

Short Communication

Accounting for variation in the required sample size in the design of group-sequential trials

Michael J. Grayling^{a,*}, Adrian P. Mander^b^a Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, UK^b Centre for Trials Research, Cardiff University, Cardiff, UK

ARTICLE INFO

Keywords:

Adaptive design
Interim analysis
Multi-stage
Optimal design
Sample size
Two-stage

ABSTRACT

Introduction: Most literature on optimal group-sequential designs focuses on minimising the expected sample size. We highlight other factors for consideration.

Methods: We discuss several quantities less-often considered in adaptive design: the median and standard deviation of the random required sample size, and the probability of committing an interim error. We consider how the optimal timing of interim analyses changes when these quantities are accounted for.

Results: Incorporating the standard deviation of the required sample size into an optimality framework, we demonstrate how and when this quantity means using a group-sequential approach is not optimal. The optimal timing of an interim analysis is shown to be highly dependent on the pre-specified preference for minimising the expected sample size relative to its standard deviation.

Conclusions: Examining multiple factors, which measure the advantages and disadvantages of group-sequential designs, helps determine the best design for a specific trial.

1. Introduction

Adaptive designs have received substantial recent attention; a response to escalating costs in the drug development process and their potential to improve efficiency [1]. Amongst available types of adaptation, group-sequential (GS) designs are the most established in practice. In a GS approach, interim analyses that can terminate the trial are conducted after pre-planned numbers of participants. With this, the average sample size required by a GS design is often lower than a corresponding fixed-sample design [2].

When designing a GS trial, the timing of the interim analyses and the interim stopping boundaries are key required factors that can greatly influence a design's operating characteristics. Accordingly, there are many papers on optimal GS designs [3–22]. The majority of these minimise a weighted combination of expected sample sizes (ESSs) at particular treatment effects.

Here, we highlight the value of considering other factors in the optimality criteria that directly address the drawbacks of a GS approach. Through this, we extend previous work on optimal interim analysis timing in a way that allows examination of when prior desires indicate a fixed-sample approach should be preferred.

2. Methods

2.1. Group-sequential designs

Consider a balanced two-arm GS trial with a normally distributed outcome. Other settings can be treated similarly; we discuss single- and two-arm trials with Bernoulli data in the Supplementary Materials.

Assume a maximum of J analyses are allowed, with analysis $j \in \{1, \dots, J\}$ conducted using data from $t_j n$ patients on each arm, $0 < t_1 < t_2 < \dots < t_J = 1$. Set $\mathbf{t} = (t_1, \dots, t_J)$. Outcome $X_{ik} \sim N(\mu_k, \sigma^2)$, where $k \in \{0, 1\}$ indicates treatment arm and $i \in \{1, \dots, n\}$ indicates participant. The variance of the outcomes, σ^2 , is assumed known. Suppose we test $H_0 : \tau = \mu_1 - \mu_0 \leq 0$, wishing to control the type-I error-rate to $\alpha \in (0, 1)$ when $\tau = 0$ and have power of $1 - \beta \in (0, 1)$ when $\tau = \delta > 0$. At analysis j , we use a Wald test-statistic, $T_j = \hat{\tau}_j \sqrt{I_j(n, \mathbf{t})}$, where

$$\hat{\tau}_j = \frac{1}{t_j n} \sum_{i=1}^{t_j n} (x_{i1} - x_{i0}),$$

$$I_j(n, \mathbf{t}) = \frac{t_j n}{2\sigma^2}.$$

* Corresponding author at: Population Health Sciences Institute, Baddiley-Clark Building, Richardson Road, Newcastle upon Tyne NE2 4AX, UK.

E-mail address: michael.grayling@newcastle.ac.uk (M.J. Grayling).

Then, $\mathbf{T}_j = (T_1, \dots, T_j)$ is multivariate normal with $\mathbb{E}(\mathbf{T}_j) = \tau\sqrt{I_j(n, \mathbf{t})} = \tau(\sqrt{I_1(n, \mathbf{t})}, \dots, \sqrt{I_j(n, \mathbf{t})})$ and $\{\text{Cov}(T_j, T_{j_2})\}_{j_1 j_2} = \{\Lambda_j(n, \mathbf{t})\}_{j_1 j_2} = \sqrt{I_{j_1}(n, \mathbf{t})/I_{j_2}(n, \mathbf{t})}$ for $1 \leq j_1 \leq j_2 \leq j$ [23].

Design requires upper (efficacy) and lower (futility) stopping boundaries, which we denote by $\mathbf{e} = (e_1, \dots, e_j)$ and $\mathbf{f} = (f_1, \dots, f_j)$, with $e_j > f_j$ for $j \in \{1, \dots, J - 1\}$. We also set $e_J = f_J$, so that the trial terminates with a decision on whether to reject H_0 after at most J stages. The decision rules at analysis j are

- if $T_j > e_j$ terminate the trial, rejecting H_0 ;
- else if $T_j \leq f_j$ terminate the trial, but do not reject H_0 ;
- else continue the trial to stage $j + 1$.

In total, a design is given by choices for $n, \mathbf{t}, \mathbf{e}$, and \mathbf{f} . Using these, and the implied distribution of \mathbf{T}_j , we can compute the probability of stopping for efficacy (E_j) or futility (F_j) at analysis j using

$$E_j(\tau|n, \mathbf{t}, \mathbf{e}, \mathbf{f}) = \begin{cases} \int_{e_1}^{\infty} \phi_1\{x, \tau\sqrt{I_1(n, \mathbf{t})}, 1\} dx & : j = 1, \\ \int_{f_1}^{e_1} \dots \int_{f_{j-1}}^{e_{j-1}} \int_{e_j}^{\infty} \phi_j\{x_j, \tau\sqrt{I_j(n, \mathbf{t})}, \Lambda_j(n, \mathbf{t})\} dx_j \dots dx_1 & : j \in \{2, \dots, J\}, \end{cases}$$

$$F_j(\tau|n, \mathbf{t}, \mathbf{e}, \mathbf{f}) = \begin{cases} \int_{-\infty}^{f_1} \phi_1\{x, \tau\sqrt{I_1(n, \mathbf{t})}, 1\} dx & : j = 1, \\ \int_{f_1}^{e_1} \dots \int_{f_{j-1}}^{e_{j-1}} \int_{-\infty}^{f_j} \phi_j\{x_j, \tau\sqrt{I_j(n, \mathbf{t})}, \Lambda_j(n, \mathbf{t})\} dx_j \dots dx_1 & : j \in \{2, \dots, J\}, \end{cases}$$

where $\phi_j(\mathbf{x}_j) = \phi_j(\mathbf{x}_j | \boldsymbol{\mu}, \Sigma)$ is the PDF of a j -dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ evaluated at (x_1, \dots, x_j) .

Furthermore, the probability H_0 is rejected and the ESS for any τ can be found with

$$P(\tau|n, \mathbf{t}, \mathbf{e}, \mathbf{f}) = \sum_{j=1}^J E_j(\tau|n, \mathbf{t}, \mathbf{e}, \mathbf{f}),$$

$$ESS(\tau|n, \mathbf{t}, \mathbf{e}, \mathbf{f}) = \sum_{j=1}^J 2nt_j S_j(\tau|n, \mathbf{t}, \mathbf{e}, \mathbf{f}),$$

for $S_j(\tau|n, \mathbf{t}, \mathbf{e}, \mathbf{f}) = E_j(\tau|n, \mathbf{t}, \mathbf{e}, \mathbf{f}) + F_j(\tau|n, \mathbf{t}, \mathbf{e}, \mathbf{f})$.

We can also calculate several other quantities, which we will examine the importance of. These are the median sample size (MSS), the standard deviation of the sample size (SDSS), and the probability of an interim error (PIE)

$$MSS(\tau|n, \mathbf{t}, \mathbf{e}, \mathbf{f}) = \begin{cases} 2n(t_j + 0.5t_{j+1}) & : \exists j \text{ such that } \tilde{S}_j(\tau|n, \mathbf{t}, \mathbf{e}, \mathbf{f}) = 0.5, \\ 2nt_j & : \text{argmin}_{j \in \{1, \dots, J\}} \{\tilde{S}_j(\tau|n, \mathbf{t}, \mathbf{e}, \mathbf{f}) \geq 0.5\}, \end{cases}$$

$$SDSS(\tau|n, \mathbf{t}, \mathbf{e}, \mathbf{f}) = \left(\sum_{j=1}^J (2nt_j)^2 S_j(\tau|n, \mathbf{t}, \mathbf{e}, \mathbf{f}) \right) - ESS(\tau|n, \mathbf{t}, \mathbf{e}, \mathbf{f})^2,$$

$$PIE(\tau|n, \mathbf{t}, \mathbf{e}, \mathbf{f}) = \begin{cases} \sum_{j=1}^{J-1} E_j(\tau|n, \mathbf{t}, \mathbf{e}, \mathbf{f}) & : \tau \leq 0, \\ \sum_{j=1}^{J-1} F_j(\tau|n, \mathbf{t}, \mathbf{e}, \mathbf{f}) & : \tau > 0, \end{cases}$$

where $\tilde{S}_j(\tau|n, \mathbf{t}, \mathbf{e}, \mathbf{f}) = S_1(\tau|n, \mathbf{t}, \mathbf{e}, \mathbf{f}) + \dots + S_j(\tau|n, \mathbf{t}, \mathbf{e}, \mathbf{f})$. While the PIE arises in error-spending designs [24], the MSS and SDSS are rarely mentioned in the GS design literature.

2.2. Optimal designs

Many methods are available for specifying \mathbf{e} and \mathbf{f} . We consider two approaches

- Wang-Tsiatis bounds [17]: $e_j = Ct_j^{\Delta-0.5}$ for $j \in \{1, \dots, J\}$ and $f_j = -e_j$ for $j \in \{1, \dots, J - 1\}$. Here, C is chosen for any n and \mathbf{t} by solving $P(0|n, \mathbf{t}, \mathbf{e}, \mathbf{f}) = \alpha$.
- Futility stopping only, with pre-specified interim bounds: We set $e_1 = \dots = e_{J-1} = \infty$ and nominate values for f_1, \dots, f_{J-1} . Like the above, $f_J = e_J$ is chosen for any n and \mathbf{t} by solving $P(0|n, \mathbf{t}, \mathbf{e}, \mathbf{f}) = \alpha$.

In either case, given \mathbf{t} , we identify n as the solution to

$$\text{argmin}_{n \in \mathbb{R}^+} P(\delta|n, \mathbf{t}, \mathbf{e}, \mathbf{f}) \geq 1 - \beta,$$

i.e., the minimal value such that the power requirement is met.

Thus, we have methods for choosing n, \mathbf{e} , and \mathbf{f} , for given \mathbf{t} . As in Togo and Iwasaki [16] and Xi et al. [22], we will focus on how to optimally determine \mathbf{t} . Togo and Iwasaki [16] focused on Wang-Tsiatis boundaries, choosing \mathbf{t} to minimise $ESS(\delta|n, \mathbf{t}, \mathbf{e}, \mathbf{f})$. Xi et al. [22] considered our other setting with futility stopping only, discussing several optimality criteria consisting of weighted sums of ESSs.

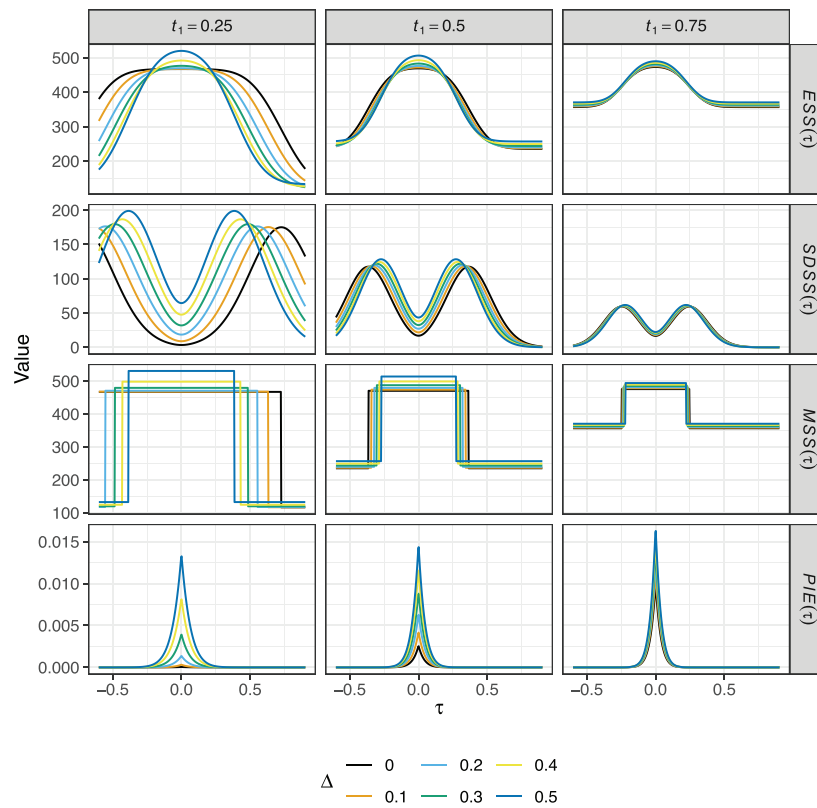


Fig. 1. The ESS, SDSS, MSS, and PIE curves are shown, in the Wang-Tsiatis designs with several values of Δ , for several values of t_1 as a function of τ .

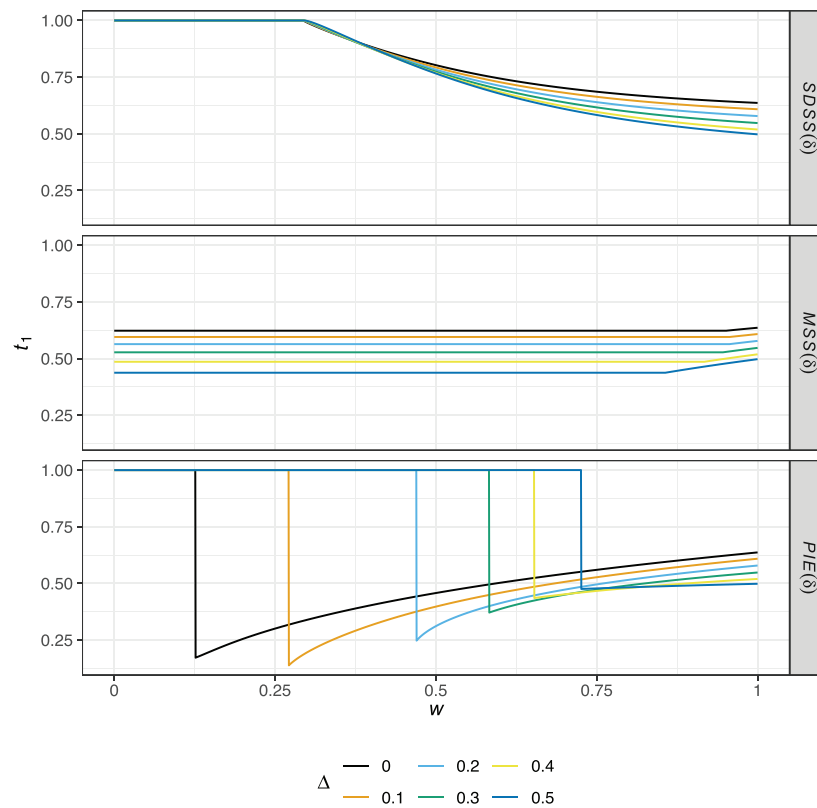


Fig. 2. The optimal value of t_1 is shown, in the Wang-Tsiatis designs with several values of Δ , as a function of $w \in [0, 1]$, the weight given to the expected sample size in the three optimality criteria, for $\theta = \delta$.

We seek to demonstrate how focusing solely on minimising ESSs can negate consideration of the drawbacks of a GS approach. Namely, any use of a GS design involves a trade-off in terms of reducing the ESS at particular treatment effects while increasing the maximal sample size and introducing non-zero variation in the required sample size. Therefore, modifying previous suggestions we consider the optimal solutions to three optimality problems

$$\begin{aligned} & \operatorname{argmin}_{t \in [0,1]^J : 0 < t_1 < \dots < t_J = 1} w \operatorname{ESS}(\theta | n, t, e, f) + (1 - w) \operatorname{SDSS}(\theta | n, t, e, f), \\ & \operatorname{argmin}_{t \in [0,1]^J : 0 < t_1 < \dots < t_J = 1} w \operatorname{ESS}(\theta | n, t, e, f) + (1 - w) \operatorname{MSS}(\theta | n, t, e, f), \\ & \operatorname{argmin}_{t \in [0,1]^J : 0 < t_1 < \dots < t_J = 1} w \frac{\operatorname{ESS}(\theta | n, t, e, f)}{\max_{t' \in [0,1]^J : 0 < t'_1 < \dots < t'_J = 1} \operatorname{ESS}(\theta | n, t', e, f)} + \\ & (1 - w) \frac{\operatorname{PIE}(\theta | n, t, e, f)}{\max_{t' \in [0,1]^J : 0 < t'_1 < \dots < t'_J = 1} \operatorname{PIE}(\theta | n, t', e, f)}, \end{aligned}$$

for different values of the weight $w \in [0, 1]$ which categorises the relative desires to minimise the two factors in the objective functions. We

consider $\theta \in \{0, \delta\}$, as these have been common choices historically in the optimal design literature. Note rescaling is included in the third optimality problem as the two factors exist on different scales.

2.3. Examples

To correspond to a typical confirmatory trial, we assume $\alpha = 0.025$ and $\beta = 0.1$. The results will not depend on the values of δ and σ ; we set $\delta = 0.3$ and $\sigma^2 = 1$ arbitrarily. We focus on two-stage designs ($J = 2$); some results for $J = 3$ are given in the Supplementary Materials. We allow a fixed-sample ($J = 1$) design to be identified as optimal by allowing $t_1 = t_2 = 1$. Code for reproduction of our results is available from https://github.com/mjg211/article_code.

3. Results

3.1. Wang-Tsiatis designs

Fig. 1 shows the ESS, SDSS, MSS, and PIE curves as a function of τ and

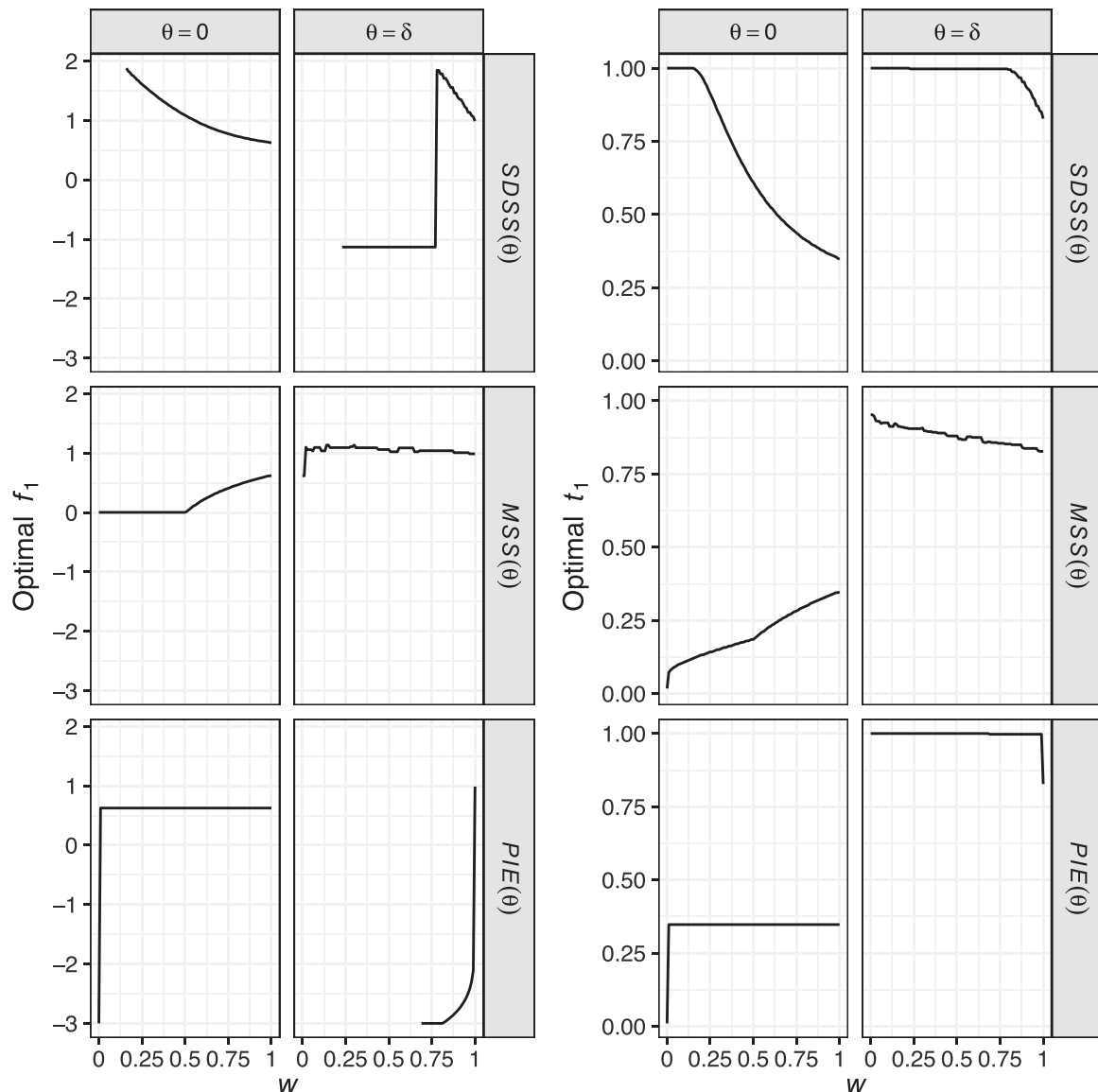


Fig. 3. The optimal values of t_1 and f_1 are shown, in the stopping for futility designs, as a function of $w \in [0, 1]$, the weight given to the expected sample size in the three optimality criteria, for $\theta \in \{0, \delta\}$.

t_1 , for several Δ . It depicts why focusing solely on the ESS may lead to designs with large values for the SDSS, MSS, or PIE. For example, the smallest considered value of t_1 produces the smallest value of $ESS(\delta)$ for all examined Δ , but its SDSS curve is substantially more variable.

Accordingly, in Fig. 2 we present the optimal value of t_1 for several Δ , as a function of w , for the three optimality criteria. We consider only $\theta = \delta$, as the symmetry of Wang-Tsiatis bounds means $\theta = 0$ will always recommend a fixed-sample design.

For the optimality criteria incorporating $SDSS(\delta)$ and $PIE(\delta)$, small w results in a single-stage design (i.e., $t_1 = 1$) being optimal. For larger w , the optimal timing of the interim analysis changes rapidly in w ; indicative of a large trade-off between the two factors that make up the optimality criteria. Results for the optimality problem involving $MSS(\delta)$ are different; they indicate regardless of Δ and w , t_1 in the region of 0.5 is optimal.

3.2. Futility stopping

Next, we consider optimising t_1 and f_1 in the designs that allow early stopping for futility alone. The solutions as a function of w are given in Fig. 3 for $\theta \in \{0, \delta\}$ for the three optimality criteria.

For $\theta = 0$, with all three optimality criteria w need not be large for a GS approach (i.e., $t_1 < 1$) to be optimal. However, the optimal timing of the interim analysis and the optimal futility bound again often change quickly in w , and vary substantially across the optimality criteria. For $\theta = \delta$, the inclusion only of early stopping for futility means a GS design is typically only optimal for larger w .

4. Discussion

Adaptive designs are not always useful [25]. In a GS design, several issues can arise, for example, it may not be clear what will happen to trial staff if a study terminates early. Costing the trial can also be more challenging, as it may be necessary to determine costs for each possible sample size. Such issues mean in some settings the advantages a GS design brings may not outweigh the drawbacks.

Therefore, we have here demonstrated the utility of directly considering factors that address disadvantages when optimising the design. The optimal design was demonstrated to be highly sensitive to the choice of weighting parameter w (Figs. 2–3). Thus, in practice this may mean a trialist could choose to compromise on the ESS for, e.g., a large reduction in SDSS. By determining what factors matter most to their study, our approach would allow a trialist to determine when an interim analysis could be most effectively timed, and indeed whether one should be conducted at all.

For two-stage single-arm trials with Bernoulli data, we also demonstrate in the Supplementary Materials how a small concession on the ESS may lead to much larger gains in terms of reducing the SDSS. This echoes similar findings from Hanfelt et al. [11] for the MSS.

In conclusion, we encourage those considering utilising a GS design to explicitly evaluate the MSS, SDSS, and PIE when choosing a design. They can have a notable impact on the optimal design, and may even indicate that a GS design would not be appropriate.

Declaration of interests

The authors declare that they have no known competing financial

interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.cct.2021.106459>.

References

- [1] M. Dimairo, P. Pallmann, J. Wason, et al., The adaptive designs CONSORT extension (ACE) statement: a checklist with explanation and elaboration guideline for reporting randomised trials that use an adaptive design, *BMJ* 369 (2020) m115.
- [2] C. Jennison, B. Turnbull, *Group Sequential Methods with Applications to Clinical Trials*, Chapman and Hall/CRC, New York, NY, 2000.
- [3] K. Anderson, Optimal spending functions for asymmetric group sequential designs, *Biom. J.* 49 (2007) 337–345.
- [4] S. Barber, C. Jennison, Optimal asymmetric one-sided group sequential tests, *Biometrika* 89 (2002) 49–60.
- [5] E. Brittain, K. Bailey, Optimization of multistage testing times and critical values in clinical trials, *Biometrics* 49 (1993) 763–772.
- [6] M. Chang, Optimal designs for group sequential clinical trials, *Commun. Stat. Theory* 25 (1996) 361–379.
- [7] J. Eales, C. Jennison, An improved method for deriving optimal one-sided group sequential tests, *Biometrika* 79 (1992) 13–24.
- [8] J. Eales, C. Jennison, Optimal two-sided group sequential tests, *Seq. Anal.* 14 (1995) 273–286.
- [9] M. Grayling, J. Wason, A. Mander, Efficient determination of optimised multi-arm multi-stage experimental designs with control of generalised error-rates, *arXiv* (2017) arXiv:1712.00229.
- [10] M. Grayling, J. Wason, A. Mander, An optimised multi-arm multi-stage clinical trial design for unknown variance, *Contemp. Clin. Trials* 67 (2018) 116–120.
- [11] J. Hanfelt, R. Slack, E. Gehan, A modification of Simon's optimal design for phase II trials when the criterion is median sample size, *Control. Clin. Trials* 20 (1999) 555–566.
- [12] S. Jung, T. Lee, K. Kim, S. George, Admissible two-stage designs for phase II cancer clinical trials, *Stat. Med.* 23 (2004) 561–569.
- [13] A. Mander, S. Thompson, Two-stage designs optimal under the alternative hypothesis for phase II cancer clinical trials, *Contemp. Clin. Trials* 31 (2010) 572–578.
- [14] A. Mander, J. Wason, M. Sweeting, S. Thompson, Admissible two-stage designs for phase II cancer clinical trials that incorporate the expected sample size under the alternative hypothesis, *Pharm. Stat.* 11 (2012) 91–96.
- [15] R. Simon, Optimal two-stage designs for phase II clinical trials, *Control. Clin. Trials* 10 (1989) 1–10.
- [16] K. Togo, M. Iwasaki, Optimal timing for interim analyses in clinical trials, *J. Biopharm. Stat.* 23 (2013) 1067–1080.
- [17] S. Wang, A. Tsiatis, Approximately optimal one-parameter boundaries for group sequential trials, *Biometrics* 43 (1987) 193–199.
- [18] J. Wason, T. Jaki, Optimal design of multi-arm multi-stage trials, *Stat. Med.* 31 (2012) 4269–4279.
- [19] J. Wason, A. Mander, S. Thompson, Optimal multi-stage designs for randomised clinical trials with continuous outcomes, *Stat. Med.* 31 (2012) 301–312.
- [20] J. Wason, A. Mander, Minimising the maximum expected sample size in two-stage phase II clinical trials with continuous outcomes, *J. Biopharm. Stat.* 22 (2012) 836–852.
- [21] J. Wason, OptGS: an R package for finding near-optimal group-sequential designs, *J. Stat. Softw.* 66 (2015) 1–13.
- [22] D. Xi, P. Gallo, D. Ohlssen, On the optimal timing of futility interim analyses, *J. Biopharm. Stat.* 9 (2017) 293–301.
- [23] C. Jennison, B. Turnbull, Group-sequential analysis incorporating covariate information, *J. Am. Stat. Assoc.* 92 (1997) 1330–1341.
- [24] K. Lan, D. DeMets, Discrete sequential boundaries for clinical trials, *Biometrika* 70 (1983) 659–663.
- [25] J. Wason, P. Brocklehurst, C. Yap, When to keep it simple – adaptive designs are not always useful, *BMC Med.* 17 (2019) 152.