24, 32, 40\}$ and $\sigma_e^2 = 6.51$, is shown with and without the use of the inflation factor. For the unblinded, null adjusted, and block randomized method with $n_B = 4$, the inflation factor increases power to above the desired level in every instance. Consequently, this simple inflation factor appears once more to be an effective adjustment to the basic procedures.

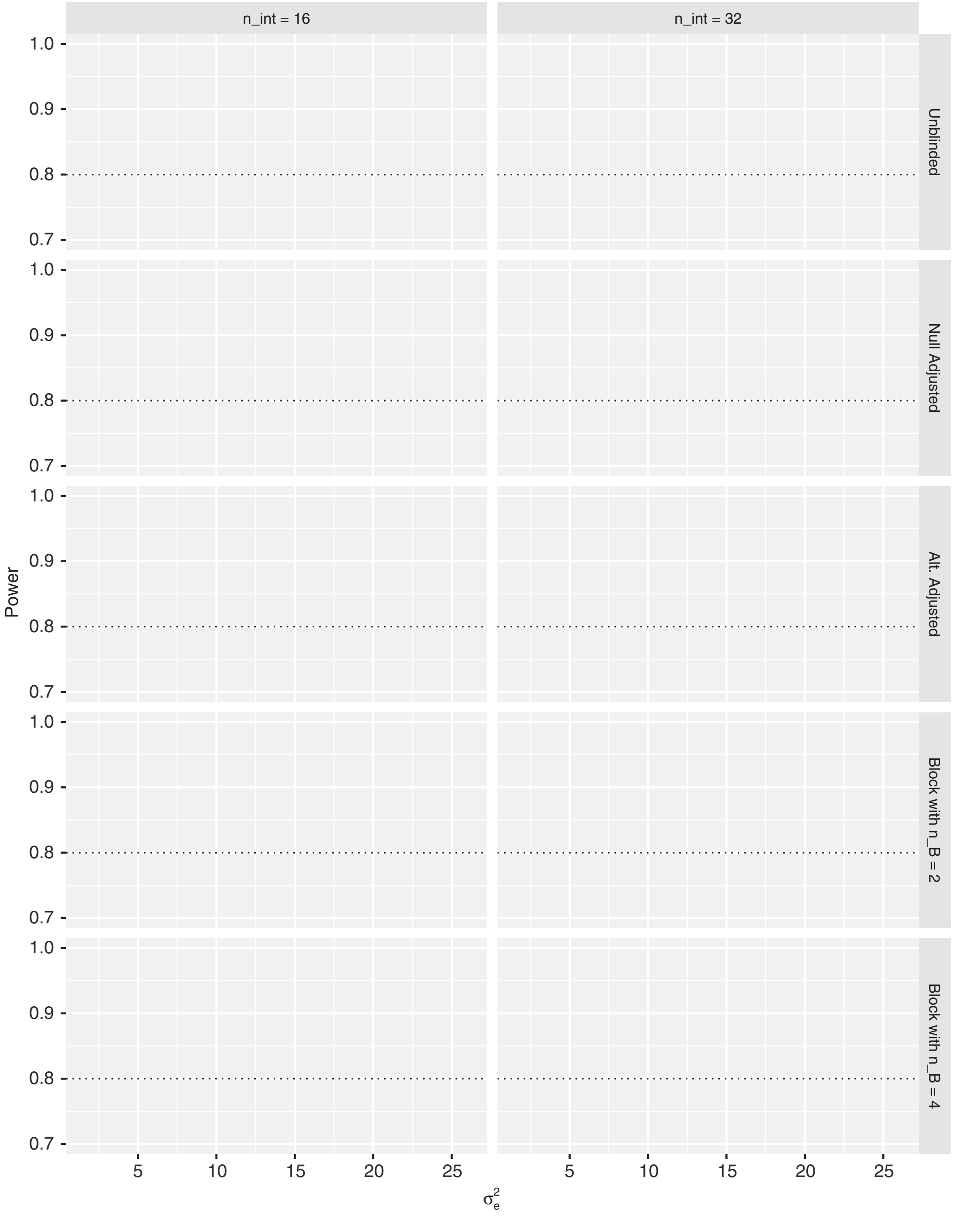

FIGURE 4 The simulated power is shown under the global alternative hypothesis for each of the reestimation procedures when $n_{\text{int}} \in \{16, 32\}$, as a function of the within person variance σ_e^2, for Example 1. The Monte Carlo error is approximately 0.0013 in each instance. The dashed line indicates the desired value of the power

4 | DISCUSSION

In this article, we have developed and explored several methods for the interim re-assessment of the sample size required by a multitreatment crossover trial. Our methodology is applicable to any trial analyzed using the linear-mixed model (1), when there is equal participant allocation to a set of period-balanced sequences. Thus while adapting the work of Golkowski et al. (2014) would be advisable in the case of an AB/BA superiority trial, given that it does not require the use of simulation, our methods

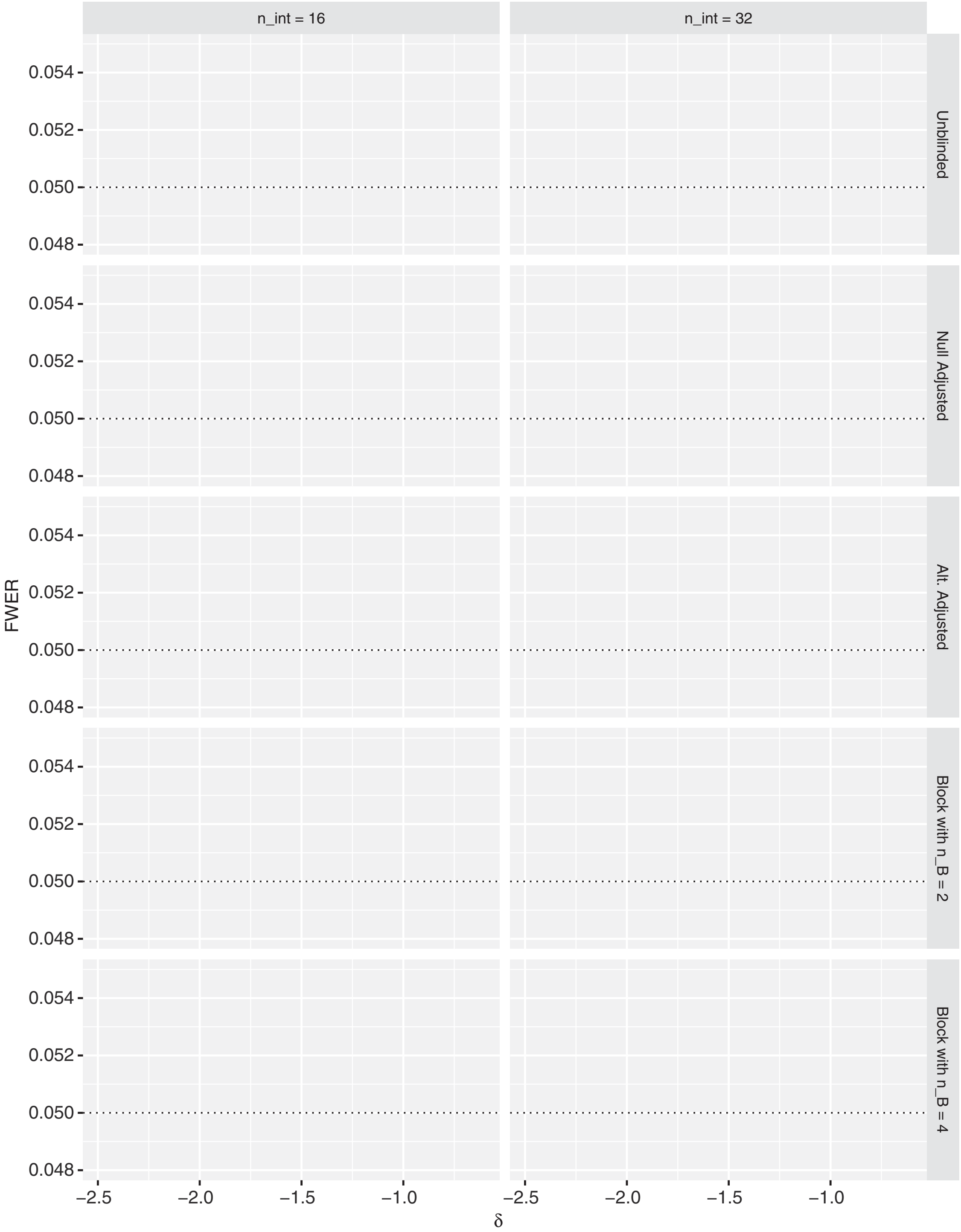

FIGURE 5 The simulated familywise error-rate (FWER) is shown under the global null hypothesis for each of the reestimation procedures when $n_{\text{int}} \in \{16, 32\}$, as a function of the clinically relevant difference δ, for Example 1. The Monte Carlo error is approximately 0.0007 in each instance. The dashed line indicates the desired value of the FWER

are pertinent to a broader set of crossover designs. Indeed, they are as readily applicable to multitreatment superiority trials as they are ones for establishing bioequivalence.

We explored performance via three motivating examples, allowing consideration of settings with different types of sequences and a range of required sample sizes. Overall, the results presented here for the TOMADO trial are similar to those provided in the Supplementary Material for Examples 2 and 3. However, larger inflation to the FWER was observed in Example 2, most likely as a consequence of its associated smaller sample sizes. Nonetheless, the methods were found to provide desirable power

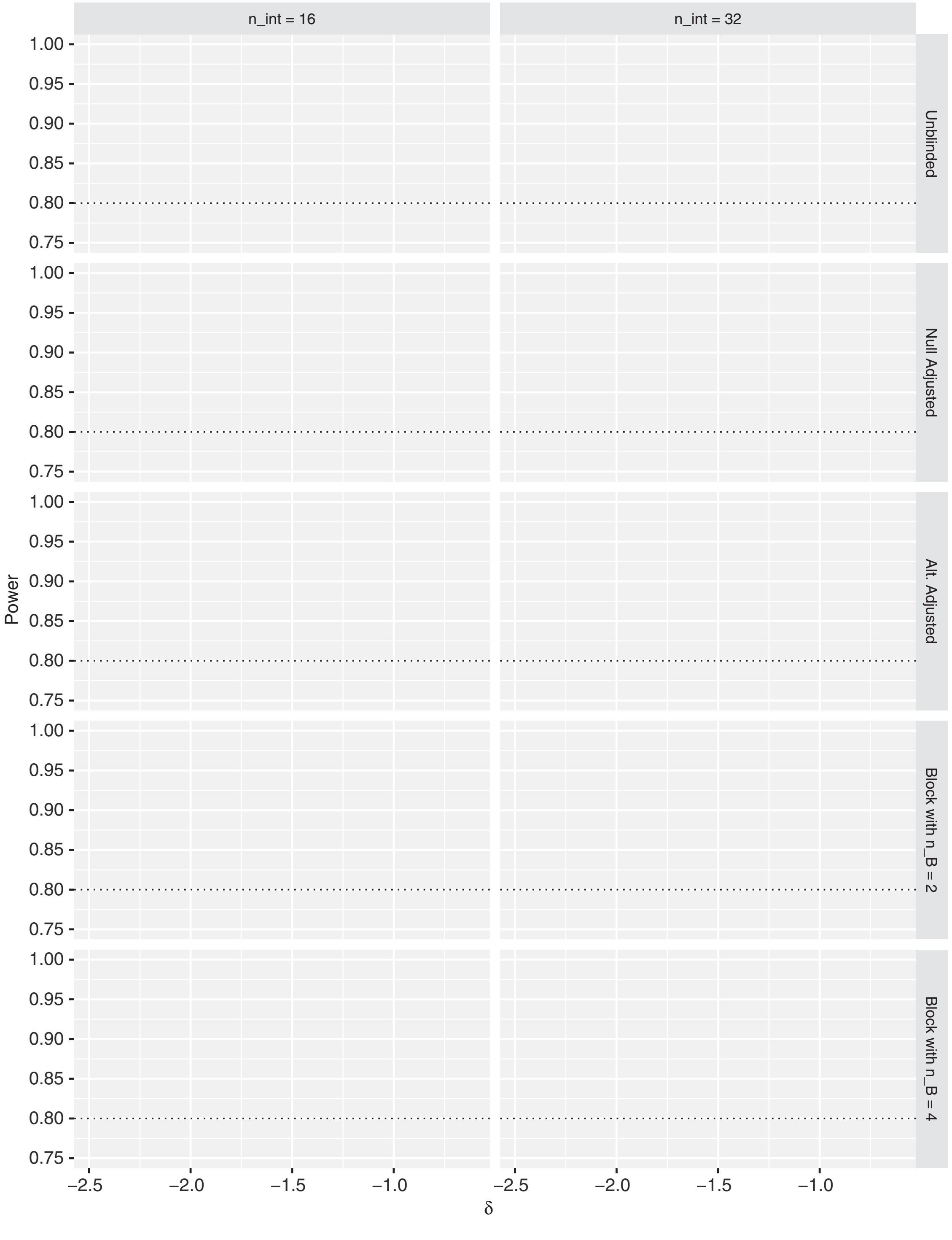

FIGURE 6 The simulated power is shown under the global alternative hypothesis for each of the reestimation procedures when $n_{\text{int}} \in \{16, 32\}$, as a function of the clinically relevant difference δ, for Example 1. The Monte Carlo error is approximately 0.0013 in each instance. The dashed line indicates the desired value of the power

characteristics with negligible inflation to the FWER in many settings. In particular, the blinded procedures provided comparable operating characteristics to the unblinded procedure, and thus can be considered viable alternatives. Following results for parallel arm trials (Friede & Kieser, 2013), the null adjusted blinded estimator arguably performed better than the other estimators in that its typical overestimation of the variance at interim led to the desired power being achieved more often. We may therefore tentatively suggest the null adjusted blinded estimator to be the preferred approach in this setting.

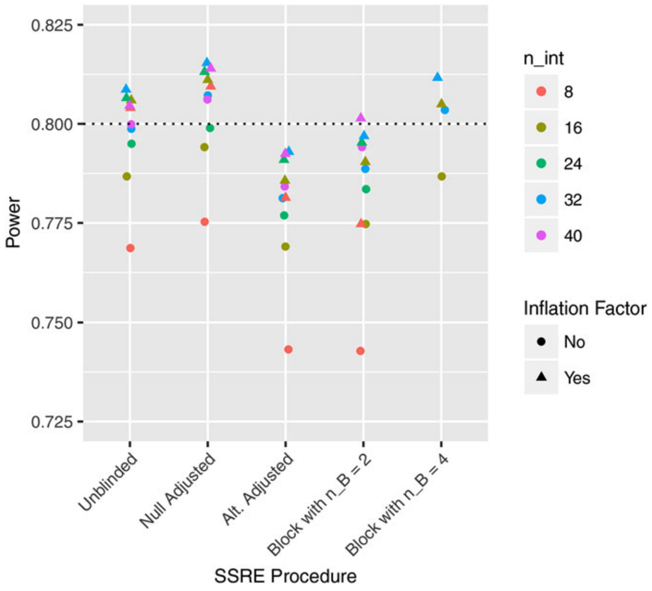

FIGURE 7 The influence of the considered inflation factor upon the power of the re-estimation procedures under the global alternative hypothesis is shown for several values of n_{int}, for Example 1. The dashed line indicates the desired value of the power

Our findings indicate that for each of the reestimation procedures, the choice of δ and the underlying values of σ_e^2 and σ_b^2 often have little effect upon the FWER and power. We may be reassured therefore that the performance of the procedures should often be relatively insensitive to the design parameters. On a similar note, it is important to recognize that one cannot be certain when utilizing these methods that the value of the period effects will not influence the performance of the reestimation procedures. While the final analysis should be asymptotically invariant to period effects, in finite samples it may influence the results of the hypothesis tests. Intuitively though one would not anticipate this effect to be large, nor would one routinely expect large period effects in many settings. In the Supplementary Material, simulations to explore this are presented for the TOMADO example. The results indicate that there is little evidence to suggest the value of the period effects influences the performance of the reestimation procedures. Trialists must be mindful however that this cannot be guaranteed, and should therefore be investigated.

We also considered the utility of a simple sample size inflation factor in ensuring the power reaches the desired level. Ultimately, we demonstrated that this was an effective extension to the basic reestimation procedures. Though the observed inflation to the FWER of our procedures was often small, if more strict control is desired, a crude α-level adjustment procedure can also be utilized. For a particular reestimation scenario, the values of σ_e^2 and σ_b^2, $\sigma_{e,\max}^2$, and $\sigma_{b,\max}^2$ say, which maximize the inflation to the FWER under the global null hypothesis can be determined via a two dimensional search. Then, the significance level used in the analysis of the trial can be adjusted to the α_{adj} that confers a FWER of α for this $\sigma_{e,\max}^2$, $\sigma_{b,\max}^2$ pair, according to further simulations. This may be useful in practice if the inflation is large for a particular trial design scenario of interest.

It is important to note the seemingly inherent advantages and disadvantages of the various reestimation procedures. The adjusted estimator is perhaps the most constrained of those considered; requiring an equal number of patients to be allocated to each sequence for any nonzero adjustment to be reasonable. This is particularly troubling because of the possibility of patient drop-out.

The estimator following block randomisation does not necessitate equal allocation to sequences (though its performance was considered here only when this was the case), but could also fall foul of patient drop-out that would prevent the estimation of the within person variance for each block. It also requires block randomization, and could not be used with a more simple randomization scheme if this was desired. The unblinded estimator of course suffers from none of these problems, but as discussed may be looked upon less favorably by regulators.

Finally, note that in conducting our work we also considered the performance of two reestimation procedures based on methodology for the clustering of longitudinal data (Fraley & Raftery, 2003; Genolini, Alacoque, Sentenac, & Arnauld, 2009). The motivation for this came from the Expectation-Maximisation algorithm approaches of Gould and Shih (1992) for parallel two-arm, and Kieser and Friede (2002) for parallel multiarm, studies. These methods may seem appealing, as they are blinded, under certain assumptions can produce unbiased estimates of the variance parameters, do not require specification of any adjustment, and in theory should be able to more readily handle small amounts of missing data. However, we found that they routinely vastly underestimated the size of within person variance, resulting in substantially lower power than that attained by the other reestimation procedures. Accordingly, especially given the associated concerns about the appropriateness of an Expectation-Maximization algorithm for blinded sample size reestimation (Friede & Kieser, 2002), we would not recommend reestimation be performed based on a clustering-based approach.

In conclusion, following findings for other trial design settings, blinded estimators can be used for sample size reestimation in multitreatment crossover trials. The operating characteristics of any chosen procedure should of course be assessed pretrial through a comprehensive simulation study. But, often, investigators can hope to find that the likelihood of correctly powering their study when there is pretrial uncertainty over the within and between person variances can be enhanced.

ACKNOWLEDGMENTS

The authors would like to thank the two anonymous reviewers whose comments helped to substantially improve the quality of this article. This work was supported by the Medical Research Council [grant number MC_UU_00002/3 to M.J.G. and A.P.M., and grant number MC_UU_00002/6 to J.M.S.W.].

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

ORCID

Michael J. Grayling http://orcid.org/0000-0002-0680-6668

REFERENCES

Bailey, R. A., & Druilhet, P. (2014). Optimal cross-over designs for full interaction models. *The Annals of Statistics*, *42*, 2282–2300.

European Medicines Agency (EMEA)—Committee for Medicinal Products for Human Use (CHMP). (2007). CHMP reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. (Accessed April 24, 2017).

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, *50*, 1096–1121.

Food and Drug Administration (FDA). (2010). Guidance for industry—Adaptive design clinical trials for drugs and biologics. (Accessed April 24, 2017).

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis*. New Jersey: John Wiley & Sons.

Fraley, C., & Raftery, A. E. (2003). Enhanced software for model-based clustering, density estimation, and discriminant analysis: MCLUST. *Journal of Classification*, *20*, 263–286.

Friede, T., & Kieser, M. (2002). On the inappropriateness of an EM algorithm based procedure for blinded sample size re-estimation. *Statistics in Medicine*, *21*, 165–176.

Friede, T., & Kieser, M. (2013). Blinded sample size re-estimation in superiority and noninferiority trials: Bias versus variance in variance estimation. *Pharmaceutical Statistics*, *12*, 141–146.

Genolini, C., Alacoque, X., Sentenac, M., & Arnauld, C. (2009). kml and kml3d: R Packages to Cluster Longitudinal Data. *Journal of Statistical Software*, *65*, 1–34.

Golkowski, D., Friede, T., & Kieser, M. (2014). Blinded sample size re-estimation in crossover bioequivalence trials. *Pharmaceutical Statistics*, *13*, 157–162.

Gould, A., & Shih, W. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics Theory and Methods*, *21*, 2833–2853.

International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). (1998). ICH Hamonised Tripartite Guideline: Statistical Principles for Clinical Trials E9. (Accessed April 24, 2017).

Jensen, K., & Kieser, M. (2010). Blinded sample size recalculation in multicentre trials with normally distributed outcome. *Biometrical Journal*, *52*, 377–399.

Jones, B., & Kenward, M. G. (2014). *Design and analysis of cross-over trials*. Boca Raton, FL: Chapman & Hall.

Kieser, M., & Friede, T. (2002). Blinded sample size re-estimation in multi-armed clinical trials. *Drug Information Journal*, *34*, 455–460.

Kieser, M., & Friede, T. (2003). Simple procedures for blinded sample size adjustment that do not affect the type I error rate. *Statistics in Medicine*, *22*, 3571–3581.

Kieser, M., & Rauch, G. (2015). Two-stage designs for cross-over bioequivalence trials. *Statistics in Medicine*, *34*, 2403–2416.

Lake, S., Kammann, E., Klar, N., & Betensky, R. (2002). Sample size re-estimation in cluster randomization trials. *Statistics in Medicine*, *21*, 1337–1350.

Lui, K. J., & Chang, K. C. (2016). Test equality in binary data for a 4-4 crossover trial under a Latin-square design. *Statistics in Medicine*, *35*, 4110–4123.

Montague, T. H., Potvin, D., Diliberti, C. E., Hauck, W. W., Parr, A. F., & Schuirmann, D. J. (2012). Additional results for "Sequential design approaches for bioequivalence studies with crossover designs. *Pharmaceutical Statistics*, *11*, 8–13.

Potvin, D., Diliberti, C. E., Hauck, W. W., Parr, A. F., Schuirmann, D. J., & Smith, R. A. (2007). Sequential design approaches for bioequivalence studies with crossover designs. *Pharmaceutical Statistics*, *7*, 245–262.

Proschan, M. (2005). Two-stage sample size re-estimation based on a nuisance parameter: A review. *Journal of Biopharmaceutical Statistics*, *15*, 559–574.

Quinnell, T. G., Bennett, M., Jordan, J., Clutterbuck-James, A. L., Davies, M. G., Smith, I. E., ... Sharples, L. D. (2014). A crossover randomised controlled trial of oral mandibular advancement devices for obstructive sleep apnoea-hypopnoea (TOMADO). *Thorax*, *69*, 938–945.

R Core Team. (2016). R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.r-project.org/.

Senn, S. J. (1992). Is the "simple carry-over" model useful? *Statistics in Medicine*, *11*, 715–726.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics*, *24*, 243–258.

Togo, K., & Iwasaki, M. (2011). Sample size re-estimation for survival data in clinical trials with an adaptive design. *Pharmaceutical Statistics*, *10*, 325–331.

van Schie, S., & Moerbeek, M. (2014). Re-estimating sample size in cluster randomised trials with active recruitment within clusters. *Statistics in Medicine*, *33*, 3253–3268.

Wittes, J., & Brittain, E. (1990). The role of internal pilot studies in increasing the eciency of clinical trials. *Statistics in Medicine*, *9*, 65–72.

Xing, B., & Ganju, J. (2005). A method to estimate the variance of an endpoint from an on-going blinded trial. *Statistics in Medicine*, *24*, 1807–1814.

Xu, J., Audet, C., DiLiberti, C. E., Hauck, W. W., Montague, T. H., Parr, A. F., ... Schuirmann, D. J. (2016). Optimal adaptive sequential designs for crossover bioequivalence studies. *Pharmaceutical Statistics*, *15*, 15–27.

Zucker, D. M., & Denne, J. (2002). Sample size redetermination for repeated measures studies. *Biometrics*, *58*, 548–559.

Zucker, D., Wittes, J., Schabenberger, O., & Brittan, E. (1999). Internal pilot studies II: comparison of various procedures. *Statistics in Medicine*, *18*, 3493–3509.

SUPPORTING INFORMATION

Additional Supporting Information including source code to reproduce the results may be found online in the supporting information tab for this article.

How to cite this article: Grayling MJ, Mander AP, Wason JMS. Blinded and unblinded sample size reestimation in crossover trials balanced for period. *Biometrical Journal*. 2018;60:917–933. https://doi.org/10.1002/bimj.201700092