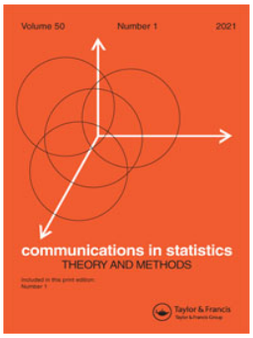

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

# Exact group sequential designs for two-arm experiments with Poisson distributed outcome variables

Michael J. Grayling, James M. S. Wason & Adrian P. Mander

**To cite this article:** Michael J. Grayling, James M. S. Wason & Adrian P. Mander (2021) Exact group sequential designs for two-arm experiments with Poisson distributed outcome variables, Communications in Statistics - Theory and Methods, 50:1, 18-34, DOI: 10.1080/03610926.2019.1628273

**To link to this article:** https://doi.org/10.1080/03610926.2019.1628273

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
2021, VOL. 50, NO. 1, 18–34
https://doi.org/10.1080/03610926.2019.1628273



# Exact group sequential designs for two-arm experiments with Poisson distributed outcome variables

Michael J. Grayling[a,b], James M. S. Wason[a,b], and Adrian P. Mander[a,c]

[a]MRC Biostatistics Unit, Hub for Trials Methodology Research, Cambridge, UK; [b]Institute of Health & Society, Newcastle University, Newcastle, UK; [c]Centre for Trials Research, Cardiff University, Cardiff, UK

**ABSTRACT**

We describe and compare two methods for the group sequential design of two-arm experiments with Poisson distributed data, which are based on a normal approximation and exact calculations respectively. A framework to determine near-optimal stopping boundaries is also presented. Using this framework, for a considered example, we demonstrate that a group sequential design could reduce the expected sample size under the null hypothesis by as much as 44% compared to a fixed sample approach. We conclude with a discussion of the advantages and disadvantages of the two presented procedures.

**ARTICLE HISTORY**
Received 26 August 2018
Accepted 31 May 2019

**KEYWORDS**
Adaptive design; exact; interim analysis; optimal; Skellam; two-stage

## 1. Introduction

It is desirable that experiments be designed and analyzed to ensure control of their type-I and type-II error-rates. For, within the context of a clinical trial, for example, an inflated type-I error-rate reduces our confidence that a significant health benefit has not been observed by chance, whilst under-powering a study raises the risk of failing to identify an efficacious treatment. It is for this reason that much research has been conducted to develop methodology for type-I error-rate control and sample size determination, in studies with many possible types of data, and many possible choices of null hypothesis. Within the context of experiments with discrete outcome variables, this research has often focused on the establishment of methods that are able to determine operating characteristics exactly. This allows us to reduce our reliance on approximate techniques that typically depend on asymptotic results. See, for example, Jung (2012), which summarizes exact methods for experiments with binary outcomes.

In contrast, for randomized two-arm experiments with Poisson distributed outcome variables, several authors have proposed approximate procedures for study design (Thode 1997; Shiue and Bain 1982; Mathews 2010), but it was only comparatively recently that Menon et al. (2011) described an alternative approach based on the exact distribution of the difference between two Poisson variables. They demonstrated that when the anticipated

CONTACT Michael J. Grayling michael.grayling@newcastle.ac.uk Institute of Health & Society, Baddiley Clark Building, Richardson Road, Newcastle upon Tyne NE2 4AX, UK.

Supplemental data for this article can be accessed here.

difference between the Poisson rates in the two arms was large, the sample size required by their approach was typically 5-10% smaller than that based on a normal approximation.

This result is a valuable one, as count outcomes occur in many experimental design settings of practical interest. In particular, within the context of clinical research, they are common in the form of the number of observed adverse events. More specifically, count outcomes occur as an endpoint of interest in epilepsy studies through seizure counts, in multiple sclerosis or Parkinson's trials via relapse counts, and in migraine treatment studies as the number of attacks. So to are they useful in cardiovascular trials as the number of hospitalizations observed over a particular time-frame, or in alcohol treatment studies as the number of drinks over a recent period of time. Whilst in asthma trials, the number of exacerbations is often of interest, and as we will discuss later, the number of apnea and hypopnea events per hour is a common primary outcome in sleep-apnea studies.

Regardless of the now available efficient exact method noted above though, it is always of interest to be able to develop methods for reducing requisite sample sizes further. One widely applicable approach to achieving this is to employ a group sequential design, through which an experiment's expected sample size (ESS) can be reduced by permitting early stopping following the repeated testing of a hypothesis of interest. Depending on the context, this can provide valuable savings in terms of time or money. For further details on group sequential methods see Jennison and Turnbull (2000).

Unfortunately, compared to settings with normally or Bernoulli distributed data, there is relatively little methodology available pertaining to the group sequential design of trials with Poisson distributed outcome variables. In fact, this is true more generally of count outcomes. Notable exceptions include the work of Cook and Lawless (1996), who described a non-parametric approach to interim monitoring in comparative studies with recurrent event outcomes. Moreover, Jiang (1999) considered the group sequential design of experiments with heterogeneous recurrent event endpoints. Xia and Hoover (2007) described a group sequential framework for Poisson outcome variables when interim analyses were timed after landmark numbers of events had occurred in each arm. They based their testing framework on the exact distribution of the number of events from one arm, conditional on the total observed number of events. Cook et al. (2010) presented methods for the sequential analysis of data from experiments with recurrent event responses observed over two periods, where one is a baseline period of observation. Recently, Mutze et al. (2018a) also considered recurrent event responses, presenting group sequential procedures for a robust semi-parametric analysis of such data. Finally, Mutze et al. (2018b) described methodology for the group sequential design of trials with negative-binomial outcomes based on Wald test statistics using maximum likelihood estimators.

Here, we focus on a different design scenario to these articles, in which each outcome is assumed to be a single Poisson distributed variable, with its precise distribution dependent only on which arm it was accrued from. First, we describe how established group sequential design theory can be applied in this setting, allowing widely available software for design determination to be employed, and operating characteristics to be approximately controlled. Following this, we detail a novel design that allows for the exact computation of a design's operating characteristics based on the extension of the approach of Menon et al. (2011) to group sequential experiments. To allow maximal efficiency gains to be attained from the group sequential tests, we then describe an

effective means of choosing stopping boundaries in a near-optimal manner. Next, we expound on the potential gains a group sequential design could bring through two hypothetical examples, before concluding with a discussion of the advantages and disadvantages of the two described procedures.

## 2. Methods

### 2.1. Notation and hypotheses

We consider an experiment in which data is to be accrued from two arms, which we index by $j \in \{1, 2\}$. In addition, we suppose that our group sequential experiment will have at most $K \in \mathbb{N}^+$ stages (explicitly permitting $K = 1$, which corresponds to a fixed sample design), and we index the stages by $k \in \{1, ..., K\}$. Thus, from here, unless stated otherwise, it can be assumed that $j \in \{1, 2\}$ and $k \in \{1, ..., K\}$.

We then denote our outcomes variables by $X_{ijk}$: the number of observed events for sample $i$, in arm $j$, in stage $k$. And, we assume that each outcome variable is accrued following observation over an equal fixed positive time, and that then $X_{ijk} \sim Po(\lambda_j)$, for $\lambda_j \in \mathbb{R}^+$ the mean event rate on arm $j$. Furthermore, for simplicity, we assume that analyses are performed after $kn$, $n \in \mathbb{N}^+$, outcome variables have been gathered on each of the two arms (implying $i \in \{1, ..., n\}$). Note however that the methods which follow could readily be extended to consider unequal allocation to the arms in and across the stages.

Our null hypothesis of interest will be $H_0 : \lambda_1 = \lambda_2 \in \Lambda_0 \subset (0, \infty)$. That is, we test a composite null hypothesis, to expressly allow for design in scenarios (see Example 1 below) in which a range of possible values of the null mean response are anticipated. Moreover, we power our experiment for a scenario in which $\lambda_1 = \lambda_2 + \delta \in \Lambda_1 \subset (0, \infty)$ for $\delta \in \mathbb{R}^+$, as we are principally motivated by a clinical scenario in which we hope the novel treatment (arm $k = 2$) will reduce the mean response. However, note that a design for $\delta \in \mathbb{R}^-$ could be identified similarly.

We allow early stopping both to reject and to not reject $H_0$, with our stopping rules dependent on vectors of boundaries that we denote by $\boldsymbol{a} = (a_1, ..., a_K)$ and $\boldsymbol{r} = (r_1, ..., r_K)$. For now, we assume that $\boldsymbol{a}, \boldsymbol{r} \in \mathbb{R}^K$. More precisely, we will specify test statistics $T_k$, which will be used with the following stopping rules

- For $k \in \{1, \ldots, K-1\}$
  - If $T_k \geq r_k$ stop the experiment, rejecting $H_0$;
  - If $T_k < a_k$ stop the experiment, not rejecting $H_0$;
  - If $T_k \in [a_k, r_k)$ continue the experiment to stage $k+1$.
- For $k = K$
  - If $T_k \geq r_k$ stop the experiment, rejecting $H_0$;
  - If $T_k < a_k$ stop the experiment, not rejecting $H_0$.

To ensure that our stopping rules make sense (i.e., that we do not simultaneously recommend to both reject and not reject $H_0$), and to guarantee that a conclusion is drawn

on $H_0$, we enforce that $a_k < r_k$ for $k \in \{1, ..., K-1\}$ and set $a_K = r_K$. We will also make use of the notation $\boldsymbol{a}_k = (a_1, ..., a_k)$, and similarly for $\boldsymbol{r}$.

Key to the identification of our group sequential designs will then be the ability to calculate the probability we stop to reject $H_0$, $R_k(\cdot)$, or stop not rejecting $H_0$, $A_k(\cdot)$, at each analysis $k$, conditional on the values of the design parameters. Formally, these are

$$A_k(\lambda_1, \lambda_2 | n, \boldsymbol{a}_k, \boldsymbol{r}_k) = \begin{cases} \mathbb{P}\big[T_1 \in (-\infty, a_1) | \lambda_1, \lambda_2, n\big] & : k = 1, \\ \mathbb{P}\{T_1 \in [a_1, r_1), ..., T_{k-1} \in [a_{k-1}, r_{k-1}), & \\ T_k \in (-\infty, a_k) | \lambda_1, \lambda_2, n\} & : k \in \{2, ..., K\} \end{cases}$$

and

$$R_k(\lambda_1, \lambda_2 | n, \boldsymbol{a}_k, \boldsymbol{r}_k) = \begin{cases} \mathbb{P}\big\{T_1 \in [r_1, \infty) | \lambda_1, \lambda_2, n\big\} & : k = 1, \\ \mathbb{P}\{T_1 \in [a_1, r_1), ..., T_{k-1} \in [a_{k-1}, r_{k-1}), & \\ T_k \in [r_k, \infty) | \lambda_1, \lambda_2, n\} & : k \in \{2, ..., K\} \end{cases}$$

respectively, where we note that the distribution of the $T_k$ will be dependent upon $\lambda_1$, $\lambda_2$, and $n$.

With the above, we can compute the maximal type-I error-rate of our test as

$$\alpha'(n, \boldsymbol{a}, \boldsymbol{r}) = \max_{\lambda \in \Lambda_0} \sum_{k=1}^{K} R_k(\lambda, \lambda | n, \boldsymbol{a}_k, \boldsymbol{r}_k)$$

Furthermore, the maximal type-II error-rate is

$$\beta'(n, \boldsymbol{a}, \boldsymbol{r}) = \max_{\lambda \in \Lambda_1} \sum_{k=1}^{K} A_k(\lambda, \lambda - \delta | n, \boldsymbol{a}_k, \boldsymbol{r}_k)$$

Consequently, our goal in what follows will be to choose values for $n$, $\boldsymbol{a}$, and $\boldsymbol{r}$ such that $\alpha'(n, \boldsymbol{a}, \boldsymbol{r}) \leq \alpha$, and $\beta'(n, \boldsymbol{a}, \boldsymbol{r}) \leq \beta$, for specified $\alpha, \beta \in (0, 1)$.

### 2.2. Design based on a normal approximation

First, we describe how we can compute group sequential designs based on the asymptotic distribution of Wald-type test statistics. To this end, denote the maximum likelihood estimate of $\lambda_j$, at analysis $k$, by $\hat{\lambda}_{jk}$. We have

$$\hat{\lambda}_{jk} = \frac{1}{kn} \sum_{l=1}^{k} \sum_{i=1}^{n} X_{ijl}$$

Next, note that the expected Fisher information of the parameter $\lambda_j$ at analysis $k$ is $I_{jk} = kn/\lambda_j$. Then, key to design determination is the notion of the information level at analysis $k$, which serves a measure of the knowledge available about the difference of the means of the two treatment arms. We denote this information by $\mathcal{I}_k$, with it given by

$$\mathcal{I}_k = \frac{1}{\frac{1}{I_{1k}} + \frac{1}{I_{2k}}} = \frac{kn}{\lambda_1 + \lambda_2}$$

Importantly, we have

$$\left(\hat{\lambda}_{1k}-\hat{\lambda}_{2k}\right)\sqrt{\mathcal{I}_k} \xrightarrow{\mathcal{D}} N(0,1) \text{ as } n \to \infty$$

Unfortunately, the information level $\mathcal{I}_k$ depends on the unknown means, $\lambda_1$ and $\lambda_2$. Therefore, to test $H_0$ using asymptotic normality we replace $\mathcal{I}_k$ by a consistent estimator, $\hat{\mathcal{I}}_k$, obtaining the Wald-type test statistic for analysis $k$ as

$$T_{Wk} = \left(\hat{\lambda}_{1k}-\hat{\lambda}_{2k}\right)\sqrt{\hat{\mathcal{I}}_k} = \left(\hat{\lambda}_{1k}-\hat{\lambda}_{2k}\right)\sqrt{\frac{kn}{\hat{\lambda}_{1k}+\hat{\lambda}_{2k}}}$$

Now, because $\mathcal{I}_k$ is estimated using a consistent estimator, Slutsky's theorem tells us that $T_{Wk}$ is still asymptotically normally distributed. Moreover, by standard results in group sequential design theory, the vector of Wald-type test statistics $\boldsymbol{T}_W = (T_{W1}, \ldots, T_{WK})$ follows what has been referred to as the 'canonical joint distribution' (Scharfstein et al. 1997; Jennison and Turnbull 2000). That is, $\boldsymbol{T}_W$ is asymptotically multivariate normal with mean vector $(\lambda_1-\lambda_2)\boldsymbol{\mathcal{I}}_K^{1/2} = (\lambda_1-\lambda_2)(\mathcal{I}_1^{1/2}, \ldots, \mathcal{I}_K^{1/2})^\top$ and $K \times K$ covariance matrix $\Sigma = \{\Sigma_{k_1k_2}\}$ with $\Sigma_{k_1k_2} = \Sigma_{k_2k_1} = \sqrt{\mathcal{I}_{k_1}/\mathcal{I}_{k_2}} = \sqrt{k_1/k_2}$, for $k_2 \in \{1, \ldots, K\}$ and $k_1 \in \{1, \ldots, k_2\}$.

Accordingly, we can compute the $A_k(\cdot)$ and $R_k(\cdot)$ via

$$A_k(\lambda_1, \lambda_2 | n, \boldsymbol{a}_k, \boldsymbol{r}_k) = \begin{cases} \int_{-\infty}^{a_1} \phi\left[x, (\lambda_1-\lambda_2)\mathcal{I}_1^{1/2}, 1\right] \mathrm{d}x & : k = 1, \\ \int_{a_1}^{r_1} \ldots \int_{-\infty}^{a_k} \phi\left[\boldsymbol{x}, (\lambda_1-\lambda_2)\boldsymbol{\mathcal{I}}_k^{1/2}, \Sigma_k\right] \mathrm{d}x_k \ldots \mathrm{d}x_1 & : k \in \{2, \ldots, K\} \end{cases}$$

and

$$R_k(\lambda_1, \lambda_2 | n, \boldsymbol{a}_k, \boldsymbol{r}_k) = \begin{cases} \int_{r_1}^{\infty} \phi\left[x, (\lambda_1-\lambda_2)\mathcal{I}_1^{1/2}, 1\right] \mathrm{d}x & : k = 1, \\ \int_{a_1}^{r_1} \ldots \int_{r_k}^{\infty} \phi\left[\boldsymbol{x}, (\lambda_1-\lambda_2)\boldsymbol{\mathcal{I}}_k^{1/2}, \Sigma_k\right] \mathrm{d}x_k \ldots \mathrm{d}x_1 & : k \in \{2, \ldots, K\} \end{cases}$$

respectively. Here, $\phi(\boldsymbol{x}, \boldsymbol{\mu}, \Omega)$ is the probability density function of a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\Omega$, and $\Sigma_k$ signifies the restriction of $\Sigma$ to its first $k$ rows and columns.

### 2.3. Design based on exact calculations

Given that the methodology described in Section 2.2 relies upon asymptotic theory, the operating characteristics computed using multivariate normal distribution functions could be a poor approximation to their empirical values. Therefore, to permit strict error control we now extend the results of Menon et al. (2011) to group sequential designs. To achieve this, note that the sum of the outcomes in arm $j$ in stage $k$, $Y_{jk}$ say, has the following distribution based on the familiar result that the sum of independent Poisson random variables is itself Poisson

$$Y_{jk} = \sum_{i=1}^{n} X_{ijk} \sim Po\left(n\lambda_j\right)$$

Then, the difference in the sum of the outcomes on each arm, $\tilde{T}_k = Y_{1k} - Y_{2k}$, as a difference between two independent Poisson random variables, has a Skellam distribution (Skellam 1946). We signify this by $\tilde{T}_k \sim Skellam(n\lambda_1, n\lambda_2)$, and denote the probability mass and cumulative distribution functions of $\tilde{T}_k$ on its support $\mathbb{Z}$ as follows

$$g(t|n,\lambda_1,\lambda_2) = \mathbb{P}\big(\tilde{T}_k = t|n,\lambda_1,\lambda_2\big) = e^{-n(\lambda_1+\lambda_2)}\left(\frac{\lambda_1}{\lambda_2}\right)^{t/2} I_{|t|}\left(2n\sqrt{\lambda_1\lambda_2}\right),$$
$$G(t|n,\lambda_1,\lambda_2) = \mathbb{P}\big(\tilde{T}_k \leq t|n,\lambda_1,\lambda_2\big) = \sum_{\{s\in\mathbb{Z}\,:\,s\leq t\}} g(s|n,\lambda_1,\lambda_2)$$

where $I_\nu(\cdot)$ is the modified Bessel function of the first kind, and we make use of the particular representation presented in Menon et al. (2011).

Our test statistic at analysis $k$ for the exact design is then $T_{Sk} = \tilde{T}_1 + \cdots + \tilde{T}_k$. Now, within the context of exact group sequential designs for Bernoulli distributed outcome variables, it is well understood that the interim test statistics do not in general have a simple distribution (Jung 2012). Similarly, it is important to note that $T_{Sk} \nsim Skellam(kn\lambda_1, kn\lambda_2)$. What is more, unlike the case for Bernoulli outcomes, we are in fact unable to compute the probability mass function of $T_{Sk}$ across its entire support, $\mathbb{Z}$. This is a consequence specifically of the fact that the support of each $\tilde{T}_k$ is infinite. However, we can compute a part of the probability mass function, which we denote by $h_k(t|n,\lambda_1,\lambda_2,\boldsymbol{a}_k,\boldsymbol{r}_k) = \mathbb{P}(T_{Sk} = t|n,\lambda_1,\lambda_2,\boldsymbol{a}_k,\boldsymbol{r}_k)$. Explicitly

$$h_k(t|n,\lambda_1,\lambda_2,\boldsymbol{a}_k,\boldsymbol{r}_k) = \begin{cases} g(t|n,\lambda_1,\lambda_2) & : k = 1, \\ \displaystyle\sum_{s=\lfloor a_{k-1}\rfloor}^{\lfloor r_{k-1}\rfloor - 1} h_{k-1}(s|n,\lambda_1,\lambda_2,\boldsymbol{a}_{k-1},\boldsymbol{r}_{k-1}) & \\ \quad g(t-s|n,\lambda_1,\lambda_2) & : k \in \{2,\ldots,K-1\}, t \in [a_k, r_k) \end{cases}$$

That is, we are able to evaluate the probability mass function for $T_{Sk} \in [a_k, r_k)$. This result is in essence an extension of that of Schultz et al. (1973) for single-arm experiments with Bernoulli outcome variables.

Fortunately, the above is all that is required to compute $A_k(\cdot)$ and $R_k(\cdot)$, because

$$R_k(\lambda_1,\lambda_2|n,\boldsymbol{a}_k,\boldsymbol{r}_k) = \begin{cases} 1-G(r_1-1|n,\lambda_1,\lambda_2) & : k = 1, \\ \displaystyle\sum_{t=\lfloor a_{k-1}\rfloor}^{\lfloor r_{k-1}\rfloor - 1} h_{k-1}(t|n,\lambda_1,\lambda_2,\boldsymbol{a}_{k-1},\boldsymbol{r}_{k-1}) & \\ \quad \big\{1-G(r_k-t-1|n,\lambda_1,\lambda_2)\big\} & : k \in \{2,\ldots,K\} \end{cases}$$

and

$$A_k(\lambda_1,\lambda_2|n,\boldsymbol{a},\boldsymbol{r}) = \begin{cases} G(a_1-1|n,\lambda_1,\lambda_2) & : k = 1, \\ \displaystyle\sum_{t=\lfloor a_{k-1}\rfloor}^{\lfloor r_{k-1}\rfloor - 1} h_{k-1}(t|n,\lambda_1,\lambda_2,\boldsymbol{a}_{k-1},\boldsymbol{r}_{k-1}) & \\ \quad G(a_k-t-1|n,\lambda_1,\lambda_2) & : k \in \{2,\ldots,K\} \end{cases}$$

Thus, we can compute the operating characteristics to arbitrary accuracy provided we are able to evaluate $g(\cdot)$ and $G(\cdot)$. We can achieve this in R using the skellam package (Lewis et al. 2016), with $G(\cdot)$ in particular computed using a relationship between the Skellam and $\chi^2$ distributions due to Johnson (1959).

### 2.4. Boundary determination

A variety of methods have been presented in the literature through which stopping boundaries and sample sizes for group sequential designs can be determined. For the method of Section 2.2, the normal approximation approach, many of these could be applied in our considered Poisson-outcome experimental design scenario. This includes comparatively simpler procedures such as that of Pampallona and Tsiatis (1994), through to more complex methods that allow design optimization (Wason et al. 2012). For our framework from Section 2.3, however, there are fewer methods available that could be directly applied to determine suitable designs. The primary reason for this is that in this setting the stopping boundaries are best treated as discrete parameters (i.e., it is best to apply the restriction $\boldsymbol{a}, \boldsymbol{r} \in \mathbb{Z}^K$), otherwise a redundancy will exist in the design space (e.g., two designs that are otherwise equal apart from their values of $a_1$ being 2.1 and 2.2 respectively would have the same operating characteristics).

This makes a brute-force approach that searches across possible designs tempting, similar to that which has been applied to both non-randomized and randomized trials with binary outcome variables. Unfortunately, such a solution will for many $K$ be computationally expensive given that the design space is $2K$-dimensional. Consequently, here, we propose an approach to determining 'near-optimal' values for these discrete bounds based on the error spending approach to group sequential design (Lan and DeMets 1983). We also utilize this method for determining designs based on the normal approximation procedure, in order to allow for a fairer comparison to the performance of the exact procedure, and as in our experience it offers an advantageous tradeoff between computational run-time and design efficiency.

However, for the exact approach with $K = 2$, we do also contrast the performance of these near-optimal designs to 'optimal' designs identified via an extensive search over possible values for $n$, $\boldsymbol{a}$ and $\boldsymbol{r}$. Through this, we aim to demonstrate that a brute-force style approach may often provide little advantage over the comparatively easy-to-identify near-optimal designs. For this search, note that the infinite support of the $\tilde{T}_{Sk}$ implies that an exhaustive assessment of all possible boundary values is impossible. Accordingly, for any $n$, a method is required for limiting the considered $\boldsymbol{a}$ and $\boldsymbol{r}$ in a logical manner. Note that these restrictions should take in to account the chosen values for $\Lambda_0, \Lambda_1$, and $\delta$. Here, for any $n$ our approach is to identify the solutions of the following equations

$$a_* = \underset{a \in \mathbb{Z}}{\operatorname{argmax}} \left[ \max_{\lambda \in \Lambda_0} \left\{ A_1(\lambda, \lambda | n, a, r) \leq \epsilon \right\} \right],$$
$$r_* = \underset{r \in \mathbb{Z}}{\operatorname{argmin}} \left[ \max_{\lambda \in \Lambda_1} \left\{ R_1(\lambda, \lambda - \delta | n, a, r) \leq \epsilon \right\} \right]$$

That is, we choose $r_*$ as the minimal integer that ensures the probability of stopping the trial after stage one to reject $H_0$, under $\lambda_1 = \lambda_2 + \delta \in \Lambda_1$, is at most $\epsilon$, and similarly

for $a_*$. We then search over boundaries such that $a_1 \in [a_*, r_*-2]$, $r_1 \in [a_1+2, r_*]$, and $a_2 \in [a_1+a_*, r_1+r_*]$. This ensures that having continued to stage two, the conditional probability of rejecting $H_0$ is at most $\epsilon$ when $\lambda_1 = \lambda_2 + \delta \in \Lambda_1$, with a similar statement holding true for the probability of not rejecting $H_0$ at the end of stage two when $\lambda_1 = \lambda_2 \in \Lambda_0$. Thus, we allow values for $\boldsymbol{r}$ that provide conditional probabilities of stopping after each stage to reject $H_0$ of at least $\epsilon$ under some scenario for which we wish to power the trial, and similarly for the values of $\boldsymbol{a}$. For, if $\epsilon \ll 1$ it is reasonable to expect that little could be gained from designs that are not captured as part of this search procedure. Accordingly, in our search we set $\epsilon = 10^{-7}$. Then, we perform our evaluations over group-sizes $n$ such that $n \in \{1, ..., 1.5n_{\text{fixed}}\}$, where $n_{\text{fixed}}$ is the group-size required by a corresponding fixed-sample ($K=1$) design. Note that the maximal type-I and type-II error-rates for any considered design, across the sets $\Lambda_0$ and $\Lambda_1$, are identified using Brent's algorithm (Brent 1973).

In contrast, using the error-spending approach, we specify error spending vectors $\boldsymbol{\pi}_A = (\pi_{A1}, ..., \pi_{AK})$ and $\boldsymbol{\pi}_R = (\pi_{R1}, ..., \pi_{RK})$, with

$$\sum_{k=1}^{K} \pi_{Rk} = \alpha, \quad \sum_{k=1}^{K} \pi_{Ak} = \beta, \quad \pi_{Rk}, \pi_{Ak} \geq 0, k \in \{1, ..., K\}$$

that then imply particular stopping boundaries, and a particular required group size. First, we describe how the boundaries are chosen for a fixed $n \in \mathbb{N}^+$, as well as fixed $\boldsymbol{\pi}_A$ and $\boldsymbol{\pi}_R$ conforming to the requirements above. We then describe how $n$ can subsequently be chosen for these $\boldsymbol{\pi}_A$ and $\boldsymbol{\pi}_R$, before describing how the spending vectors themselves can be specified.

Put simply, for a given $n$, $\boldsymbol{\pi}_A$ and $\boldsymbol{\pi}_R$, we recursively identify the boundaries, beginning with $r_1$. The precise nature of the calculations differ for the exact and normal approximation based approaches. We proceed by describing the differences before expanding on the reasons that they are present.

First, for the exact procedure, for $k \in \{1, ..., K-1\}$, we iterate between finding $r_k$ and $a_k$ as the solutions to

$$\underset{r_k \in \mathbb{Z}}{\text{argmin}} \left[ \max_{\lambda \in \Lambda_0} R_k(\lambda, \lambda | n, \boldsymbol{a}_{k-1}, \boldsymbol{r}_k) \leq \pi_{Rk} \right]$$

and

$$\underset{a_k \in \mathbb{Z}}{\text{argmax}} \left[ \max_{\lambda \in \Lambda_1} A_k(\lambda, \lambda - \delta | n, \boldsymbol{a}_k, \boldsymbol{r}_{k-1}) \leq \pi_{Ak} \right]$$

respectively, where we may arbitrary take $\boldsymbol{a}_0 = \boldsymbol{r}_0 = 0$ since these are not used in our evaluations of $A_1(\cdot)$ and $R_1(\cdot)$.

Whilst, for the normal approximation approach, for $k \in \{1, ..., K-1\}$, we iterate between finding $r_k$ and $a_k$ as the solutions to

$$\underset{r_k \in \mathbb{R}}{\text{argmin}} \left\{ R_k(\lambda, \lambda | n, \boldsymbol{a}_{k-1}, \boldsymbol{r}_k) \leq \pi_{Rk} \right\}$$

for some arbitrarily chosen $\lambda \in \Lambda_0$, and

$$\underset{a_k \in \mathbb{R}}{\text{argmax}} \left[ A_k \left\{ \sup(\Lambda_1), \sup(\Lambda_1) - \delta | n, \boldsymbol{a}_k, \boldsymbol{r}_{k-1} \right\} \leq \pi_{Ak} \right]$$

In either case, we then utilize the formula to specify $r_K$, but not $a_K$, setting it instead as $a_K = r_K$. This ensures that we control the theoretical type-I error-rate to the desired level $\alpha$, and as discussed earlier it guarantees that a decision is made on $H_0$ during the course of the experiment.

Now, we note the reasons for the differences in the approaches utilized for the two methods. They arise because of our desire to control the type-I and type-II error-rates over the sets $\Lambda_0$ and $\Lambda_1$ respectively. Specifically, for the normal approximation approach, note that the distribution of the $\boldsymbol{T}_W$ when $\lambda_1 = \lambda_2 \in \Lambda_0$ does not depend on the specific shared value of $\lambda_1$ and $\lambda_2$. Thus, the type-I error-rate can be found for an arbitrarily chosen value of these parameters, and in turn we need only base our values $r_k$ on controlling the $R_k(\cdot)$ to below $\pi_{Rk}$ for this arbitrary value. Similarly, the theoretical type-II error-rate for the normal approximation approach when $\lambda_1 = \lambda_2 + \delta \in \Lambda_1$ is maximized when $\lambda_1 = \sup(\Lambda_1)$, since this provides the minimal possible information over the set $\Lambda_1$. Thus, if we wish to control our type-II error-rate to at most $\beta$ over all possible scenarios $\lambda_1 = \lambda_2 + \delta \in \Lambda_1$, we can simply choose our $a_k$ to constrain the $A_k(\cdot)$ to at most $\pi_{Ak}$ when $\lambda_1 = \lambda_2 + \delta = \sup(\Lambda_1)$.

In contrast, for the exact test the maximal type-I and type-II error-rates are not easy to identify. For, we may hope that an analytical formulae for the location of the maximal error-rates could be derived (e.g., by proving monotonicity of rejection probabilities across the sets $\Lambda_0$ and $\Lambda_1$). However, as we discuss further in the Supplementary Material, we believe it is unlikely that this could be achieved, given that such a result was only recently derived for the comparatively simple case of a single-arm trial with Bernoulli outcomes (Shan et al. 2017), and at least for the type-I error-rate our explorations suggest there is no simple pattern to the location of the maxima. For this reason our determination of the $a_k$ and $r_k$ retains a one-dimensional numerical search for the maximal values of $A_k(\cdot)$ and $R_k(\cdot)$ respectively. With this, however, we are guaranteed to control the error-rates to the desired level, whilst the empirical error-rates of normal approximation designs may be above their nominal levels.

The above completes our computation of the stopping boundaries for fixed $n$, $\boldsymbol{\pi}_A$ and $\boldsymbol{\pi}_R$. For any such error spending vectors, the value of $n$ that provides the desired power can then be determined as

$$\underset{n\in\mathbb{N}^+}{\operatorname{argmin}} \left\{ \beta'(n, \boldsymbol{a}, \boldsymbol{r}) \leq \beta \right\}$$

where the $\boldsymbol{a}$ and $\boldsymbol{r}$ here are those specifically derived for the particular $n$ under assessment to control the type-I error-rate, using the methods above.

Thus, by the above we are able to determine the $n$, $\boldsymbol{a}$, and $\boldsymbol{r}$ that correspond to any choice of $\boldsymbol{\pi}_A$ and $\boldsymbol{\pi}_R$. Near-optimal designs are then identified by searching over possible choices for these two error spending vectors: by searching over a large number of such vectors, an approximately exhaustive search over possible designs can be executed.

All that is additionally required to choose a design is an optimality criteria to choose amongst potential designs. Here, we use the following, which has been considered extensively in the past for similar group sequential design settings (Mander et al. 2012; Wason et al. 2012)

$$w_1 ESS(\lambda_{\text{ESS}}, \lambda_{\text{ESS}} | n, \boldsymbol{a}, \boldsymbol{r}) + w_2 ESS(\lambda_{\text{ESS}}, \lambda_{\text{ESS}} - \delta | n, \boldsymbol{a}, \boldsymbol{r}) + w_3 2Kn$$

where $ESS(\cdot)$ is a function that computes the ESS, given by

$$ESS(\lambda_1, \lambda_2 \mid n, \boldsymbol{a}, \boldsymbol{r}) = 2n \sum_{k=1}^{K} k[A_k(\lambda_1, \lambda_2 \mid n, \boldsymbol{a}_k, \boldsymbol{r}_k) + R_k(\lambda_1, \lambda_2 \mid n, \boldsymbol{a}_k, \boldsymbol{r}_k)]$$

and $\lambda_{\text{ESS}}$ is a specified mean response in arm 1. Furthermore, the $w_l \in [0, \infty), l \in \{1, 2, 3\}$, are then weights given to the different components of the optimality function. Note we should generally ensure that $w_1 + w_2 \in (0, \infty)$ as multiple designs will often have the same minimal maximal sample size, $2Kn$. For brevity in what follows, we set $\boldsymbol{w} = (w_1, w_2, w_3)$.

## 3 Results

### 3.1. Example group-sequential designs

To demonstrate the potential efficiency gains from utilizing a group-sequential design, and to compare the exact and normal approximation approaches, we consider two example trial design scenarios. The results for Example 1 are presented below, whilst those for Example 2 are included in the Supplementary Material.

We motivate the design parameters for Example 1 by considering a hypothetical obstructive sleep apnea-hypopnea (OSAH) trial. Specifically, continuous positive airway pressure is typically the first line treatment for severe sufferers of OSAH (National Institute for Health and Care Excellence 2008). However, its benefits in milder disease are less certain (Patel et al. 2003; Weaver et al. 2012), and intolerance of continuous positive airway pressure is also common (Weaver and Grunstein 2008). Therefore, we consider the design of a trial that aims to examine the efficacy of an alternative treatment option for OSAH in more moderate disease cases, such as for example the oral mandibular advancement devices that were examined by Quinnell et al. (2014) amongst others. Thus arm 1 will correspond to no treatment, and arm 2 to the new treatment option of interest.

We assume that, as is common in OSAH studies, the apnea-hypopnea index (AHI, the combined average number of apneas and hypopneas that occur per hour of sleep) will be the primary endpoint of interest. To account for variability in the mean AHI of enrolled patients on the control arm, we specify $\Lambda_0 = \Lambda_1 = [15, 30]$ (corresponding to the established definition of moderate OSAH disease). Finally, we assume that it is desired to control the type-I error-rate to $\alpha = 0.05$ and have power of $1-\beta = 0.8$ when $\delta = 2.25$ (corresponding to at least a 15% reduction in AHI for moderate disease sufferers).

We consider the optimal designs for

$$\boldsymbol{w}_1 = (1, 0, 0), \boldsymbol{w}_2 = (0, 1, 0), \boldsymbol{w}_3 = (1/2, 1/2, 0),$$
$$\boldsymbol{w}_4 = (1/2, 0, 1/2), \boldsymbol{w}_5 = (0, 1/2, 1/2), \boldsymbol{w}_6 = (1/3, 0, 1/3)$$

when $K \in \{2, 3\}$, taking $\lambda_{\text{ESS}} = 15$ as an example.

Finally, for the case $K = 2$, when utilizing the error-spending approach, we consider all combinations of $\boldsymbol{\pi}_A$ and $\boldsymbol{\pi}_R$, conforming to our requirements from earlier, with $(\pi_{A1}, \pi_{R1}) \in \{0.02, 0.04, \ldots, 0.16, 0.18\} \times \{0.005, 0.01, \ldots, 0.045\}$. Similarly, for $K = 3$ we

examine all permissible combinations with $(\pi_{A1}, \pi_{A2}, \pi_{R1}, \pi_{R2}) \in \{0.03, 0.06, 0.09, 0.12\}^2 \times \{0.01, 0.015, ..., 0.035\}^2$. Note that code to reproduce our results is available from https://github.com/mjg211/article_code.

Now, the optimal designs for $K \in \{2, 3\}$, amongst the considered $\boldsymbol{\pi}_A$ and $\boldsymbol{\pi}_R$, and using the (approximately) exhaustive search to determine optimal designs, were determined for the six stated values of $\boldsymbol{w}$. They are displayed, along with the corresponding single-stage designs based on the exact and normal approximation methods, in Table 1.

We observe that, as would be expected, utilizing a group sequential approach reduces the ESS when $\lambda_1 = \lambda_2 = \lambda_{\text{ESS}}$ and when $\lambda_1 = \lambda_2 + \delta = \lambda_{\text{ESS}}$ relative to using a single-stage design. In particular, the ESS when $\lambda_1 = \lambda_2 = \lambda_{\text{ESS}}$ can be reduced by as much as 41% and 44% when using the normal approximation or exact approaches respectively (for $K = 3$, compared to their respective required sample sizes when $K = 1$).

Moreover, as is typical for group sequential designs, increasing the value of $K$ allows us to increase efficiency further in terms of the ESS, but this comes at a cost to the maximal possible sample sizes. Interestingly though, using either design approach, there do exist designs that require only very minor increases to the maximal possible sample size that bring sizeable reductions to the considered ESSs. In particular, the optimal designs for $\boldsymbol{w}_6$, which place a non-zero weight on each of the three components of the optimality criteria, appear to perform particularly well in comparison to a single-stage approach under both $\lambda_1 = \lambda_2 = \lambda_{\text{ESS}}$ and $\lambda_1 = \lambda_2 + \delta = \lambda_{\text{ESS}}$, without requiring a substantial increase to the maximal possible sample size.

Finally, observe that the optimal design search is able to identify for each considered value of $\boldsymbol{w}$ a design that out-performs the corresponding near-optimal design. However, the difference between the operating characteristics of these designs is typically small. For example, for $\boldsymbol{w}_1$, the optimal design has $ESS(\lambda_{\text{ESS}}, \lambda_{\text{ESS}}) = 92.8$, whilst the near-optimal design has $ESS(\lambda_{\text{ESS}}, \lambda_{\text{ESS}}) = 94.6$, an increase of only 2%.

### 3.2. Empirical performance of normal approximation designs

Here, we expand on the potential problems associated with use of the normal approximation approach, utilizing simulation to assess the difference between the theoretical (estimated) error-rates of normal approximation designs in comparison to their empirical values. Specifically, to examine in detail the potential differences between the estimated and empirical values, we identified the normal-approximation error-spending designs for $\Lambda_0 = \Lambda_1 = \lambda_0 \in \{1, 1.5, ..., 9.5, 10\}$, when $\delta \in \{0.25\lambda, 0.275\lambda, ..., 0.725\lambda, 0.75\lambda\}$, $\boldsymbol{\pi}_A = (0.1, 0.1)$, and $\boldsymbol{\pi}_R = (0.025, 0.025)$. For each of these 399 designs, 100,000 simulations were then used to evaluate the empirical type-I and type-II error-rates for comparison to their values estimated via the formulae of Section 2.2.

The difference between the estimated and empirical type-I and type-II error-rates of the designs are presented in Figures 1 and 2. In general, the differences are small on the depicted raw scale, with the maximal absolute difference in type-I and type-II error-rates being 0.0029 and 0.0319 to 4 dp respectively. However, when considered as a percentage difference from the estimated value, the maximal differences are 5.8% and 24.4%. These figures, along with other considerations, impact the applicability of the normal approximation approach.

**Table 1.** The optimal two and three-stage designs for Example 1 are shown, based on the exact (optimal and near-optimal) and normal approximation approaches. For comparison, the single-stage designs are also given. Note that for brevity we write $\alpha' = \alpha'(n, \boldsymbol{a}, \boldsymbol{r}), \beta' = \beta'(n, \boldsymbol{a}, \boldsymbol{r}), ESS(\lambda_{\text{ESS}}, \lambda_{\text{ESS}}) = ESS(\lambda_{\text{ESS}}, \lambda_{\text{ESS}} \mid n, \boldsymbol{a}, \boldsymbol{r})$, and similarly for $ESS(\lambda_{\text{ESS}}, \lambda_{\text{ESS}} - \delta)$. The type-I error-rates and power figures are given to 3 dp, whilst ESSs are given to 1 dp.

| $K$ | $w$ | $\pi_A$ | $\pi_R$ | $n$ | $a$ | $r$ | $\alpha'$ | $1-\beta'$ | $ESS(\lambda_{\text{ESS}}, \lambda_{\text{ESS}})$ | $ESS(\lambda_{\text{ESS}}, \lambda_{\text{ESS}} - \delta)$ | $2Kn$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Exact: Optimal | | | | | | |
| 1 | N/A | N/A | N/A | 73 | 110 | 110 | 0.049 | 0.800 | 146.0 | 146.0 | 146 |
| 2 | $\boldsymbol{w}_1$ | N/A | N/A | 40 | (35, 108) | (141, 108) | 0.050 | 0.801 | 92.8 | 151.0 | 160 |
| 2 | $\boldsymbol{w}_2$ | N/A | N/A | 42 | (−26,138) | (92, 138) | 0.050 | 0.800 | 148.5 | 123.1 | 168 |
| 2 | $\boldsymbol{w}_3$ | N/A | N/A | 43 | (38, 126) | (98, 126) | 0.050 | 0.800 | 98.5 | 126.1 | 172 |
| 2 | $\boldsymbol{w}_4$ | N/A | N/A | 38 | (22, 109) | (127, 109) | 0.050 | 0.800 | 95.9 | 142.3 | 152 |
| 2 | $\boldsymbol{w}_5$ | N/A | N/A | 38 | (−52,118) | (101, 118) | 0.050 | 0.800 | 147.3 | 127.5 | 152 |
| 2 | $\boldsymbol{w}_6$ | N/A | N/A | 39 | (25, 113) | (109, 113) | 0.050 | 0.800 | 96.4 | 133.3 | 156 |
| | | | | | Exact: Near-optimal | | | | | | |
| 1 | N/A | 0.2 | 0.05 | 73 | 110 | 110 | 0.049 | 0.800 | 146.0 | 146.0 | 146 |
| 2 | $\boldsymbol{w}_1$ | (0.14,0.06) | (0.01,0.04) | 42 | (41, 112) | (118, 112) | 0.049 | 0.802 | 94.6 | 142.2 | 168 |
| 2 | $\boldsymbol{w}_2$ | (0.02,0.18) | (0.03,0.02) | 41 | (−8,132) | (94, 132) | 0.049 | 0.800 | 130.5 | 124.1 | 164 |
| 2 | $\boldsymbol{w}_3$ | (0.12,0.08) | (0.03,0.02) | 44 | (40, 130) | (98, 130) | 0.049 | 0.800 | 99.9 | 126.6 | 176 |
| 2 | $\boldsymbol{w}_4$ | (0.08,0.12) | (0.005,0.045) | 38 | (20, 110) | (124, 110) | 0.049 | 0.800 | 97.4 | 141.2 | 152 |
| 2 | $\boldsymbol{w}_5$ | (0.04,0.16) | (0.02,0.03) | 39 | (5, 120) | (100, 120) | 0.050 | 0.802 | 112.8 | 127.5 | 156 |
| 2 | $\boldsymbol{w}_6$ | (0.1,0.1) | (0.015,0.035) | 40 | (28, 116) | (107, 116) | 0.049 | 0.801 | 97.0 | 132.8 | 160 |
| 3 | $\boldsymbol{w}_1$ | (0.12,0.03,0.05) | (0.01,0.015,0.025) | 30 | (19, 49, 121) | (100, 125, 121) | 0.049 | 0.800 | 81.7 | 129.9 | 180 |
| 3 | $\boldsymbol{w}_2$ | (0.03,0.06,0.11) | (0.015,0.02,0.015) | 28 | (−13,45,133) | (90, 113, 133) | 0.049 | 0.800 | 101.4 | 121.3 | 168 |
| 3 | $\boldsymbol{w}_3$ | (0.12,0.03,0.05) | (0.02,0.02,0.01) | 33 | (23, 59, 144) | (92, 119, 144) | 0.049 | 0.803 | 86.0 | 122.9 | 198 |
| 3 | $\boldsymbol{w}_4, \boldsymbol{w}_6$ | (0.06,0.06,0.08) | (0.01,0.01,0.03) | 27 | (−1,47,117) | (95, 127, 117) | 0.049 | 0.801 | 88.4 | 129.1 | 162 |
| 3 | $\boldsymbol{w}_5$ | (0.03,0.06,0.11) | (0.01,0.02,0.02) | 27 | (−14,42,125) | (95, 113, 125) | 0.049 | 0.801 | 99.4 | 122.7 | 162 |
| | | | | | Normal approximation | | | | | | |
| 1 | N/A | 0.2 | 0.05 | 71 | 1.64 | 1.64 | 0.050 | 0.802 | 142.0 | 142.0 | 142 |
| 2 | $\boldsymbol{w}_1$ | (0.12,0.08) | (0.005,0.045) | 39 | (0.67,1.57) | (2.58,1.57) | 0.050 | 0.801 | 97.1 | 112.4 | 156 |
| 2 | $\boldsymbol{w}_2$ | (0.04,0.16) | (0.03,0.02) | 40 | (0.12,1.86) | (1.88,1.86) | 0.050 | 0.802 | 113.7 | 96.1 | 160 |
| 2 | $\boldsymbol{w}_3$ | (0.1,0.1) | (0.015,0.035) | 39 | (0.57,1.65) | (2.17,1.65) | 0.050 | 0.802 | 99.1 | 100.7 | 156 |
| 2 | $\boldsymbol{w}_4$ | (0.08,0.12) | (0.005,0.045) | 37 | (0.40,1.62) | (2.58,1.62) | 0.050 | 0.800 | 99.2 | 109.3 | 148 |
| 2 | $\boldsymbol{w}_5$ | (0.04,0.16) | (0.015,0.035) | 37 | (0.05,1.70) | (2.17,1.70) | 0.050 | 0.801 | 108.4 | 98.3 | 148 |
| 2 | $\boldsymbol{w}_6$ | (0.08,0.12) | (0.015,0.035) | 38 | (0.42,1.67) | (2.17,1.67) | 0.050 | 0.801 | 100.5 | 99.4 | 152 |
| 3 | $\boldsymbol{w}_1$ | (0.12,0.03,0.05) | (0.01,0.01,0.03) | 29 | (0.42,0.80,1.60) | (2.33,2.21,1.60) | 0.050 | 0.802 | 85.2 | 92.3 | 174 |
| 3 | $\boldsymbol{w}_2$ | (0.03,0.03,0.14) | (0.025,0.01,0.015) | 28 | (−0.31,0.50,1.92) | (1.96,2.13,1.92) | 0.050 | 0.803 | 103.5 | 83.6 | 168 |
| 3 | $\boldsymbol{w}_3$ | (0.09,0.06,0.05) | (0.02,0.015,0.015) | 30 | (0.28,1.06,1.78) | (2.05,2.00,1.78) | 0.050 | 0.800 | 87.8 | 86.0 | 180 |
| 3 | $\boldsymbol{w}_4, \boldsymbol{w}_6$ | (0.06,0.03,0.11) | (0.01,0.01,0.03) | 26 | (−0.05,0.52,1.69) | (2.33,2.22,1.69) | 0.050 | 0.802 | 90.9 | 89.0 | 156 |
| 3 | $\boldsymbol{w}_5$ | (0.03,0.03,0.14) | (0.015,0.01,0.025) | 26 | (−0.37,0.42,1.77) | (2.17,2.19,1.77) | 0.050 | 0.803 | 99.7 | 85.9 | 156 |

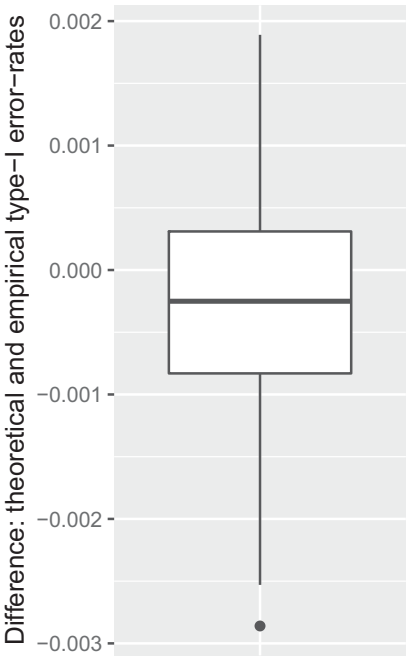


**Figure 1.** The difference between the theoretical and empirical type-I ($\lambda_1 = \lambda_2$) error-rate of exact designs is shown.

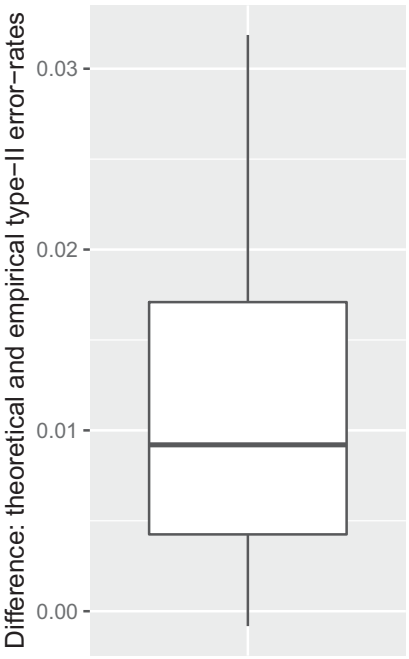


**Figure 2.** The difference between the theoretical and empirical type-II ($\lambda_1 = \lambda_2 + \delta$) error-rate of exact designs is shown.

## 4. Discussion

Here, we have described how we can use well established methodology to design group sequential experiments with Poisson distributed outcomes. Furthermore, in order to provide a method that is not dependent on asymptotic theory, we also utilized the Skellam distribution to determine operating characteristics exactly. This exact design required, in particular, careful consideration of how to evaluate the probability of stopping at each interim analysis.

Furthermore, to permit efficient group sequential designs to be identified for any value of $K$, we presented a method for design determination based on the error spending approach to group sequential design. Specifically, we searched over possible error spending vectors to find which amongst these minimized a particular optimality function. Whilst this is unlikely to identify the best possible design, we believe it is an effective means of finding efficient designs, particularly if parallelization is used to search over a large number of possible spending vectors. Indeed, from Table 1 we observed that at least for Example 1, the near-optimal designs for $K=2$ where only slightly less efficient than the optimal designs identified for an extensive search over possible values of $n$, $\boldsymbol{a}$, and $\boldsymbol{r}$.

Overall, for our considered example, we demonstrated that using a group sequential design had the potential to improve efficiency substantially; with the ESS for $\lambda_1 = \lambda_2 = \lambda_{\text{ESS}}$ in Example 1 reducible by as much as 44%. This unfortunately came at a cost of an increase to the maximal possible sample size of 23%. However, we did identify several designs that required only minimal increases to the maximal possible sample size, whilst still reducing the ESS when $\lambda_1 = \lambda_2 = \lambda_{\text{ESS}}$ and when $\lambda_1 = \lambda_2 + \delta = \lambda_{\text{ESS}}$ notably. Statistically speaking, such designs have almost no disadvantages. Furthermore, as is typical with group sequential designs, the identified efficiency gains increased with the value of $K$. Note that we presented results here for $K \in \{2, 3\}$, as our investigations have suggested the value of setting $K > 3$, in terms of reducing the ESS, is often small. This should not be interpreted as design determination for $K > 3$ being intractable.

Note that one limitation of our sequential methodology, as is typical for adaptive designs, is that it is most effective in practice when outcome accruement following inclusion in the experiment is fast relative to the entire length of the study (i.e., the enrollment rate), as we make the underlying assumption that outcome accruement completes between interim analyses, removing potential issues caused by delayed outcomes. That is, the theoretical efficiency gains are most likely to be realized in reality in this case. This is not to say, however, that a group-sequential approach cannot be useful when the study duration is short, as accruement could be paused in each stage when the required sample size has been achieved. This though would likely come at a cost to the overall time taken to complete the experiment. Thus, the utility of a group-sequential approach may often depend on the willingness to trade an increased study length for a smaller required sample size.

Moreover, we here assume that all observations are accrued following a fixed and common period of observation. This may make application in fields such as manufacturing easier than many clinical research settings where observation time may be more difficult to standardize in this manner. For scenarios with highly variable observation times, alternative methodology for group sequential design (e.g., Xia and Hoover 2007)

would likely be more appropriate. At the design stage of an experiment, however, provided only small variability in the observation times is anticipated, our methods could provide a simple means of determining the approximately required sample size.

We specified our null hypothesis in a composite form, and also powered the trial across a range of possible mean event rates. We made these choices as in many settings, in particular in clinical research, it would likely difficult to nominate single points at which to control the error-rates. This came at a cost, however, in that a method to control the maximal error-rates across the sets $\Lambda_0$ and $\Lambda_1$ was required. Whilst this was readily achieved, at least theoretically, for the normal approximation design, this was not the case for the exact approach. We thus retained one-dimensional searches for maximal stopping probabilities in our specification of the $a_k$ and $r_k$. It is important to acknowledge that this approach may in general lead to conservative designs as, for example, the value of $\lambda \in \Lambda_0$ that maximizes $R_k(\cdot)$ may not be equal across $k \in \{1, ..., K\}$. In practice, this appears to not be the case, as we were able to identify highly efficient designs with maximal error-rates close to their desired level. Nonetheless, one possible solution to reducing this conservatism could be to conduct a test conditional on the number of observed events, similar to the approaches of Xia and Hoover (2007) and Grayling et al. (2017). This remains as a possible avenue for future research.

We finish with a discussion of an important point: the inherent pros and cons of the two described design determination procedures. The exact procedure is, as discussed, preferable precisely because it is exact. That is, it guarantees control of the type-I and type-II error-rates. In contrast, the normal approximation approach may not provide such error-rate control in practice, and verifying that it does via simulation may be particularly time consuming. However, our presented simulation study does suggest that in many settings the error-rates of normal approximation designs may often be close to their nominal levels. Furthermore, from Table 1 we observe that in certain cases, for a given $\boldsymbol{w}$, the normal approximation designs have lower ESSs than the corresponding exact designs. This is a consequence of the fact that $\boldsymbol{a}, \boldsymbol{r} \in \mathbb{R}^K$, rather than $\mathbb{Z}^K$ for the normal approximation design, and so can be more precisely tailored to each possible value of $n$. In addition, provided additional simulations are not required to confirm the theoretical operating characteristics, the optimal normal approximation designs are typically faster to identify than their exact counterparts. This may lead us to conclude that the normal approximation approach may be preferred when only approximate error-rate control is desired, or when a search for an exact design would be time consuming. In contrast, if one wishes to be certain that the error-rates are controlled, or wishes to know for sure that the correct optimal or near-optimal design has been chosen, the exact approach may be preferred. Finally, in particular, whilst there does not appear to be a simple rule through which we can specify the estimated error-rates of a normal approximation design will begin to deviate from their empirical values, similar to design for Bernoulli outcome data, utilization of an exact approach would be most expedient when the requisite sample size is small (i.e., when the anticipated mean response is small and/or $\delta$ is large). Heeding this advice, group sequential tests for Poisson distributed outcome variables can then be determined which substantially reduce the requisite sample size compared to single-stage approaches.

## Funding

This work was supported by the Medical Research Council under Grant MC_UU_00002/3 to M.J.G. and A.P.M., and Grant MC_UU_00002/6 to J.M.S.W.

## ORCID

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

## References

Brent, R. 1973. *Algorithms for minimization without derivatives*. Englewood Cliffs, NJ: Prentice-Hall.

Cook, R., and J. F. Lawless. 1996. Interim monitoring of longitudinal comparative studies with recurrent event responses. *Biometrics* 52 (4):1311–23. doi:10.2307/2532846.

Cook, R., G. Yi, and K. A. Lee. 2010. Sequential testing with recurrent events over multiple treatment periods. *Statistics in Biosciences* 2 (2):137–53. doi:10.1007/s12561-010-9030-1.

Grayling, M. J., A. P. Mander, and W. J.M.S. 2017. A two-stage fisher exact test for multi-arm studies with binary outcome variables. arXiv:1711.10199v1.

Jennison, C., and B. W. Turnbull. 2000. *Group sequential methods with applications to clinical trials*. Boca Raton, FL: Chapman & Hall/CRC Press.

Jiang, W. 1999. Group sequential procedures for repeated events data with frailty. *Journal of Biopharmaceutical Statistics* 9 (3):379–99. doi:10.1081/BIP-100101183.

Johnson, N. L. 1959. On ad extension of the connexion between Poisson and $\chi^2$ distributions. *Biometrika* 46 (3/4):352–64. doi:10.2307/2333532.

Jung, S. H. 2012. *Randomized phase II cancer clinical trials*. Boca Raton, FL: CRC Press.

Lan, K. K. G., and D. L. DeMets. 1983. Discrete sequential boundaries for clinical trials. *Biometrika* 70 (3):659–63. doi:10.1093/biomet/70.3.659.

Lewis, J. W., P. E. Brown, and M. Tsagris. 2016. skellam: Densities and sampling for the Skellam distribution, 2016. Accessed August 24, 2018. https://cran.r-project.org/web/packages/skellam/.

Mander, A. P., J. M. S. Wason, M. J. Sweeting, and S. G. Thompson. 2012. Admissible two-stage designs for phase II cancer clinical trials that incorporate the expected sample size under the alternative hypothesis. *Pharmaceutical Statistics* 11 (2):91–6. doi:10.1002/sim.1600.

Mathews, P. 2010. *Sample size calculations: Practical methods for engineers and scientists*. Harbor, OH: Mathews Malnar and Bailey, Inc.

Menon, S., J. Massaro, J. Lewis, M. Pencina, Y. C. Wang, and P. Lavin. 2011. Sample size calculation for Poisson endpoint using the exact distribution of difference between two Poisson random variables. *Statistics in Biopharmaceutical Research* 3 (3):497–504. doi:10.1198/sbr.2011.10015.

Mutze, T., E. Glimm, H. Schmidli, and T. Friede. 2018a. Group sequential designs with robust semiparametric recurrent event models. *Statistical Methods in Medical Research* doi:10.1177/0962280218780538.

Mutze, T., E. Glimm, H. Schmidli, and T. Friede. 2018b. Group sequential designs for negative binomial outcomes. *Statistical Methods in Medical Research*. doi:10.1177/0962280218773115.

National Institute for Health and Care Excellence. 2008. Continuous positive airway pressure for the treatment of obstructive sleep apnoea/hypopnoea syndrome. URL: https://www.nice.org.uk/guidance/ta139.

Pampallona, S., and A. A. Tsiatis. 1994. Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference* 42 (1–2):19–35. doi:10.1016/0378-3758(94)90187-2.

Patel, S. R., D. P. White, A. Malhotra, M. L. Stanchina, and N. T. Ayas. 2003. Continuous positive airway pressure therapy for treating sleepiness in a diverse population with obstructive

sleep apnea: results of a meta-analysis. *Archives of Internal Medicine* 163 (5):565–71. doi: 10.1001/archinte.163.5.565.

Quinnell, T. G., M. Bennett, J. Jordan, A. L. Clutterbuck-James, M. G. Davies, I. E. Smith, N. Oscroft, M. A. Pittman, M. Cameron, and R. Chadwick, et al. 2014. A crossover randomised controlled trial of oral mandibular advancement devices for obstructive sleep apnoea-hypopnoea (TOMADO). *Thorax* 69(10):938–45.

Scharfstein, D. O., A. A. Tsiatis, and J. M. Robins. 1997. Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *Journal of the American Statistical Association* 92 (440):1342–50. doi:10.1080/01621459.1997.10473655.

Schultz, J. R., F. R. Nichol, G. L. Elfring, and S. D. Weed. 1973. Multiple-stage procedures for drug screening. *Biometrics* 29 (2):293–300.

Shan, G., J. J. Chen, and C. Ma. 2017. Boundary problem in Simon's two-stage clinical trial designs. *Journal of Biopharmaceutical Statistics* 27 (1):25–33. doi:10.1080/10543406.2016.1148716.

Shiue, W. K., and L. J. Bain. 1982. Experiment size and power comparisons for two-sample Poisson tests. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 31 (2):103–34. doi:10.2307/2347975.

Skellam, J. G. 1946. The frequency distribution of the difference between two poisson variates belonging to different populations. *Journal of the Royal Statistical Society: Series A (General)* 109 (3):296.

Thode, H. C. 1997. Power and sample size requirements for tests of differences between two Poisson rates. *Journal of the Royal Statistical Society: Series D (the Statistician)* 46:227–30. doi: 10.2307/2981372.

Wason, J. M. S., A. P. Mander, and S. G. Thompson. 2012. Optimal multistage designs for randomised clinical trials with continuous outcomes. *Statistics in Medicine* 31 (4):301–12. doi: 10.1111/1467-9884.00078.

Weaver, T. E., and R. R. Grunstein. 2008. Adherence to continuous positive airway pressure therapy: the challenge to effective treatment. *Proceedings of the American Thoracic Society* 5 (2): 173–8. doi:10.1513/pats.200708-119MG.

Weaver, T. E., C. Mancini, G. Maislin, J. Cater, B. Staley, J. R. Landis, K. A. Ferguson, C. F. George, D. A. Schulman, H. Greenberg, et al. 2012. Continuous positive airway pressure treatment of sleepy patients with milder obstructive sleep apnea: results of the CPAP Apnea Trial North American Program (CATNAP) randomized clinical trial. *American Journal of Respiratory and Critical Care Medicine* 186 (7):677–83. doi:10.1164/rccm.201202-0200OC.

Xia, Q., and D. R. Hoover. 2007. A procedure for group sequential comparative Poisson trials. *Journal of Biopharmaceutical Statistics* 17 (5):869–81. doi:10.1080/10543400701514015.