Two-stage adaptive designs for three-treatment bioequivalence studies

Michael J. Grayling[1,2*], Adrian P. Mander[1,3], James M. S. Wason[1,2],

1. Hub for Trials Methodology Research, MRC Biostatistics Unit, Cambridge, UK,

2. Institute of Health & Society, Newcastle University, Newcastle, UK,

3. Centre for Trials Research, Cardiff University, Cardiff, UK.

*Address correspondence to Michael J. Grayling, Institute of Health & Society, Newcastle University, Baddiley-Clark Building, Richardson Road, Newcastle upon Tyne NE2 4AX, UK; Tel: +44(0)191 208 7045; E-mail: michael.grayling@newcastle.ac.uk.

Abstract: Bioequivalence studies are most often conducted as crossover trials, and therefore establishing their required sample size necessitates specification of the within-person variance. Given that this specification is often difficult in practice, there has been great interest in recent years in the use of adaptive designs for bioequivalence trials. However, whilst numerous methods for this have now been presented, their focus has been solely on two-treatment bioequivalence studies. In some instances, it will be desired to incorporate more than a single test and reference formulation into a bioequivalence trial. It would therefore be useful to establish methodology for the design of adaptive multi-treatment bioequivalence trials, in order to acquire the benefits in the two-treatment setting in this more complex situation. Here, we achieve this for three-treatment studies by extending previously proposed designs for two-treatment trials. First, we discuss the additional design considerations that arise when multiple comparisons are made. Next, an extensive simulation study is employed to compare the performance of the proposed procedures. With this, we demonstrate that two-stage designs with desirable statistical operating

characteristics can be readily identified for three-treatment bioequivalence trials.

Keywords: Interim analysis; Re-estimation; Sequential; Unblinded.

1. Introduction

Bioequivalence (BE) studies are conducted for several reasons. For example, they may be carried out when a generic product is proposed to replace an innovator product following patent expiry, when a product's manufacturing process or site changes, or when a product's formulation is altered in some way. In each case, their goal is to indirectly demonstrate the clinical equivalence of several pharmaceutical products by establishing their equivalence in bioavailability (U.S. Food and Drug Administration, 2000).

Precisely, BE refers to the equivalence in rate and extent to which the active ingredient in a drug is absorbed and becomes available to a system (U.S. Food and Drug Administration, 2000). This is usually characterized by the pharmacokinetic parameters area under the concentration time-curve (AUC) and maximum plasma concentration ($C_{\max}$). Accordingly, BE is concluded if the 90% confidence interval of the geometric mean ratio (GMR), for each pharmacokinetic parameter, lies within a specific range, often $[0.8, 1.25]$ (U.S. Food and Drug Administration, 2001; European Medicines Agency Committee for Medicinal Products for Human Use, 2010). There is a long history of methodology proposed to assist in establishing BE, comprehensive overviews of which have been provided by Chow and Liu (2008) and Patterson and Jones (2017). Bioequivalence also remains an active area of research, with interest in recent years fuelled by the highly related problem of developing biosimilars. For example, Kang and Kim (2014), Burdick et al. (2016), and Mielke et al. (2017) each discuss statistical methodology for biosimilar development. In contrast, Mielke et al. (2018) reflect on the clinical development programmes of numerous approved biosimilars, whilst LaVange (2019) notes recent regulatory considerations around biosimilar development, and Tang and Gallo (2019) discuss the important concept of interchangeability.

Returning to the issue of establishing BE, estimating the sample size required for a BE study depends on a variance parameter. Generally, BE studies use crossover designs (Patterson and Jones, 2017), and therefore it is the within-person variance that is required, often re-parameterized as the within-person coefficient of variation (CV). Unfortunately, at the planning stage this variance will typically be subject to great uncertainty. For example, one review of 21 BE studies reported within-subject CVs for the AUC for Ibuprofen of 0.06 to 0.27 (Steinijans et al. 1995). Therefore, with a fixed sample design, there is substantial risk of conducting an under- or over-powered study.

Group-sequential designs (Jennison and Turnbull, 1999) are one possible approach that can provide defence against assuming a larger variance in the sample size calculation than the true value, since the trial is likely to cross a stopping boundary when there is a larger maximum sample size than that truly required. However, because the sample size of each stage is fixed a priori in a group-sequential design, they do not afford any defence against a variance parameter being specified as a smaller value than the truth. Sample size re-estimation designs (e.g., Stein (1945), Gould and Shih (1992), Kieser and Friede (2003)), which allow the planned sample size to be adjusted according to a trials accrued data, are more suited to handling situations where a variance parameter is under-specified. Therefore, if one combines the core features of group-sequential and sample size re-estimation designs, permitting the interim termination of a trial or the adjustment of its required sample size, they can readily guard against both under- and over-powering a study. This type of design is typically given the more generic title adaptive design.

As a result of the issues faced in specifying the variance parameter in BE studies, there has been substantial interest in recent years in methodology for conducting adaptive BE trials. Since BE studies are typically small, the focus has been limited in practice to two-stage designs, even though the principles allow for any number of interim assessments. The first such methods were presented by Potvin et al. (2008) and Montague et al. (2010), who explored, via simulation, the statistical characteristics of a simple unblinded sample size re-estimation design and three two-stage adaptive designs. Following this, further adaptive designs were proposed by Karalis and Macheras (2013), who considered utilising

the interim estimate of the GMR in the re-estimation procedure, and Fuglsang (2014), who presented a design in which two-stages were mandatory. More recently, Zheng et al. (2015) offered an adaptive design that aimed to more accurately control the type-I error-rate, and Xu et al. (2016) described optimized adaptive designs. In addition, several recent articles have sought to provide additional methodology that assists with the strict control of the type-I error-rate in two-stage adaptive BE trials (Maurer et al., 2018; Rasmussen, et al., 2018). Finally, two additional articles have added to this literature by comparing group sequential and fixed sample size designs for bioequivalence trials with highly variable drugs (Knahl et al., 2018) and by providing Bayesian two-stage adaptive designs for bioequivalence (Liu et al., 2019). All such considerations arguably now hold greater importance because of the aforementioned increased interest in methods for establishing the equivalence of biosimilars, for which adaptive designs have been proposed and utilized (Uozumi and Hamada, 2017; Mielke et al., 2018). Ultimately, acknowledging the difficulties in specifying a sample size for a BE trial, regulatory guidelines allow the application of such two-stage designs (European Medicines Agency Committee for Medicinal Products for Human Use, 2010; Health Canada, 2012). Adaptive designs have therefore established themselves as a useful and important tool for establishing treatment BE.

However, currently, the available adaptive BE trial designs are limited to two-treatment two-sequence BE trials. In some settings it may desirable to incorporate more than a single test and reference formulation in to a BE trial (Patterson and Jones, 2006; Zheng et al., 2012). For example, this may occur when multiple possible dosing formulations of a drug are to be tested against some common reference, or when it is desired to confirm a single test formulation is bioequivalent to multiple reference formulations licensed for different markets. This latter scenario has also been a common one in completed biosimilar studies (Mielke et al., 2018). Alternatively, multiplicity may arise in drug-drug interaction studies from desiring to assess more than one interaction, whilst it can also occur in food effect studies when there are multiple drug intake time points, or multiple meal types. In these multi-arm BE trial design settings, the issues around specifying the within-person CV persist. It would therefore be highly advantageous to

establish methodology for determining adaptive two-stage designs for multi-arm BE trials.

Here, we demonstrate how this can be achieved for designs with two test, and a single reference, treatment formulation. Extension to settings with additional test or reference formulations should then be clear. Based on previous proposals for two-arm trials, we specify several possible methods for conducting an adaptive two-stage three-arm design, and then explore their performance using an extensive simulation study. Following Xu et al. (2016), we also describe how our three-arm designs can be optimized to obtain notable efficiency gains. Central to each of our designs is a newly incorporated design parameter, denoted by R, which controls whether a trial is terminated when BE is declared for one of the test formulations. This paper now proceeds by introducing the considered two-stage designs, focusing on the role of the novel parameter R. The formal notation required is then presented along with the hypothesis testing framework. Next, the methodology necessitated by the considered designs is detailed. The simulation study is then described, and its results presented. We conclude with a brief discussion.

2. Methods

2.1. Two-stage designs

We consider five two-stage designs, extending methods A-C of Potvin et al. (2008), and methods E and F of Xu et al. (2016), to a three-treatment setting. We will refer to our designs as A_3, B_3, and so on. Each will desire to control the familywise error-rate to level α, and the type-II error-rate (defined precisely in the next section) to level β. Note that the parameters that must be specified are generally the same as for two-arm designs. However, for methods B_3, C_3, E_3 and F_3, as discussed, an additional parameter that we denote by $R \in \{1, 2\}$ must be provided. This signifies the number of null hypotheses that must be rejected at the interim analysis for the trial to be terminated (i.e., the number of test formulations that must be declared bioequivalent to the reference formulation). Thus, $R = 1$ implies that if any test formulation is declared bioequivalent the trial would be terminated. Whereas, $R = 2$ requires that a decision must be made for each test for-

mulation (i.e., accept or reject BE) for the trial to be stopped. Schematic representations of the methods are given in Figures 1 and 2.

In brief, method A_3 is an unblinded sample size re-estimation procedure, where the first stage data is used to estimate power. If the power is at least that required $(1-\beta)$, the test formulations are assessed for BE and the trial is terminated, otherwise the required sample size is determined and the second stage carried out accordingly.

In contrast, methods B_3 and C_3 are adaptive procedures in which analyses are performed at the end of stage 1 regardless of the estimated power. Specifically, in method B_3, BE is evaluated after stage 1 at a specified level α'. If BE is met for both of the test formulations, the trial is terminated. If BE is met for one of the test formulations and $R = 1$, the trial is also terminated. However, if BE is met for only one test formulation and $R = 2$, or if BE is not met for either of the test formulations, method B_3 evaluates the trial's power at level α'. Power greater than or equal to $1 - \beta$ brings about the termination of the trial, otherwise stage 2 is conducted (with the reference formulation and either one or both of the test formulations present according to the results of the previous significance tests - that is, with the test formulations for which BE has not yet been accepted or rejected) after determining its required sample size. BE is then assessed on completion of stage 2 at level α' for the test formulations remaining in the trial in stage 2. Here, α' is a significance level that should in general be smaller than α to account for multiple testing.

In method C_3, power is evaluated after stage 1 at level α. In the event that power is greater than or equal to $1 - \beta$, BE is assessed for each of the test formulations at level α and the trial is terminated. Alternatively, if power is less than $1 - \beta$, BE is instead evaluated at level α'. If BE is met for both test formulations when $R = 2$, or at least one test formulation in the case $R = 1$, the trial is stopped. Otherwise, stage 2 is conducted using a sample size calculated based on level α' and the number of formulations remaining. Finally, BE is assessed after trial completion at level α' for the test formulations remaining in the trial in stage 2.

Methods E_3 and F_3 extend methods B_3 and C_3 respectively to include the possibility

of declaring retaining particular test formulations to be futile, according to a pre-specified futility bound f. Additionally, they allow the significance levels used after each stage to differ. That is, values α_1 and α_2 are employed instead of simply α'.

2.2. Notation, hypotheses, and analysis

Formally, we consider a randomized BE trial design with three treatments, indexed by $d = 0, 1, 2$, and a maximum of two stages, indexed by $l = 1, 2$. Treatments $d = 1, 2$ are considered test formulations, to be compared to the reference $d = 0$. The log of a single outcome variable (e.g., AUC or $C_{\max}$) will be analyzed using the following linear mixed model

$$y_{ijk} = \mu + \pi_j + \tau_{\mathrm{d}(j,k)} + s_{ik} + \epsilon_{ijk}, \tag{2.1}$$

where

- y_{ijk} is the log of the response for individual i, in period j, on sequence k;
- μ is an intercept term;
- π_j is a fixed effect for period j, with the identifiability constraint $\pi_1 = 0$;
- $\tau_{\mathrm{d}(j,k)}$, $\mathrm{d}(j,k) \in \{0, 1, 2\}$, is the fixed direct treatment effect (the log of the GMR noted earlier) for the treatment administered to an individual in period j, on sequence k, with the identifiability constraint $\tau_0 = 0$;
- $s_{ik} \sim N(0, \sigma_b^2)$ is a random effect for individual i on sequence k;
- $\epsilon_{ijk} \sim N(0, \sigma_e^2)$ is the residual for the response from individual i, in period j, on sequence k. The within-patient CV is then given by $\sqrt{\exp(\sigma_e^2) - 1}$ (Hauschke et al., 1994).

We denote the number of patients recruited in stage $l = 1, 2$ by n_l, and similarly the number of observations accrued in stage $l = 1, 2$ is referred to as o_l. For each of the designs, all three treatments are present in stage 1, and we restrict ourselves to a scenario in which the $n_1 > 0$ patients are allocated treatments using sequences obtained

from a Latin Square (e.g., $k = 1, 2, 3$ corresponds to the sequences $\{0, 1, 2\}$, $\{1, 2, 0\}$, and $\{2, 0, 1\}$). Consequently, $o_1 = 3n_1$. Additionally, for balance, n_1 is always pre-specified as some multiple of three. It is supposed that an interim analysis is conducted after the o_1 observations have been accrued. Then, according to the design's particular specifications, the trial may be terminated (rejecting or declaring BE for the test formulations), or stage 2 will be conducted with the reference formulation and one or more of the test formulations present. That is, in all instances, $n_2 \geq 0$ is determined at interim according to the accrued data. We once more restrict ourselves such that if stage 2 is conducted, regardless of whether one or two of the test formulations have been carried forward, an equal number of patients are allocated to several treatment sequences obtained from a Latin Square. That is, if both test formulations are carried forward to stage 2, then sequences $k = 1, 2, 3$ are again utilized. Otherwise, if only one test formulation is present in stage two, a 2×2 Latin Square is employed (e.g., $k = 4, 5$ correspond to the sequences $\{0, 1\}$ and $\{1, 0\}$ when $d = 1$ is present and $d = 2$ is not). Therefore, when $n_2 > 0$ (i.e., when stage 2 is conducted), $o_2 \in \{2n_2, 3n_2\}$, and n_2 is either divisible by two or three according to the number of treatments present. Otherwise, $o_2 = n_2 = 0$.

Note that following the approach taken for two-treatment BE trials, we class the first period of the second stage ($l = 2$) of a trial as period one, and similarly for its second and third periods. In some situations, there may be reason to alter these to be considered periods four through six, for example when a seasonal effect is anticipated in the trial. Exploration of designs for this setting could be achieved similarly. We consider our approach most sensible, however, given for fixed sample crossover trials the first observation for each individual is typically treated as period one, even if there was a lengthy recruitment rate and other patients have completed several time periods.

To evaluate the two test formulations for BE, the following intersection-union test problems are assessed, using the familiar two one-sided t-tests (TOST) procedure (see, e.g., Schuirmann (1987) for an extensive description)

$$H_{0d} : H_{0d}^{(-)} \cup H_{0d}^{(+)}, \qquad H_{1d} : H_{1d}^{(-)} \cap H_{1d}^{(+)}, \qquad d = 1, 2,$$

where

$$H_{0d}^{(-)} : \tau_d \leq -B, \qquad H_{1d}^{(-)} : \tau_d > -B, \qquad H_{0d}^{(+)} : \tau_d \geq B, \qquad H_{1d}^{(+)} : \tau_d < B, \qquad d = 1, 2,$$

and $B > 0$ is a specified BE margin. Note that corresponding to the confidence interval discussed earlier, $[0.8, 1.25]$, typically $B = -\log(0.8) = \log(1.25)$.

As described above, we assume that we would like each of the two-stage designs to have a familywise error-rate (FWER), the probability of incorrectly rejecting at least one of the null hypotheses H_{01} and H_{02}, of at most α. Additionally, we suppose that without loss of generality we desire power of at least $1 - \beta$ to reject H_{01} when $\tau_1 = \delta \in (-B, B)$. That is, we power the trial to reject a particular null hypothesis, rather than either null hypothesis.

Across the designs, as can be seen in Figures 1 and 2, several common features can be observed. We now proceed to explain how each of these features are carried out mathematically.

Firstly, when designated, the designs test BE at the end of stage l, at level α_*, by performing the following calculations. The following test-statistics are computed to perform the TOST procedure

$$T_{dl}^{(-)} = \frac{\hat{\tau}_{dl} + B}{\sqrt{\text{Var}(\hat{\tau}_{dl})}}, \qquad T_{dl}^{(+)} = \frac{\hat{\tau}_{dl} - B}{\sqrt{\text{Var}(\hat{\tau}_{dl})}}, \qquad d = 1, 2,$$

where $\hat{\tau}_{dl}$ is the REML estimate of τ_d obtained by fitting Equation (2.1) to the accumulated data. Then, they test for the BE of the test formulations present in stage l. That is, if a treatment is dropped at the end of stage 1 to either accept or reject its null hypothesis, they do not test again for whether it is bioequivalent at the end of stage 2. For the tested hypotheses d, H_{0d} is rejected if $T_{dl}^{(-)} > c_l$ and $T_{dl}^{(+)} < -c_l$, for critical boundary c_l. The form of c_l is dependent on the number of test formulations for which BE is to be assessed. Thus, when both formulations are to be tested, employing a Dunnett test as has been previously recommended for BE trials (Hauschke et al., 2007), c_l is the solution of the

following integral

$$\alpha_* = 1 - \Psi\left\{\begin{pmatrix} c_l \\ c_l \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \nu_l\right\}. \tag{2.2}$$

Otherwise

$$\alpha_* = 1 - \Psi\left\{c_l, 1, \nu_l\right\}. \tag{2.3}$$

In the above, $\Psi\{\boldsymbol{c}, \boldsymbol{\Lambda}, \nu\}$ is the cumulative distribution function of a central multivariate t-distribution with covariance matrix $\boldsymbol{\Lambda}$ and degrees of freedom ν, up to vector $\boldsymbol{c}$. The particular form given for $\boldsymbol{\Lambda}$ in Equation (2.2) arises from the well-known covariance structure for the treatment effects in crossover trials with Latin Square treatment allocation (Jones and Kenward, 2014). Moreover, we take ν_l to be the degrees of freedom for a corresponding balanced multi-level ANOVA design. Therefore

$$\nu_l = \sum_{m=1}^{l} (o_m - n_m) - 4. \tag{2.4}$$

Next, define the following function

$$P(\sigma_e^2, n, c, \nu) = \Psi\left\{(B - \delta)\sqrt{\frac{n}{2\sigma_e^2}} - c, 1, \nu\right\} - \Psi\left\{(-B - \delta)\sqrt{\frac{n}{2\sigma_e^2}} + c, 1, \nu\right\}.$$

With this, when required by a design, the power to reject each of the H_{0d} at the end of stage 1 can be estimated by $P(\hat{\sigma}_e^2, n_1, c_1, \nu_1)$, with $\hat{\sigma}_e^2$ the REML estimate of the within-person variance, and c_1 and ν_1 determined using Equations (2.2) and (2.4) respectively.

Furthermore, when a design must estimate the sample size needed for the second stage to attain a specified power, the result proven in the Appendix can be used. It is shown that whether or not both of the test formulations are present in stage 2, provided there is balance in both stages, the variance of the maximum likelihood estimator of τ_d, for those d present in stage 2, is

$$\mathrm{Var}(\hat{\tau}_{d2}) = \frac{2\sigma_e^2}{n_1 + n_2}. \tag{2.5}$$

Thus a one-dimensional search can be used to identify the minimal n_2, divisible by the number of formulations to be present in the second stage, such that $P(\hat{\sigma}_e^2, n_1 + n_2, c_2, \nu_2) \geq$

$1-\beta$. In this, ν_2 and c_2 should be determined using Equation (2.4) and either Equation (2.2) or (2.3), respectively. Note that to make the designs more useful in practice, bounds could also be placed on the value of n_2, so that $n_{\min} \leq n_1 + n_2 \leq n_{\max}$. However, we will not utilize such an approach here.

Finally, methods E_3 and F_3 also incorporate a futility rule after stage 1. With this, if $T_{d1}^{(-)} < f$ and $T_{d1}^{(+)} > -f$ for specified f, continuing to assess treatment d for BE is deemed to be futile. It is therefore dropped from the trial with the conclusion test formulation d is not bioequivalent to the reference formulation. In the case where this is the last test formulation for which a decision has not been made, this brings about the termination of the entire trial.

2.3. Simulation study

To assess the performance of the considered two-stage designs, the results of a large simulation study are presented. Throughout we set $B = \log(1.25)$, $\alpha = 0.05$, and $\beta = 0.2$. Finally, as in Potvin et al. (2008), all power calculations are performed with $\delta = \log(100/95)$, to allow some difference from reference.

This leaves n_1, R, CV, $\boldsymbol{\tau} = (\tau_1, \tau_2)^\top$, α', α_1, α_2 and f unspecified, the influence of which will be explored. Precisely, the empirical rejection rate, average required sample size (AVN) and average number of observations (AVO) required by each method, for possible choices of these parameters, will be estimated via 100,000 replicate trial simulations. Note that we simulate individual-period level data (i.e., the outcomes y_{ijk} are simulated). This is in contrast to previously presented two-treatment results, which for their AB/BA crossover design framework simulated patient differences in response, and used basic estimators to compute test statistics. We use this method as for our three-treatment framework with the mixed model Equation (2.1) for analysis, we are unaware of any results that suggest a similar basic-estimator approach would allow for the accurate estimation of rejection rates, particularly given the requirement to analyse datasets that consist of outcomes from patients who have received different numbers of formulations. Moreover, in simulating data, we set $\mu_0 = \pi_2 = \pi_3 = 0$ and $\sigma_b^2 = 2\sigma_e^2$ in all instances.

Given that the distribution of the estimators of $\boldsymbol{\tau}$ and σ_e^2 are asymptotically invariant to these choices (Jiang, 1996), this should have minimal impact upon the presented results.

We consider three particular choices for $\boldsymbol{\tau}$: $\boldsymbol{\tau} = (B, B)^\top$ (the 'null treatment effects scenario'), $\boldsymbol{\tau} = (\delta, B)^\top$ (the 'mixed treatment effects scenario'), and $\boldsymbol{\tau} = (\delta, \delta)^\top$ (the 'alternative treatment effects scenario'). For the null treatment effects scenario, we estimate the FWER as the proportion of times H_{01} or H_{02} were rejected. In contrast, for the mixed treatment effects scenario, we estimate the FWER instead as the proportion of times H_{02} is rejected (as H_{01} is in this case bioequivalent). Similarly, for the mixed and alternative treatment effects, we estimate power as the proportion of times in which H_{01} is rejected.

Programmes to repeat our analyses are available in Supplementary File 1 and from https://github.com/mjg211/article_code.

3. Results

3.1. Familywise error-rate, power, average required sample size and number of observations

Here, we present the results of our simulation study, in which $n_1 \in \{12, 24, 36\}$, $CV \in \{0.1, 0.2, 0.3.0.4\}$, and $R \in \{1, 2\}$. Precisely, Tables 1-3 exhibit our findings for the three considered values of $\boldsymbol{\tau}$. In methods B_3 and C_3, as in Potvin et al. (2008), we set $\alpha' = 0.0294$. Similarly, in methods E_3 and F_3, we take $\alpha_1 = \alpha_2 = 0.0294$, and $f = 0$.

Under the null treatment effects scenario, for each method, we see that there is no clear pattern as to how the FWER is affected by a change in either n_1, R, or the CV, when all other design parameters are held constant (Table 1). It is also difficult to argue that any of the methods is the most conservative, though method A_3 can readily be claimed to be the least conservative (exhibiting a maximum FWER of 0.0564, and failing to control to the desired level in 8 of 12 considered scenarios). Importantly, methods B_3 and E_3 control the FWER to the nominal level across all considered values for n_1, CV, and R. Whilst this is not true for methods C_3 and F_3, the inflation in these designs is typically small. Precisely, it is maximized for method C_3 when $n_1 = 12$, $CV = 0.2$, and $R = 2$ at

0.0507, and it is maximized for method C_3 when $n_1 = 12$, $CV = 0.1$, and $R = 2$ at 0.0512. Allowing for small Monte Carlo errors, these designs thus also approximately achieve the desired level of control across all considered null treatment effect scenarios. Additionally, as one may anticipate, for the mixed treatment effects scenario the fact that only one null hypothesis is true means the designs typically control the FWER conservatively.

From Tables 2 and 3, it is evident that increasing n_1 typically results in an increase in power. In addition, in the alternative treatment effects scenario, $R = 2$ provides increased power relative to $R = 1$. This is not the case for the mixed treatment effects scenario, however, as in this instance it is unlikely BE would be declared for formulation 2. Moreover, by contrasting Tables 2 and 3, it is evident that power is larger in the mixed treatment effect scenario when $R = 1$, but as we would anticipate it is similar for the mixed and alternative treatment effect scenarios when $R = 2$. Note that none of the methods attain the desired power across the full range of considered parameter values.

Interestingly, whilst increasing the CV typically reduces power, there are several instances in which power is smaller for $CV = 0.4$ than $CV = 0.3$. This is likely a consequence of the fact that interim stopping rules are less likely to be utilized when $CV = 0.4$ because of decreased power relative to $CV = 0.3$.

From Table 1, for the null treatment effects scenario there is also evidence to suggest that increasing the value of n_1 in general comes at a cost to the AVN or AVO for methods E_3 and F_3, though this result is obscured for $CV \in \{0.1, 0.2\}$ by the fact that the requisite sample size of a single-stage design (see below) is smaller than several of the considered values for n_1. However, this finding is not present in either the mixed or alternative treatment effect scenarios.

Similarly, in some scenarios $R = 2$ increases the AVN and AVO as compared to $R = 1$. This result is most clear when (i) $CV \in \{0.3, 0.4\}$, (ii) $\boldsymbol{\tau}$ is either the mixed and alternative treatment effect scenarios, and (iii) when n_1 is large. This is a consequence of the fact that (i) for small CV values stage 2 is not typically required, (ii) that it is unlikely BE will be declared at the interim analysis for the null treatment effect scenario, and (iii) that for larger n_1 values we would in general have the power required to stop at the interim

analysis. As an example of the magnitude of the effect, when $n_1 = 36$ and $CV = 0.4$, in the alternative treatment effects scenario method F_3 has an AVO of 228.37 when $R = 1$, but this increases by 11.7% to 255.15 with $R = 2$.

Note that $\alpha' = \alpha_1 = \alpha_2$ means that in general methods E_3 and F_3 have lower rejection probabilities and smaller AVN and AVO values than their corresponding method B_3 and C_3 designs. The differences in performance between methods B_3 and E_3, and methods C_3 and F_3, are most evident in the null and mixed treatment effect scenarios, when the futility stopping rules of methods E_3 and F_3 are more likely to be employed. However, there are also some cases of large differences in the alternative treatment effects scenario. For example, when $n_1 = 12$, $CV = 0.4$, and $R = 1$, method B_3 has power of 0.7695 and an AVN of 93.15, whilst method E_3 has power of only 0.6559 and an AVN of 88.45.

Finally, it is informative to contrast the performance of the adaptive designs to correctly specified single-stage designs. To this end, note that the minimal sample sizes, divisible by three, which provide power of 0.8 are nine, 24, 48, and 81, for CVs of 0.1, 0.2, 0.3, and 0.4 respectively. Thus, the adaptive designs cannot compare favourably in the settings with $CV = 0.1$. However, for $CV = 0.2$, when $n_1 \in \{12, 24\}$, the maximal value of the AVN across all methods and scenarios presented in Tables 1-3 is 27.52, an increase of 14.7% over the 24 patients required by the single-stage design. Similarly, when $CV = 0.3$, the maximal value of the AVN across all methods and scenarios when $n_1 \in \{12, 24, 36\}$, is 55.31, an increase of 15.2% over the 48 patients required by the single-stage design. Finally, for $CV = 0.4$ the corresponding maximal value is 93.39, an increase of 15.3% over the 81 patients required by the single-stage design. Thus, the cost of the flexibility introduced by the adaptive designs appears in general to be at most 15%.

3.2. Design optimisation

Whilst the performance of several of the methods seems to be highly desirable based on the results in Tables 1-3, there remain instances where each does not attain the desired power, or does not accurately control the FWER. To address this in the two-treatment setting, Xu et al. (2016) presented methods E and F as those with optimized values for

n_1, α_1, α_2, and f. Specifically, the values that would confer the desired type-I error-rate and power across a range of possible CV values with the smallest weighted sum of average sample sizes. We now consider a similar approach in our three treatment setting, in order to demonstrate how the adaptive designs can be made as efficient as possible. This is particularly relevant given that it was recently pointed out that the values of α_1 and α_2 proposed by Potvin et al. (2008) have no theoretical basis, and are not actually required for control of the FWER (Kieser and Rauch, 2015).

First, one chooses a discrete set of CV values, $\mathscr{CV} = \{CV_1, \ldots, CV_{|\mathscr{CV}|}\}$ (e.g., $\mathscr{CV} = \{0.1, 0.2, 0.3\}$). Ideally, this set should contain as many values as possible, with the minimal and maximal values chosen according to the range that may be realistically anticipated for the formulations under investigation. Next, a discrete set of values for $\boldsymbol{\tau}$, $\mathscr{I}_{\text{FWER}} = \{\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_{|\mathscr{I}_{\text{FWER}}|}\}$ is chosen, such that the empirical FWER is required to be below α when $\{\boldsymbol{\tau}, CV\} \in \mathscr{I}_{\text{FWER}} \times \mathscr{CV}$. Similarly, a set $\mathscr{I}_{\text{power}} = \{\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_{|\mathscr{I}_{\text{power}}|}\}$ is nominated, such that the empirical power is required to be above $1 - \beta$ when $\{\boldsymbol{\tau}, CV\} \in \mathscr{I}_{\text{power}} \times \mathscr{CV}$. All that is then required is an optimality criterion that possible choices for the design parameters can be evaluated upon. As stated, Xu et al. (2016) considered a weighted sum of AVNs for this. We modify its specification for our three-formulation setting, giving

$$Cost(n_1, \alpha_1, \alpha_2, f) = \frac{1}{|\mathscr{I}||\mathscr{CV}|} \sum_{i=1}^{|\mathscr{I}|} \sum_{j=1}^{|\mathscr{CV}|} Cost(\boldsymbol{\tau}_i, CV_j | n_1, \alpha_1, \alpha_2, f),$$

for

$$Cost(\boldsymbol{\tau}_i, CV_j | n_1, \alpha_1, \alpha_2, f) = \begin{cases} (AVN_{\{\boldsymbol{\tau}_i, CV_j\}} - n_{\{0.95, CV_j\}})^2 & \text{if } AVN_{\{\boldsymbol{\tau}_i, CV_j\}} - n_{\{0.95, CV_j\}} > 1, \\ (AVN_{\{\boldsymbol{\tau}_i, CV_j\}} - n_{\{0.95, CV_j\}}) & \text{if } AVN_{\{\boldsymbol{\tau}_i, CV_j\}} - n_{\{0.95, CV_j\}} \leq 1, \end{cases}$$

where $\mathscr{I} = \{\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_{|\mathscr{I}|}\}$ is a set of values for $\boldsymbol{\tau}$ at which the AVN will be taken in to consideration, and $n_{\{0.95, CV\}}$ is the sample size required by a single-stage design for a GMR of 0.95, when the CV is CV. Thus, the cost function proposed by Xu et al. (2016) penalizes designs that have an average sample size greater than a corresponding single-

stage design more heavily than it rewards those that are more efficient than a single-stage study. This is to prioritize minimizing the possible inflation of the required sample size. Note it is important to recall that the AVN is dependent on n_1, α_1, α_2, f, but this is omitted here for brevity.

To identify the optimal choices for the design parameters, ideally a numerical optimisation routine would be employed. However, this is difficult in practice because the search space has a discrete part (n_1) and a continuous part (α_1, α_2, f). Furthermore, there is no guarantee that the value of the optimality criteria will be smooth in the search parameters. A global stochastic search procedure could address both of these issues. However, the evaluation of a single set of design parameters is computationally intensive given the requirement for simulation, and thus searching over thousands of designs as would be required for convergence in such an approach will not typically be realistic. A more tractable solution is to find a near-optimal design. This can be achieved by designating sets of values for the four parameters, N_1, A_1, A_2, and F, say, and then evaluating the optimality criteria for $\{n_1, \alpha_1, \alpha_2, f\} \in \mathrm{N}_1 \times \mathrm{A}_1 \times \mathrm{A}_2 \times \mathrm{F}$, with $|\mathrm{N}_1 \times \mathrm{A}_1 \times \mathrm{A}_2 \times \mathrm{F}|$ small enough that this grid search is possible. The choice of parameters that minimizes the optimality criteria subject to the designated constraints is then utilized.

We demonstrate our suggested approach for Method E_3 with $R = 2$ only. As a simple example, we consider $\mathscr{CV} = \{0.2, 0.3\}$, $\mathscr{I}_{\mathrm{FWER}} = \{(B, B)^\top\}$, and $\mathscr{I}_{\mathrm{power}} = \{(\delta, \delta)^\top\}$, with $\mathscr{I} = \{(B, B)^\top, (\delta, \delta)^\top\}$ in the modified optimality criteria of Xu et al. (2016). In addition, we take $\mathrm{N}_1 = \{12, 24, 36\}$, $\mathrm{A}_1 = \mathrm{A}_2 = \{0.01, 0.02, 0.03, 0.04\}$, and $\mathrm{F} = \{-0.5, 0, 0.5\}$. This implies $|\mathrm{N}_1 \times \mathrm{A}_1 \times \mathrm{A}_2 \times \mathrm{F}| = 144$. Whilst this is a relatively large number of designs, in practice when designing one of these trials it would be advisable to consider larger sets if the available computational resources allow for this. Note also that the values within N_1 should be a multiple of three by our previous restrictions on balance. Furthermore, it is logical for the values within A_1 and A_2 to be smaller than α, and those in F to be near zero such that the futility rule is likely to be utilized but not be overly detrimental to the trial's power.

In Table 4 we present the optimal design determined via our search, along with those

designs presented in the previous section for method E_3 with $R = 2$. We can see that from our simple search it was possible to determine a design that reduced the value of the cost function substantially compared to the previously presented designs. Specifically, the best performing of the designs from earlier that controlled the FWER, whilst simultaneously achieving the desired power, was that with $n_1 = 24$, which had a value for the cost function of 107.25. The optimized design, which takes $\alpha_1 = \alpha_2 = 0.03$ and $f = 0.5$, has a cost function value of 43.84; a reduction of 59.1% compared to the former design. Examining Table 4 in greater detail, it is clear that much of this reduction in cost comes from a reduced AVN under the null treatment effect scenario when $CV = 0.3$. In this setting, the aforementioned non-optimal design has an AVN of 44.69, whilst the optimal design has an AVN of 37.62, a reduction of 15.8%.

An important final consideration is to examine the performance of the optimal design in scenarios that were not incorporated in to the optimization problem. Thus, in Table 5, we present the operating characteristics of the optimal design from Table 4 when $CV \in \{0.1, 0.2, 0.3, 0.4\}$ and $\boldsymbol{\tau}$ is given by either the null or alternative treatment effects scenarios. It can be seen that even though neither were accounted for in the optimization routine, the optimal design controls the FWER to the desired level for both $CV = 0.1$ and $CV = 0.4$. However, as would be expected, power drops to 0.6967 for $CV = 0.4$ under the alternative treatment effects scenario. This highlights the need to account for as wide a range of CV values as are considered likely for the planned study.

4. Discussion

Difficulties in specifying the within-person CV pre-trial have led to a wealth of literature on sequential BE study designs. These designs have been demonstrated to have highly favourable operating characteristics. However, they are all only relevant to two-treatment trials. Given it is much more efficient to incorporate more than two treatments into a BE trial than carry out several two-arm trials (Zheng et al., 2012), here we extended several previously proposed methods to three-treatment BE trials.

After having discussed the various additional design considerations that arise in this

setting, we were able to demonstrate through a simulation study that the considered two-stage designs, for the examined design parameters, frequently carried a FWER under the null treatment effects scenario of close to the nominal level, and attained the desired power under the alternative treatment effects scenario.

To improve efficiency, we next discussed how the chosen designs could be optimized, considering a particular design scenario for method E_3. It was evident that choosing values for the design parameters should ideally be done with care, as the optimal design may provide noteworthy reductions to the examined AVN and AVO.

Thus in all, it is evident that efficient two-stage designs can be identified for three-treatment BE studies. However, it is difficult to prescribe a single method that should be preferred. Nonetheless, method A_3 should arguably be avoided without refinement to the value of α given it displayed inflation of the FWER in most instances. In addition, given that methods B_3 and C_3 are essentially special cases of E_3 and F_3, the choice is fundamentally between E_3 and F_3. Further research with additional values of CV, n_1, and $\boldsymbol{\tau}$, may be useful for elucidating which of these methods should in general be preferred. However, it is hard to imagine that one will out-perform the other in most considered scenarios. We therefore advise that a method be chosen based upon a simulation study that focuses on values of the design parameters that are believed to be reasonable for the planned study. Smaller values of n_1 may also be advantageous to improve performance for small values of the CV. In addition, design optimisation should be employed to maximize trial efficiency.

It is informative to discuss potential reasons why one may choose a particular value for the important parameter R. We were able to observe that, as would be expected, the choice of the value of R was particularly important to the power under the alternative treatment effects scenario. Therefore, in practice, the decision over the appropriate value for R should be made based on the trial's objectives. Principally, in settings in which it is desired simply to identify a bioequivalent test formulation, and no preference is held for which test formulation is bioequivalent (for example, because there are no cost differences between the test formulations), then one would likely choose $R = 1$. For, the observed

scenarios in which $R = 2$ increases the AVN and AVO implies $R = 1$ would be preferable to minimize the cost of the trial. However, if it is desired to identify all bioequivalent test formulations, $R = 2$ should be preferred to increase power (Table 3).

It is important to note the limitations of our work. Firstly, whilst it seems intuitively reasonable that the maximal value of the FWER would occur under the considered null treatment effects, there is no available theoretical result to confirm this. Consequently, when using one of these designs for a real trial it would be important to explore the FWER for a range of possible true treatment effects. Furthermore, as with all simulation studies, our findings are only relevant to the considered scenarios. As for the original two-stage two-treatment BE trial designs, there are also issues with employing these designs in practice. Namely, the inclusion of an interim analysis necessitates the accrual of the data for the first n_1 patients before the second stage can begin. This means that the length of these trials could be longer than a corresponding single-stage design. Furthermore, as is similarly discussed in relation to the problem of demonstrating interchangeability in biosimilar development (Tang and Gallo, 2019), the increased number of time periods may lead to a higher drop-out rate, and thus more missing data to contend with, due to the longer study duration burden on participants.

Moreover, we employed a linear mixed model for data analysis, as has been recommended by several regulatory agencies (U.S. Food and Drug Administration, 2001; Health Canada, 2012). For maximum efficiency, we also utilized all available data at each analysis point. In some trial design scenarios, these designations may not be appropriate and alteration to our methods would be required. In particular, the European Medicines Agency has recommended analysis for multi-formulation BE studies to be performed using separate ANOVA tests with the data relevant only to the comparison under consideration (European Medicines Agency Committee for Medicinal Products for Human Use, 2010).

Finally, we restricted ourselves to a situation in which treatment sequences where allocated using a Latin Square. It is worth noting however that allowance for other possible sequence choices for treatment allocation could be made by amending Equation (2) and Equations (4-6) to take in to account the change in the covariance structure of

the treatment effect estimators.

Several possible extensions to our procedures now present themselves. Primarily, designs that utilize the interim estimate of the treatment effect for determination of the second stage sample size have been discussed in the two-treatment setting (Karalis and Macheras, 2013). Such extension would appear useful to three-treatment trials, but a decision would then need to be made on how the multiple estimated treatment effects would be used. For example, the smallest of the estimated treatment effects could be used to minimize the required sample size, but this would have potential implications on power for the other test formulation.

Similarly, we also assume that the trial specifies the same value of δ for each of the test formulations. In some instances, it may desired to relax this assumption, perhaps based upon prior knowledge that is held about the various formulations. Such an extension could readily be added to our designs, in order to allow power to be evaluated for each of the test formulations independently. However, this would increase the complexity of the underlying method, as scenarios in which power of $1 - \beta$ is achieved for one test formulation, but not the other, would necessitate a new set of rules on how to proceed.

Furthermore, in some settings, for example in food effect studies, it may be required that both tests be successful. With this, a modification to the definition of power would be required, and the advantages and disadvantages of requiring BE to be declared for both test formulations at the same time point would need to be considered, as in group-sequential methods for co-primary endpoints (Hamasaki et al., 2016).

Finally, there may be scenarios in which different variabilities may be anticipated for the two test formulations. As above, this could be accounted for in a simulation study, but the likely substantially differing values of power for the test formulations would then need to be accounted for carefully.

In conclusion, two-stage designs for three-treatment BE studies can be determined that confer a FWER below the nominal level when both test formulations are not bioequivalent, and provide the desired power when they are bioequivalent. An optimisation procedure can be employed to improve the efficiency of these designs. Consequently, when

it is desired to incorporate two test formulations into a BE trial, and there is pre-trial uncertainty over the required sample size, researchers should consider employing a two-stage design.

Acknowledgements

This work was supported by the Wellcome Trust [grant number 099770/Z/12/Z to M.J.G.] and 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.].

References

Burdick, R.K., Thomas, N., and Cheng, A. (2016), "Statistical Considerations in Demonstrating CMC Analytical Similarity for a Biosimilar Product", *Statistics in Biopharmaceutical Research*, 9, 249-257.

Chow, S.C. and Liu, J.P. (2008), *Design and Analysis of Bioavailability and Bioequivalence Studies*, New York: CRC Press.

European Medicines Agency Committee for Medicinal Products for Human Use (2010), "Guideline on the investigation of bioequivalence", http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/01/WC500070039.pdf, accessed 13 November 2018.

Fuglsang, A. (2014), "A sequential bioequivalence design with a potential ethical advantage", *AAPS Journal*, 16, 373-378.

Gould, A. and 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.

Hamasaki, T., Asakura, K., Evans, S.R., and Ochiai, T. (2016), *Group-Sequential Clinical Trials with Multiple Co-Objectives*, Springer.

Hauschke, D., Steinijans, V.W., Diletti, E., Schall, R., Luus, H.G., Elze, M., and Blume, H. (1994), "Presentation of the intrasubject coefficient of variation for sample size planning in bioequivalence studies", *International Journal of Clinical Pharmacology and Therapeutics*, 32, 376-378.

Hauschke, D., Steinijans, V., and Pigeot, I. (2007), *Bioequivalence Studies in Drug Development: Methods and Applications*, England: Wiley.

Health Canada (2012), "Guidance document. Conduct and analysis of comparative bioavailability studies", http://www.hc-sc.gc.ca/dhp-mps/alt_formats/pdf/prodpharma/applic-demande/guideld/ bio/gd_cbs_ebc_ld-eng.pdf, accessed 13 November 2018.

Jennison, C. and Turnbull, B.W. (1999), *Group sequential methods with applications to clinical trials*, Boca Raton: CRC Press.

Jiang, J. (1996), "REML Estimation: Asymptotic Behavior and Related Topics", *Annals of Statistics*, 24, 255-286.

Jones, B. and Kenward, M.G. (2014), *Design and analysis of cross-over trials*, Boca Raton: CRC Press.

Kang, S.H. and Kim, Y. (2014), "Sample size calculations for the development of biosimilar products", *Journal of Biopharmaceutical Statistics*, 24, 1215-1224.

Karalis, V., and Macheras, P. (2013), "An insight into the properties of a two-stage design in bioequivalence studies", *Pharmaceutical Research*, 30, 1824-1835.

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

Knahl, S.I.E., Lang, B., Fleischer, F., and Kieser, M. (2018), "A comparison of group sequential and fixed sample size designs for bioequivalence trials with highly variable

drugs", *European Journal of Clinical Pharmacology*, 74, 549.

LaVange, L.M. (2019), "Statistics at FDA: Reflections on the Past Six Years", *Statistics in Biopharmaceutical Research*, 11, 1-12.

Liu, S., Gao, J., Zheng, Y., Huang, L., and Yan, F. (2019), "Bayesian Two-Stage Adaptive Design in Bioequivalence", *The International Journal of Biostatistics*, DOI:10.1515/ijb-2018-0105.

Maurer, W., Jones, B., and Chen, Y. (2018), "Controlling the type I error rate in twostage sequential adaptive designs when testing for average bioequivalence", *Statistics in Medicine*, 37, 1587-1607.

Mielke, J., Jones, B., Jilma, B., and Koenig, F. (2017), "Sample Size for Multiple Hypothesis Testing in Biosimilar Development", *Statistics in Biopharmaceutical Research*, 10, 39-49.

Mielke, J., Jilma, B., Jones, B., and Koenig, F. (2018), "An update on the clinical evidence that supports biosimilar approvals in Europe", *British Journal of Clinical Pharmacology*, 84.

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

Patterson, S. and Jones, B. (2017), *Bioequivalence and Statistics in Clinical Pharmacology*, New York: CRC Press.

Potvin, D., DiLiberti, C.E., Hauck, W.W., Parr, A.F., Schuirmann, D.J., and Smith, R.A. (2008), "Sequential design approaches for bioequivalence studies with crossover designs", *Pharmaceutical Statistics*, 7, 245-262.

Rsamussen, H.E., Ma, R., and Wang, J.J. (2018), "Controlling type 1 error rate for sequential, bioequivalence studies with crossover designs", *Pharmaceutical Statistics*, DOI:10.1002/pst.1911.

Schuirmann, D.J. (1987), "A comparison of the Two One-Sided Tests Procedure and the Power Approach for assessing the equivalence of average bioavailability", *Journal of Pharmacokinetics and Biopharmaceutics*, 15, 657-680.

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.

Steinijans, V.W., Sauter, R., Hauschke, D., Diletti, E., Schall, R., Luus, H.G., Elze, M., Blume, H., Hoffmann, C., Franke, G., and Siegmund, W. (1995), "Reference tables for the intrasubject coefficient of variation in bioequivalence studies", *International Journal of Clinical Pharmacology and Therapeutics*, 33, 427-430.

Tang, D. and Gallo, P. (2019), "Discussion on Interchangeability and Adaptation in Biosimilar Development", *Statistics in Biopharmaceutical Research*, 11, 79-84.

U.S. Food and Drug Administration (2000), "Guidance for industry. Bioavailability and bioequivalence studies for orally administered drug products general considerations", http://www.fda.gov/downloads/Drugs/.../Guidances/ucm201790.pdf, accessed 13 November 2018.

U.S. Food and Drug Administration (2001), "Guidance for industry. Statistical approaches to establishing bioequivalence", http://www.fda.gov/downloads/Drugs/Guidances/ucm070244.pdf, accessed 13 November 2018.

Uozumi, R. and Hamada, C. (2017), "Adaptive Seamless Design for Establishing Pharmacokinetic and Efficacy Equivalence in Developing Biosimilars", *Therapeutic Innovation & Regulatory Science*, 51, 761-769.

Wason, J.M. and Jaki, T. (2012), "Optimal design of multi-arm multi-stage trials", *Statistics in Medicine*, 31, 4269-4279.

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

Zheng, C., Wang, J., and Zhao, L. (2012), "Testing bioequivalence for multiple formulations with power and sample size calculations". *Pharmaceutical Statistics*, 11, 334-341.

Zheng, C., Zhao, L., and Wang, J. (2015), "Modifications of sequential designs in bioequivalence trials", *Pharmaceutical Statistics*, 14:180-188.

Appendix

In this section we expand on a property about the variance of the treatment effect estimators claimed in the main paper. Namely, suppose that n_1 patients are recruited in stage 1 with $\text{mod}(n_1, 3) = 0$. Then, n_2 patients are recruited in stage 2 with $\text{mod}(n_2, r) = 0$, where $r = 3$ if both test formulations are present in stage 2, and $r = 2$ if only one test formulation is present. It was claimed that in this case

$$\text{Var}(\hat{\tau}_{d2}) = \frac{2\sigma_e^2}{n_1 + n_2},$$

for those $d \in \{1, 2\}$ present in stage 2, where $\hat{\tau}_{d2}$ is the maximum likelihood estimate of τ_d. This result is well known for the case where both test formulations are present in stage 2. In this case

$$\mathbf{Cov}(\hat{\boldsymbol{\tau}}_2, \hat{\boldsymbol{\tau}}_2) = \frac{2\sigma_e^2}{n_1 + n_2} \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

for $\hat{\boldsymbol{\tau}}_2 = (\hat{\tau}_{12}, \hat{\tau}_{22})^\top$ (Jones and Kenward, 2014). This is the reason for the stated form of $\boldsymbol{\Lambda}$ in Equation (2.2).

It remains to prove this result for the case where a single test formulation, along with the reference formulation, is present in stage 2. Without loss of generality suppose that this test formulation is treatment $d = 1$. Denote the Latin Square sequences used for treatment allocation in stage 1 when the three formulations are present by $k = 1, 2, 3$, and the sequences used in stage 2 by $k = 4, 5$. Now, cast our linear mixed model Equation (2.1) in the form

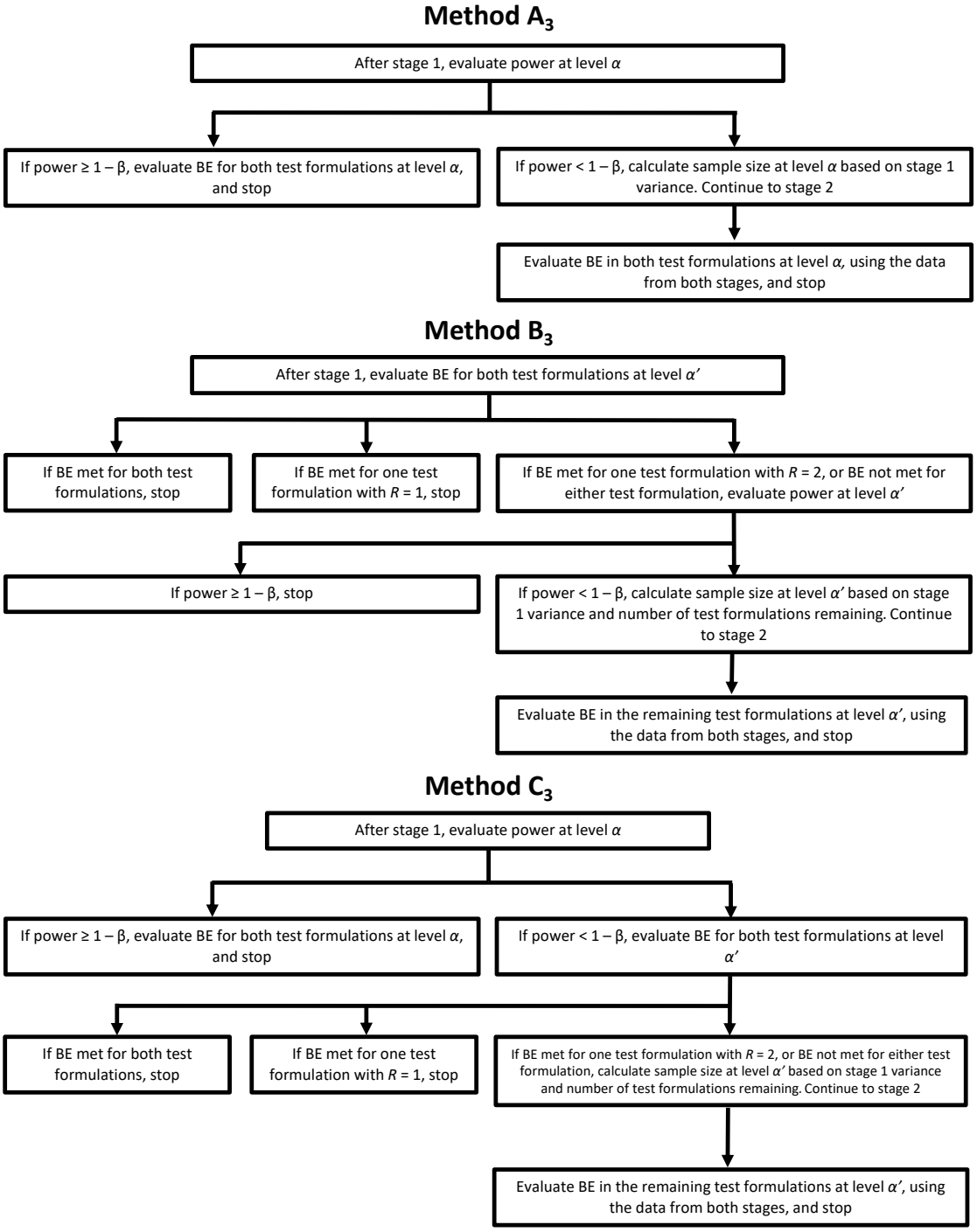

Figure 1: Flow diagrams demonstrating the steps of methods A_3, B_3, and C_3.

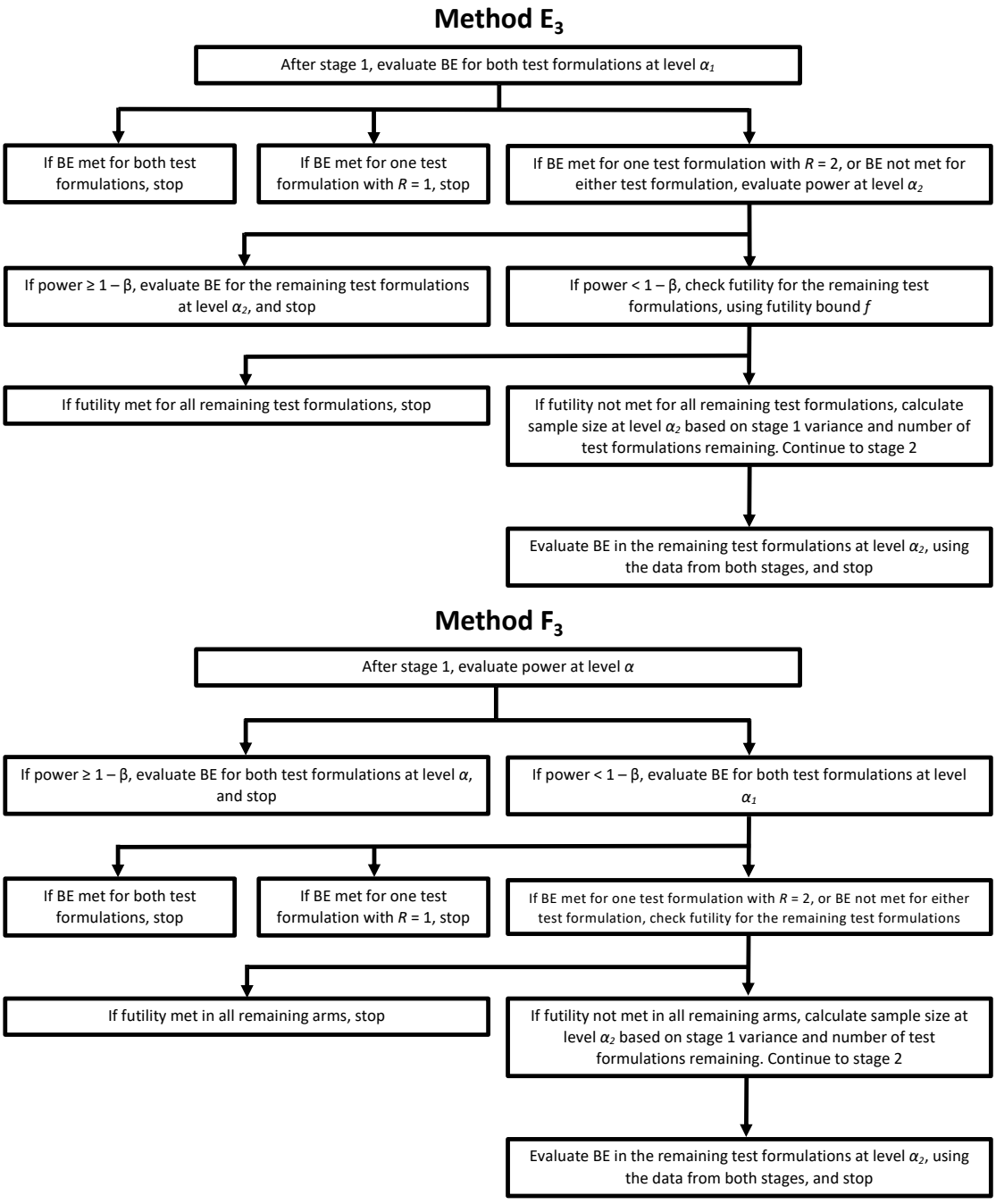

Figure 2: Flow diagrams demonstrating the steps of methods E_3 and F_3.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta} = (\mu, \pi_2, \pi_3, \tau_1, \tau_2)^\top$. Then, the maximum likelihood estimator of the fixed effects $\boldsymbol{\beta}$ at the end of stage 2 is

$$\hat{\boldsymbol{\beta}}_2 = (\hat{\mu}_2, \hat{\pi}_{22}, \hat{\pi}_{32}, \hat{\tau}_{12}, \hat{\tau}_{22})^\top = (\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y},$$

with $\boldsymbol{\Sigma} = \boldsymbol{Z}\mathbf{Cov}(\boldsymbol{b}, \boldsymbol{b})\boldsymbol{Z}^\top + \mathbf{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})$. We then have

$$\mathbf{Cov}(\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_2) = (\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1}.$$

See, for example, Fitzmaurice et al. (2014) for further details.

In our case, $\mathbf{Cov}(\boldsymbol{u}, \boldsymbol{u}) = \sigma_b^2 \boldsymbol{I}_{n_1+n_2}$ and $\mathbf{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}) = \sigma_e^2 \boldsymbol{I}_{o_1+o_2} = \sigma_e^2 \boldsymbol{I}_{3n_1+2n_2}$, where $\boldsymbol{I}_a$ is the $a \times a$ identity matrix. This implies that $\boldsymbol{\Sigma}$ is block diagonal, with n_1 blocks of the 3×3 matrix $\boldsymbol{\Sigma}_3$, say, followed by n_2 blocks of the 2×2 matrix $\boldsymbol{\Sigma}_2$, say, where

$$\boldsymbol{\Sigma}_3 = \begin{pmatrix} \sigma_e^2 + \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_e^2 + \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_e^2 + \sigma_b^2 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} \sigma_e^2 + \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_e^2 + \sigma_b^2 \end{pmatrix}$$

It can then easily be shown by verifying $\boldsymbol{\Sigma}_3 \boldsymbol{\Sigma}_3^{-1} = \boldsymbol{I}_3$ and $\boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_2^{-1} = \boldsymbol{I}_2$ that

$$\boldsymbol{\Sigma}_3^{-1} = \frac{1}{\sigma_e^2(\sigma_e^2 + 3\sigma_b^2)} \begin{pmatrix} \sigma_e^2 + 2\sigma_b^2 & -\sigma_b^2 & -\sigma_b^2 \\ -\sigma_b^2 & \sigma_e^2 + 2\sigma_b^2 & -\sigma_b^2 \\ -\sigma_b^2 & -\sigma_b^2 & \sigma_e^2 + 2\sigma_b^2 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_2^{-1} = \frac{1}{\sigma_e^2(\sigma_e^2 + 2\sigma_b^2)} \begin{pmatrix} \sigma_e^2 + \sigma_b^2 & -\sigma_b^2 \\ -\sigma_b^2 & \sigma_e^2 + \sigma_b^2 \end{pmatrix}$$

Now

$$\begin{aligned}\mathbf{Cov}(\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_2) &= (\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1}, \\ &= \left(\frac{n_1}{3} \sum_{k=1}^{3} \boldsymbol{X}_k^\top \boldsymbol{\Sigma}_3^{-1} \boldsymbol{X}_k + \frac{n_2}{2} \sum_{k=4}^{5} \boldsymbol{X}_k^\top \boldsymbol{\Sigma}_2^{-1} \boldsymbol{X}_k \right)^{-1}\end{aligned}$$

where $\boldsymbol{X}_k$ is the design matrix for a single individual on sequence k. Supposing without loss of generality that the sequences $k = 1, \ldots, 5$ are as follows

$$\begin{aligned} k &= 1 : \{0, 1, 2\}, \\ k &= 2 : \{1, 2, 0\}, \\ k &= 3 : \{2, 0, 1\}, \\ k &= 4 : \{0, 1\}, \\ k &= 5 : \{1, 0\}, \end{aligned}$$

,

we have

$$\boldsymbol{X}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}, \qquad \boldsymbol{X}_2 = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \end{pmatrix}, \qquad \boldsymbol{X}_3 = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \end{pmatrix},$$

$$\boldsymbol{X}_4 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \end{pmatrix}, \qquad \boldsymbol{X}_5 = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

Matrix multiplication then gives

$$\frac{n_1}{3}\sum_{k=1}^{3}\boldsymbol{X}_k^\top\boldsymbol{\Sigma}_3^{-1}\boldsymbol{X}_k = \frac{n_1}{3\sigma_e^2(3\sigma_b^2+\sigma_e^2)}\begin{pmatrix} 9\sigma_e^2 & 3\sigma_e^2 & 3\sigma_e^2 & 3\sigma_e^2 & 3\sigma_e^2 \\ 3\sigma_e^2 & 3(2\sigma_b^2+\sigma_e^2) & -3\sigma_b^2 & \sigma_e^2 & \sigma_e^2 \\ 3\sigma_e^2 & -3\sigma_b^2 & 3(2\sigma_b^2+\sigma_e^2) & \sigma_e^2 & \sigma_e^2 \\ 3\sigma_e^2 & \sigma_e^2 & \sigma_e^2 & 3(2\sigma_b^2+\sigma_e^2) & -3\sigma_b^2 \\ 3\sigma_e^2 & \sigma_e^2 & \sigma_e^2 & -3\sigma_b^2 & 3(2\sigma_b^2+\sigma_e^2) \end{pmatrix},$$

$$\frac{n_2}{2}\sum_{k=4}^{5}\boldsymbol{X}_k^\top\boldsymbol{\Sigma}_2^{-1}\boldsymbol{X}_k = \frac{n_2}{2\sigma_e^2(2\sigma_b^2+\sigma_e^2)}\begin{pmatrix} 4\sigma_e^2 & 2\sigma_e^2 & 0 & 2\sigma_e^2 & 0 \\ 2\sigma_e^2 & 2(\sigma_b^2+\sigma_e^2) & 0 & \sigma_e^2 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 2\sigma_e^2 & \sigma_e^2 & 0 & 2(\sigma_b^2+\sigma_e^2) & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Thus $\mathbf{Cov}(\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_2)^{-1}$ is found by adding these matrices. Then, finally, $\mathbf{Cov}(\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_2)$ can be identified as its inverse. This inverse does not have, in general, a simple form. However, it can be computed using a symbolic algebra package. Matlab code for this, and subsequently to verify the proposed form of $\mathrm{Var}(\hat{\tau}_{12})$, is available in Supplementary File 2, and from https://github.com/mjg211/article_code.

Table 1: The empirical familywise error-rate (P(Reject H_{01} or H_{02})), average required number of patients (AVN), and average required number of observations (AVO) of the five considered methods are shown for the null treatment effects scenario ($\boldsymbol{\tau} = (B, B)^\top$). Precisely, $n_1 \in \{12, 24, 36\}$, $CV \in \{0.1, 0.2, 0.3, 0.4\}$ and $R \in \{1, 2\}$ are considered. All rejection probabilities are given to four decimal places, and all AVN and AVO values are given to two decimal places.

				P(Reject H_{01} or H_{02})				AVN				AVO			
Method	n_1	R	CV:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
A_3	12	N/A		0.0491	0.0564	0.0532	0.0504	12.01	23.05	47.51	80.01	36.03	69.14	142.52	240.02
A_3	24	N/A		0.0494	0.0504	0.0532	0.0518	24.00	25.30	47.58	80.03	72.00	75.91	142.73	240.09
A_3	36	N/A		0.0500	0.0492	0.0527	0.0501	36.00	36.00	47.73	80.05	108.00	108.00	143.20	240.15
B_3	12	1		0.0304	0.0500	0.0356	0.0307	12.06	26.42	55.24	93.23	36.18	79.26	165.73	279.68
B_3	12	2		0.0294	0.0498	0.0348	0.0301	12.06	26.68	55.27	93.24	36.18	79.81	165.74	279.70
B_3	24	1		0.0290	0.0345	0.0496	0.0348	24.00	27.45	54.59	93.04	72.00	82.36	163.77	279.13
B_3	24	2		0.0290	0.0346	0.0498	0.0340	24.00	27.52	55.22	93.17	72.00	82.51	165.05	279.31
B_3	36	1		0.0290	0.0297	0.0457	0.0486	36.00	36.03	54.85	92.10	108.00	108.09	164.55	276.29
B_3	36	2		0.0300	0.0300	0.0455	0.0479	36.00	36.03	55.25	93.10	108.00	108.10	165.34	278.26
C_3	12	1		0.0502	0.0499	0.0345	0.0307	12.01	26.36	55.09	93.04	36.04	79.09	165.28	279.13
C_3	12	2		0.0502	0.0507	0.0360	0.0304	12.01	26.56	55.31	93.20	36.03	79.47	165.85	279.59
C_3	24	1		0.0494	0.0494	0.0498	0.0331	24.00	26.43	54.56	93.11	72.00	79.30	163.68	279.32
C_3	24	2		0.0497	0.0486	0.0491	0.0353	24.00	26.43	55.21	93.20	72.00	79.28	165.02	279.39
C_3	36	1		0.0500	0.0498	0.0471	0.0481	36.00	36.00	54.51	92.10	108.00	108.01	163.53	276.29
C_3	36	2		0.0497	0.0488	0.0483	0.0475	36.00	36.00	54.85	93.16	108.00	108.01	164.13	278.44
E_3	12	1		0.0295	0.0498	0.0327	0.0255	12.03	21.57	40.81	66.44	36.09	59.86	107.93	171.72
E_3	12	2		0.0286	0.0491	0.0332	0.0246	12.04	21.73	40.89	66.08	36.09	60.13	108.10	170.90
E_3	24	1		0.0296	0.0352	0.0483	0.0331	24.00	26.21	44.11	70.04	72.00	77.56	121.96	186.97
E_3	24	2		0.0291	0.0350	0.0490	0.0322	24.00	26.26	44.69	70.30	72.00	77.66	123.19	187.56
E_3	36	1		0.0299	0.0301	0.0466	0.0472	36.00	36.02	48.29	72.95	108.00	108.05	138.58	199.88
E_3	36	2		0.0303	0.0291	0.0446	0.0465	36.00	36.02	48.67	74.02	108.00	108.05	139.33	202.14
F_3	12	1		0.0494	0.0504	0.0338	0.0253	12.01	21.50	40.86	66.29	36.02	59.63	108.17	171.34
F_3	12	2		0.0512	0.0497	0.0338	0.0253	12.01	21.68	40.85	66.12	36.02	59.99	107.99	171.01
F_3	24	1		0.0509	0.0489	0.0496	0.0334	24.00	25.59	44.19	70.09	72.00	75.99	122.27	187.06
F_3	24	2		0.0492	0.0496	0.0493	0.0328	24.00	25.61	44.72	70.05	72.00	76.02	123.24	187.01
F_3	36	1		0.0492	0.0510	0.0470	0.0472	36.00	36.00	48.08	72.95	108.00	108.00	138.12	199.96
F_3	36	2		0.0494	0.0504	0.0459	0.0481	36.00	36.00	48.47	73.99	108.00	108.01	138.88	202.05

Table 2: The empirical familywise error-rate ($P(\text{Reject } H_{02})$), power ($P(\text{Reject } H_{01})$), average required number of patients (AVN), and average required number of observations (AVO) of the five considered methods are shown for the mixed treatment effects scenario ($\boldsymbol{\tau} = (\delta, B)^\top$). Precisely, $n_1 \in \{12, 24, 36\}$, $CV \in \{0.1, 0.2, 0.3, 0.4\}$ and $R \in \{1, 2\}$ are considered. All rejection probabilities are given to four decimal places, and all AVN and AVO values are given to two decimal places.

				$P(\text{Reject } H_{01})$				$P(\text{Reject } H_{02})$				AVN				AVO			
Method	n_1	R	CV:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
A_3	12	N/A		0.9831	0.8285	0.7905	0.7751	0.0277	0.0320	0.0292	0.0284	12.01	23.10	47.57	79.96	36.02	69.29	142.71	239.87
A_3	24	N/A		1.0000	0.8713	0.8106	0.7957	0.0284	0.0286	0.0297	0.0283	24.00	25.29	47.60	80.07	72.00	75.87	142.81	240.21
A_3	36	N/A		1.0000	0.9564	0.8171	0.8057	0.0276	0.0284	0.0285	0.0280	36.00	36.00	47.68	80.11	108.00	108.00	143.05	240.34
B_3	12	1		0.9685	0.8332	0.7875	0.7726	0.0169	0.0227	0.0191	0.0171	12.01	23.41	55.01	93.26	36.03	70.23	165.04	279.78
B_3	12	2		0.9680	0.8369	0.7883	0.7686	0.0168	0.0282	0.0204	0.0162	12.05	26.56	55.25	93.26	36.10	76.54	165.47	279.78
B_3	24	1		0.9998	0.8542	0.8196	0.7958	0.0165	0.0160	0.0227	0.0190	24.00	25.05	47.83	92.39	72.00	75.15	143.48	277.16
B_3	24	2		0.9999	0.8559	0.8269	0.7989	0.0154	0.0183	0.0280	0.0190	24.00	27.31	55.17	93.30	72.00	79.70	158.24	279.11
B_3	36	1		1.0000	0.9299	0.8317	0.8112	0.0159	0.0155	0.0188	0.0225	36.00	36.01	44.88	83.91	108.00	108.02	134.65	251.73
B_3	36	2		1.0000	0.9290	0.8347	0.8165	0.0157	0.0157	0.0243	0.0252	36.00	36.03	54.96	93.15	108.00	108.06	154.75	270.11
C_3	12	1		0.9827	0.8338	0.7845	0.7694	0.0285	0.0237	0.0186	0.0166	12.00	23.35	54.98	93.16	36.01	70.06	164.95	279.47
C_3	12	2		0.9832	0.8387	0.7905	0.7705	0.0285	0.0297	0.0192	0.0166	12.01	26.48	55.28	93.39	36.03	76.33	165.57	280.17
C_3	24	1		1.0000	0.8733	0.8192	0.7982	0.0280	0.0268	0.0227	0.0179	24.00	24.81	47.93	92.48	72.00	74.44	143.79	277.43
C_3	24	2		1.0000	0.8722	0.8236	0.7982	0.0284	0.0278	0.0270	0.0109	24.00	26.37	55.06	93.21	72.00	77.55	157.89	278.82
C_3	36	1		1.0000	0.9560	0.8313	0.8081	0.0274	0.0274	0.0191	0.0225	36.00	36.00	44.73	83.81	108.00	108.00	134.19	251.42
C_3	36	2		1.0000	0.9560	0.8338	0.8183	0.0273	0.0288	0.0250	0.0271	36.00	36.00	54.60	93.16	108.00	108.01	153.98	270.11
E_3	12	1		0.9678	0.8240	0.7434	0.6556	0.0165	0.0222	0.0179	0.0133	12.01	23.21	52.24	84.03	36.02	63.68	136.91	215.71
E_3	12	2		0.9680	0.8327	0.7448	0.6562	0.0161	0.0261	0.0186	0.0138	12.03	25.19	52.36	83.92	36.06	67.65	137.14	215.27
E_3	24	1		0.9999	0.8489	0.8114	0.7655	0.0167	0.0164	0.0229	0.0177	24.00	25.01	47.40	88.79	72.00	74.33	129.75	234.78
E_3	24	2		0.9999	0.8533	0.8212	0.7684	0.0162	0.0184	0.0265	0.0179	24.00	26.36	51.87	89.40	72.00	77.02	138.65	235.96
E_3	36	1		1.0000	0.9313	0.8271	0.7999	0.0156	0.0157	0.0179	0.0224	36.00	36.00	44.65	82.24	108.00	108.01	128.42	222.86
E_3	36	2		1.0000	0.9290	0.8296	0.8086	0.0156	0.0160	0.0245	0.0261	36.00	36.02	50.83	88.07	108.00	108.03	140.79	234.44
F_3	12	1		0.9833	0.8267	0.7438	0.6582	0.0273	0.0228	0.0172	0.0131	12.00	23.18	52.28	84.13	36.00	63.61	137.11	215.89
F_3	12	2		0.9825	0.8320	0.7432	0.6573	0.0286	0.0284	0.0172	0.0137	12.01	25.11	52.38	83.91	36.02	67.48	137.20	215.37
F_3	24	1		1.0000	0.8707	0.8102	0.7647	0.0276	0.0268	0.0232	0.0169	24.00	24.81	47.24	88.78	72.00	73.86	129.28	234.56
F_3	24	2		1.0000	0.8721	0.8197	0.7663	0.0280	0.0273	0.0277	0.0181	24.00	25.75	51.73	89.36	72.00	75.75	138.33	235.79
F_3	36	1		1.0000	0.9576	0.8264	0.8025	0.0273	0.0271	0.0198	0.0235	36.00	36.00	44.60	82.31	108.00	108.00	128.32	223.04
F_3	36	2		1.0000	0.9582	0.8302	0.8095	0.0282	0.0276	0.0262	0.0256	36.00	36.00	50.64	88.19	108.00	108.00	140.41	234.76

Table 3: The empirical power ($P(\text{Reject } H_{01})$), average required number of patients (AVN), and average required number of observations (AVO) of the five considered methods are shown for the alternative treatment effects scenario ($\boldsymbol{\tau} = (\delta, \delta)^\top$). Precisely, $n_1 \in \{12, 24, 36\}$, $CV \in \{0.1, 0.2, 0.3, 0.4\}$ and $R \in \{1, 2\}$ are considered. All rejection probabilities are given to four decimal places, and all AVN and AVO values are given to two decimal places.

				$P(\text{Reject } H_{01})$				AVN				AVO			
Method	n_1	R	CV:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
A_3	12	N/A		0.9838	0.8276	0.7907	0.7759	12.01	23.09	47.56	80.07	36.03	69.26	142.68	240.20
A_3	24	N/A		1.0000	0.8700	0.8108	0.7993	24.00	25.30	47.58	80.05	72.00	75.90	142.75	240.15
A_3	36	N/A		1.0000	0.9565	0.8184	0.8057	36.00	36.00	47.68	80.15	108.00	108.00	143.04	240.45
B_3	12	1		0.9677	0.7254	0.7811	0.7695	12.00	21.69	54.82	93.15	36.01	65.07	164.45	279.44
B_3	12	2		0.9686	0.8375	0.7876	0.7692	12.01	25.25	55.25	93.25	36.03	72.20	165.37	279.74
B_3	24	1		0.9999	0.8121	0.7026	0.7885	24.00	24.55	43.65	91.94	72.00	73.64	130.94	275.83
B_3	24	2		0.9999	0.8521	0.8255	0.7989	24.00	25.52	52.56	93.11	72.00	75.58	148.91	278.15
B_3	36	1		1.0000	0.9314	0.7148	0.7147	36.00	36.00	41.32	77.82	108.00	108.01	123.95	233.45
B_3	36	2		1.0000	0.9313	0.8329	0.8182	36.00	36.01	48.42	90.67	108.00	108.02	138.16	259.10
C_3	12	1		0.9824	0.7292	0.7820	0.7711	12.00	21.72	54.87	93.23	36.00	65.15	164.62	279.70
C_3	12	2		0.9828	0.8388	0.7898	0.7710	12.00	25.20	55.26	93.25	36.01	72.09	165.39	279.75
C_3	24	1		1.0000	0.8451	0.7019	0.7901	24.00	24.44	43.63	91.86	72.00	73.33	130.89	275.58
C_3	24	2		0.9999	0.8723	0.8247	0.7989	24.00	25.18	52.46	93.09	72.00	74.80	148.57	278.07
C_3	36	1		1.0000	0.9578	0.7188	0.7152	36.00	36.00	41.33	77.80	108.00	108.00	124.00	233.39
C_3	36	2		1.0000	0.9566	0.8347	0.8177	36.00	36.00	48.30	90.64	108.00	108.00	137.91	259.20
E_3	12	1		0.9678	0.7231	0.7393	0.6559	12.00	21.65	53.90	88.45	36.01	64.55	156.25	246.32
E_3	12	2		0.9676	0.8350	0.7448	0.6546	12.01	25.21	54.14	88.37	36.03	71.70	156.60	245.94
E_3	24	1		0.9999	0.8128	0.6970	0.7557	24.00	24.55	43.61	90.88	72.00	73.66	129.78	265.59
E_3	24	2		0.9998	0.8531	0.8214	0.7680	24.00	25.50	52.28	91.93	72.00	75.54	147.09	267.56
E_3	36	1		1.0000	0.9287	0.7133	0.7085	36.00	36.00	41.33	77.36	108.00	108.01	123.77	229.37
E_3	36	2		1.0000	0.9291	0.8318	0.8106	36.00	36.01	48.38	90.17	108.00	108.02	137.87	255.14
F_3	12	1		0.9821	0.7238	0.7376	0.6542	12.00	21.68	53.78	88.33	36.00	64.64	155.90	245.99
F_3	12	2		0.9835	0.8340	0.7441	0.6560	12.00	25.16	54.26	88.34	36.01	71.56	156.94	246.24
F_3	24	1		0.9999	0.8448	0.6987	0.7534	24.00	24.44	43.56	90.75	72.00	73.33	129.62	265.19
F_3	24	2		0.9999	0.8744	0.8224	0.7657	24.00	25.18	52.31	91.90	72.00	74.79	147.10	267.55
F_3	36	1		1.0000	0.9557	0.7183	0.7073	36.00	36.00	41.28	77.33	108.00	108.00	123.64	229.37
F_3	36	2		1.0000	0.9561	0.8332	0.8111	36.00	36.00	48.24	90.19	108.00	108.00	137.57	255.15

Table 4: The performance of Method E_3 for different choices of its design parameters is displayed. Specifically, a selection of results on the empirical familywise error-rate ($P(\text{Reject } H_{01} \text{ or } H_{02})$) when $\boldsymbol{\tau} = (B, B)^\top$, and the empirical power ($P(\text{Reject } H_{01})$) when $\boldsymbol{\tau} = (\delta, \delta)^\top$, for $R = 2$ and $CV \in \{0.2, 0.3\}$ are given. In addition, the average required number of patients (AVN), and average required number of required observations (AVO) are presented. The optimal design identified from the conducted grid search is also listed. All rejection probabilities are given to four decimal places, and all AVN and AVO values are given to two decimal places.

							$\boldsymbol{\tau} = (B, B)^\top$			$\boldsymbol{\tau} = (\delta, \delta)^\top$		
Design	n_1	α_1	α_2	f	Cost	CV	$P(\text{Reject } H_{01} \text{ or } H_{02})$	AVN	AVO	$P(\text{Reject } H_{01})$	AVN	AVO
Non-optimized	12	0.0294	0.0294	0	N/A	0.2	0.0491	21.73	60.13	0.8350	25.21	71.70
						0.3	0.0332	40.89	108.10	0.7448	54.14	156.60
	24	0.0294	0.0294	0	107.25	0.2	0.0350	26.26	77.66	0.8531	25.50	75.54
						0.3	0.0490	44.69	123.19	0.8214	52.28	147.09
	36	0.0294	0.0294	0	185.37	0.2	0.0291	36.02	108.05	0.9291	36.01	108.02
						0.3	0.0446	48.67	139.33	0.8318	48.38	137.87
Optimized	24	0.0300	0.0300	0.5	43.84	0.2	0.0362	25.38	75.22	0.8536	25.43	75.35
						0.3	0.0489	37.62	103.67	0.8043	51.23	142.28

Table 5: The performance of the optimized design presented in Table 4 is given. Specifically, a selection of results on the empirical familywise error-rate (FWER) when $\boldsymbol{\tau} = (B, B)^\top$, and the empirical power when $\boldsymbol{\tau} = (\delta, \delta)^\top$, for $CV \in \{0.1, 0.2, 0.3, 0.4\}$ are listed. In addition, the average number of required observations (AVO) when $\boldsymbol{\tau} = (B, B)^\top$ is presented. All rejection probabilities are given to four decimal places, and all AVN and AVO values are given to two decimal places.

	$\boldsymbol{\tau} = (B, B)^\top$			$\boldsymbol{\tau} = (\delta, \delta)^\top$		
CV	P(Reject H_{01} or H_{02})	AVN	AVO	P(Reject H_{01})	AVN	AVO
0.1	0.0299	24.00	72.00	0.9999	24.00	72.00
0.2	0.0362	25.38	75.22	0.8536	25.43	75.35
0.3	0.0489	37.62	103.67	0.8043	51.23	142.28
0.4	0.0301	54.70	144.17	0.6967	88.06	247.57