Chapter 6
Recent Developments in Group-Sequential Designs

James M.S. Wason

Abstract In a group-sequential trial, patients are recruited in groups, and their response to treatment is assessed. After each group is assessed, an interim analysis is conducted. At each interim analysis, the trial can stop for futility, stop for efficacy, or continue. The main advantage of group-sequential designs is that the expected number of patients is reduced compared to a design without interim analyses. There are infinitely many possible group-sequential designs to use, and the choice strongly affects the operating characteristics of the trial. This chapter discusses optimal and admissible group-sequential designs. Optimal designs minimise the expected sample size at some specified treatment effect; admissible designs optimise a weighted sum of trial properties of interest, such as expected sample size and maximum sample size. Methods for finding such designs are discussed, including a detailed description of an R package that implements a quick search procedure. Recent applications of group-sequential methodology to trials with multiple experimental treatments being tested against a single control treatment are also described.

6.1 Group-Sequential Designs Background

The traditional approach to analysing a randomised controlled trial is to conduct a statistical test of some null-hypothesis after a planned number of patients are recruited. In most disease areas, the number of patients is limited and so recruitment is generally time-consuming. Thus, data on the effect of treatment on early patients are available before recruitment is finished. A group-sequential design allows for multiple tests of the null-hypothesis as data is accrued. These earlier tests are referred to as interim analyses. The trial design may allow for early stopping if results from an interim analysis suggest the experimental treatment is significantly better than the control treatment. This is referred to as stopping for *efficacy*. The design may also allow for early stopping for *futility* if the results at an interim analysis suggest the trial is unlikely to end in success. A third reason for stopping

J.M.S. Wason (✉)
MRC Biostatistics Unit Hub for Trials Methodology Research,
Institute of Public Health, Cambridge, United Kingdom
e-mail: james.wason@mrc-bsu.cam.ac.uk

K. van Montfort et al. (eds.), *Developments in Statistical Evaluation of Clinical Trials*, DOI 10.1007/978-3-642-55345-5_6

is for safety – for example if the new treatment causes unacceptable side-effects. We just consider designs that allow stopping on the basis of whether or not the new treatment is effective, but one can also incorporate safety monitoring [5] into group-sequential designs.

The main advantages of group-sequential designs over designs that have no interim analyses (referred to as fixed sample-size trials) are:

1. Due to the possibility of early stopping, the expected sample size used in a trial will be lower than a fixed sample-size trial with the same significance level and power;
2. If the experimental treatment is less effective than the control treatment, the trial may stop early, meaning fewer patients are subjected to an ineffective experimental treatment;
3. In the long run, due to lower expected sample sizes, a limited set of patients can support more trials.

Group-sequential designs also have disadvantages:

1. More analyses means more statistical and data-management support is required;
2. Interim analyses require data to be unblinded before the end of the trial, meaning more potential for bias;
3. Since the null hypothesis is tested multiple times, the significance level of each analysis must be lower than that of the fixed sample-size trial in order to control the overall significance level; thus, if the trial continues to the end without stopping, the sample size used in the group-sequential trial will be larger than the fixed sample size trial.

Group-sequential designs are less useful when the outcome of interest takes a long time to observe, since recruitment will often be completed before the data on the effect of treatment on early patients are available. In settings where the treatment outcome is observed relatively quickly, the efficiency and ethical advantages of group-sequential designs are generally thought to outweigh the disadvantages.

In this chapter we will restrict attention to one-sided group-sequential designs. These are used when the null-hypothesis is tested against a one-sided alternative hypothesis. One-sided group-sequential tests are more relevant in clinical trials, as the experimental treatment is generally not of interest if it is worse than the control treatment.

A one-sided group-sequential design is parametrised by: (1) the number of patients to be recruited at each stage; (2) the futility boundaries, determining the threshold for futility stopping at each analysis; and (3) the efficacy boundaries, determining the thresholds for efficacy stopping at each analysis. The constraints on the design are the overall type-I error rate and power of the design. Since there are more parameters than constraints, there are an infinite number of possible designs to choose from. The choice of design is extremely important as it affects the statistical properties of the design, such as expected sample size.

There are three main approaches to choosing a design. The first is to constrain the stopping boundaries using some shape function. Commonly used functions are

those of Pocock [22], O'Brien and Flemming [20], and Whitehead [33]. The main advantage of this approach is that it is quick to find a design; the main disadvantage is that the properties of the design, such as expected sample size, may not be desirable for the investigator. A second approach is to use a more flexible family of stopping boundary functions. For example, the power-family of group-sequential tests, proposed by Pampallona and Tsiatis [21], is a single-parameter family of stopping boundary shapes. By varying the parameter, the properties of the resulting design differ. A third approach, is to search over the full set of parameters in order to choose the design that best matches the desired properties of the investigator.

This chapter provides an overview of some recent methodological developments in group-sequential designs, and is organised as follows: in Sect. 6.2, notation for the rest of the chapter is given; in Sect. 6.3 some background on optimal designs is provided; in Sect. 6.4 the δ-minimax design is motivated, and a simulated annealing technique to find optimal designs is discussed; in Sect. 6.5 the concept of admissible designs is motivated and discussed; in Sect. 6.6 the problem of not knowing the variance of the treatment response at the design stage is addressed; in Sect. 6.7 an R package which allows quick finding of admissible designs is described; in Sect. 6.8, extensions of group-sequential methods to multi-arm multi-stage designs are discussed; finally in Sect. 6.9, some limitations and possible extensions of the methods in the chapter are discussed.

6.2 Notation

Consider a randomised two-arm group-sequential design with up to J analyses. The jth analysis takes place after n_j patients have been randomised to each arm, and their treatment response measured. The response of patient i on the control arm, X_{0i}, is assumed to be distributed as $N(\mu_0, \sigma^2)$, with the response of patient i on the experimental arm, X_{1i}, is assumed to be distributed as $N(\mu_1, \sigma^2)$. Here, the value of σ^2 is assumed to be known, although unknown variance will be addressed in Sect. 6.6. The parameter of interest is the difference in mean response between the experimental and control arms, $\mu_1 - \mu_0$, and is labelled δ. The null-hypothesis tested is $H_0 : \delta \leq \delta_0$. A design is required such that the probability of rejecting the null is at most α when H_0 is true, and at least $1 - \beta$ when $\delta \geq \delta_1$, where δ_1 is the clinically relevant difference (CRD). The value of δ_0 will generally be set to 0, indicating that any improvement is of interest. These two constraints are referred to as the type-I error and power constraints respectively. A design which meets both constraints is called *feasible*.

At a given interim analysis j, the z-statistic for testing H_0, Z_j, is calculated:

$$Z_j = \sqrt{\frac{n_j}{2\sigma^2}} \frac{\sum_{i=1}^{n_j} X_{i1} - \sum_{i=1}^{n_j} X_{i0}}{n_j} . \tag{6.1}$$

If $Z_j > e_j$, the trial stops for efficacy; if $Z_j \leq f_j$, the trial stops for futility. If it is between the two thresholds, the trial continues to stage $j+1$. The value of e_J is set to f_J to ensure that a decision is made at the last interim analysis.

The number of design parameters is $3J-1$: J parameters for the sample size at each stage, J efficacy parameters $e = (e_1, \ldots, e_J)$, and J futility parameters $f = (f_1, \ldots, f_{J-1})$ (actually $J-1$ free parameters as $f_J = e_J$). Generally the number of parameters is reduced by assuming a constant number of patients recruited per stage to each treatment arm, n, called the *group-size*. With this assumption, the value of n_j will be equal to jn. This reduces the number of parameters to $2J$.

The vector of random variables $(Z_1, Z_2, \ldots, Z_J)$ has a multivariate normal distribution with mean vector $\left(\sqrt{\frac{n}{2\sigma^2}}\delta, \sqrt{\frac{2n}{2\sigma^2}}\delta, \ldots, \sqrt{\frac{Jn}{2\sigma^2}}\delta\right)$, and covariance matrix Σ, where the (i,j)th entry of Σ, Σ_{ij}, is equal to $\sqrt{\frac{\min(i,j)}{\max(i,j)}}$, [31]. Finding the probability of stopping for efficacy at stage j, Π_j, involves multivariate integration. Stopping for efficacy at the jth stage happens if and only if $(Z_1, \ldots, Z_{j-1})$ were all between the futility and efficacy stopping boundaries, and Z_j is above e_j. The probability of this is:

$$\Pi_J(\delta) = \int_{f_1}^{e_1} \int_{f_2}^{e_2} \ldots \int_{f_{j-1}}^{e_{j-1}} \int_{e_j}^{\infty} f_{Z_{(j)}}(z_1, \ldots, z_j) dz_j \ldots dz_1 \,, \tag{6.2}$$

where $f_{Z_{(j)}}$ is the pdf of $(Z_1, \ldots, Z_j)$. Note that the mean of $Z_{(j)}$ depends on δ, but the covariance does not. Equation (6.2) can be evaluated using the technique of Genz and Bretz [10], or the technique of Armitage [2, 18], described further in Chap. 19 of Jennison and Turnbull [13]. Note that the normality of the test statistics is the main assumption used and not the normality of the treatment endpoint – therefore other types of endpoints such as binary and time-to-event for which there are asymptotically normally distributed test statistics can be considered within this framework [13].

The probabilities $\Pi_1(\delta), \ldots, \Pi_J(\delta)$ can be summed to give the total probability of stopping for efficacy. Setting $\delta = \delta_0$ will give the type-I error rate, and setting $\delta = \delta_1$ will give the power. A similar formula as (6.2) can be used to find the probability of stopping for futility at each stage. From the probabilities of stopping for futility and efficacy at each stage, the expected sample size can be straightforwardly found.

6.3 Optimal Group-Sequential Designs

Within the context of group-sequential designs, an optimal design is one that satisfies the required type-I error rate and power (i.e. it is *feasible*), and out of all possible feasible designs, it optimises some criterion of interest. Criteria considered tend to be some function of the sample size, for example the expected sample size at some value of δ.

Finding an optimal design involves searching over the stopping boundary parameters as well as the sample size parameters. With the constraints described in Sect. 6.2, searching for an optimal J-stage group-sequential design involves searching over $2 \times J$ parameters, as the final futility and efficacy threshold are set to be the same. There are just two constraints: the type-I error and the power. This is a computationally challenging problem when $J > 2$, as the number of parameters is large and there are many local optima in the set of designs to be searched.

The method of dynamic programming was proposed for finding symmetric (i.e. the type-I error rate, α, is equal to the type-II error rate, β) optimal one-sided group-sequential designs [7] and optimal two-sided designs [8]. This was extended to non-symmetric designs by Barber and Eales [3]. The method works by defining a Bayes decision theory problem for which the optimal group-sequential design is the solution. The decision theory problem is to decide between $D_0 : \delta = 0$ and $D_\delta : \delta = \delta^*$, with the cost of making decision D with true treatment difference δ equal to $C(D_0, \delta^*) = d_\delta$ for $D = D_0$, and $C(D_\delta, 0) = d_0$ for $D = D_\delta$. For any other value of δ, $C(D, \delta)$ is set to be 0. Backwards induction can be used to find the design that minimises a given objective function, such as expected sample size at the null hypothesis. A numerical search over (d_0, d_δ) is conducted in order to find the design giving the correct type-I error rate and power. This final design will then be the optimal one.

Generally this method can be used to find an optimal design when the optimality criterion is the expected sample size at a specific value of δ (or sums of expected sample sizes at different values of δ). However, in the next section an optimality criteria is proposed that is of potential interest and that cannot be optimised using dynamic programming.

6.4 δ-Minimax Design and Simulated Annealing

The expected sample size of a group-sequential design depends on the true treatment effect. If an optimal design is chosen for a particular treatment effect, then the design may perform poorly when the true treatment effect varies from the design value. For designs allowing stopping for both futility and efficacy, the expected sample size increases in δ monotonically to a maximum and then decreases monotonically. Intuitively this is because as δ increases, the probability of the trial stopping early for futility decreases monotonically, but the probability of the trial stopping early for efficacy increases monotonically. A slightly more formal explanation is given in Wason, Mander and Thompson [31].

Thus each design has a treatment effect, $\tilde{\delta}$, that leads to the design having the maximum expected sample size over all possible values of δ. This is called the worst-case-scenario treatment effect. The optimality criterion of choosing the feasible design with the lowest maximum expected sample size was proposed for two-stage trials with binary outcomes by Shuster [25]. The design showed some good properties such as low expected sample sizes at the null treatment effect

and CRD. The design was extended to two-stage trials with normally distributed outcomes by Wason and Mander [30] and named the δ-minimax design, as it has the lowest maximum expected sample size over δ. To find the δ-minimax design for two-stages, it is feasible to use a grid-search technique, as the number of parameters (i.e. futility and efficacy boundary parameters, and group-size) is low. For more than two-stages, there are too many parameters to perform a grid-search. The dynamic programming algorithm proposed by Barber and Jennison [3] works when the optimality criterion is independent of the design; however the value of $\tilde{\delta}$ depends on the design, thus a different method must be used for $J > 2$. In Wason et al. [31], use of a stochastic search technique called simulated annealing was proposed to find the δ-minimax design.

The simulated annealing algorithm is described in detail in the supplementary material of Wason et al. [31], and C code is available on the author's website (http://sites.google.com/site/jmswason). Each iteration of the simulated annealing process consists of two steps. The first step is to generate a new candidate design from the current design (i.e. the design which the process is currently at). The second step is to decide whether the process should move from the current design to the candidate design. Both steps rely on so-called 'temperature' parameters. At the end of each iteration, the temperature parameters are reduced. As the temperature parameters fall: (1) the candidate design generated at each iteration will, on average, be closer to the current design; and (2) the process is less likely to move to a design that is worse. In this way, the process is more likely to explore the space of designs towards the beginning, with the aim of avoiding getting stuck at a local optimum.

For two and three-stage designs, simulated annealing is quick and reliable, with results not varying considerably between independent runs. However, for four or more stages, the process takes longer and becomes less reliable. The reason that it takes longer is that evaluating the operating characteristics of a design is more time-consuming when there are more stages. The process is less reliable because the number of parameters is greater and there are more local optima in the space of possible designs. One can run the simulated annealing process for longer in order to improve reliability, but of course this takes longer. With four or five stages, it is recommended that a number of independent simulated annealing processes with different random number seeds are run. The best resulting design can then be chosen.

The δ-minimax design is comparable to the triangular design proposed by Whitehead and Stratton [33]. In the case of a *symmetric* ($\alpha = \beta$) and fully sequential (i.e. interim analyses after each patient), as the type-I error rate converges to 0, the resulting triangular stopping boundaries minimise the maximum expected sample size. It is thus of interest to see whether the δ-minimax design adds anything over the use of the triangular stopping boundaries. Table 6.1 shows, for $\delta_0 = 0$, $\delta_1 = 1$, $\sigma = 3$ and various values of J, the expected sample size of the null-optimal design (optimal at $\delta = \delta_0$), the CRD-optimal design (optimal at $\delta = \delta_1$), the δ-minimax design, and the triangular test at: (1) the null treatment effect, i.e. 0; (2) the CRD, i.e. 1; (3) the worst-case-scenario treatment effect. Also shown is the maximum sample size used by the design if early stopping does not take place.

Table 6.1 Expected and maximum sample sizes per arm of investigated designs for different numbers of stages. The random variable N denotes the sample size per arm used with a specified design

		Null-optimal	CRD-optimal	δ-minimax	Triangular design
J = 2	$\mathbb{E}(N\|\delta = \delta_0)$	107.6	118.0	110.9	111.2
	$\mathbb{E}(N\|\delta = \delta_1)$	130.5	117.1	119.4	117.6
	$\mathbb{E}(N\|\delta = \tilde{\delta})$	138.9	136.8	133.3	132.2
	Maximum sample size	170	172	180	180
J = 3	$\mathbb{E}(N\|\delta = \delta_0)$	94.9	105.7	98.0	100.4
	$\mathbb{E}(N\|\delta = \delta_1)$	128.9	107.0	109.2	108.4
	$\mathbb{E}(N\|\delta = \tilde{\delta})$	137.3	130.0	125.9	125.5
	Maximum sample size	183	186	189	192
J = 4	$\mathbb{E}(N\|\delta = \delta_0)$	88.7	98.0	92.7	98.3
	$\mathbb{E}(N\|\delta = \delta_1)$	119.1	102.2	105.0	106.1
	$\mathbb{E}(N\|\delta = \tilde{\delta})$	130.6	125.5	122.0	124.9
	Maximum sample size	192	196	196	204
J = 5	$\mathbb{E}(N\|\delta = \delta_0)$	85.4	92.1	89.2	96.0
	$\mathbb{E}(N\|\delta = \delta_1)$	113.1	99.3	102.8	103.9
	$\mathbb{E}(N\|\delta = \tilde{\delta})$	126.8	122.5	119.6	123.0
	Maximum sample size	200	210	205	210

Table 6.2 Group-size, futility stopping boundaries, and efficacy stopping boundaries of five-stage optimal and triangular designs

Design	n	f	e
Null-optimal	40	$(-0.24, 0.37, 0.76, 1.09, 1.56)$	$(3.01, 2.47, 2.23, 2.03, 1.56)$
CRD-optimal	42	$(-0.51, 0.29, 0.83, 1.33, 2.05)$	$(2.14, 2.05, 2.09, 2.15, 2.05)$
δ-minimax	41	$(-0.52, 0.34, 0.92, 1.38, 1.83)$	$(2.54, 2.09, 2.03, 1.96, 1.83)$
Triangular	42	$(-0.85, 0.30, 0.98, 1.49, 1.90)$	$(2.55, 2.10, 1.96, 1.91, 1.90)$

When $J = 2$ or 3, the δ-minimax and triangular designs have very similar expected sample size properties. The δ-minimax design in fact has a higher maximum expected sample size for $J = 2, 3$, but this is because the equations determining the triangular design, given in Jennison and Turnbull [13], for given $\alpha \neq \beta$ do not result in the feasibility constraints being met exactly (the triangular design has $\alpha = 0.0517$ and 0.0512 for $J = 2$ and $J = 3$ respectively). For $J = 4$ and 5 the δ-minimax design is more distinct, having a 7.1 % reduction in expected sample size under the null, and a 5.7 % reduction for $J = 4$, compared to the triangular design.

Table 6.2 shows the design parameters for each five-stage design, and Fig. 6.1 shows the stopping boundaries of the three optimal designs graphically. Although the expected sample size patterns are similar, the stopping boundaries of the δ-minimax and triangular designs are somewhat different. Generally the δ-minimax design is marginally more likely to stop at the first stage, although this is balanced

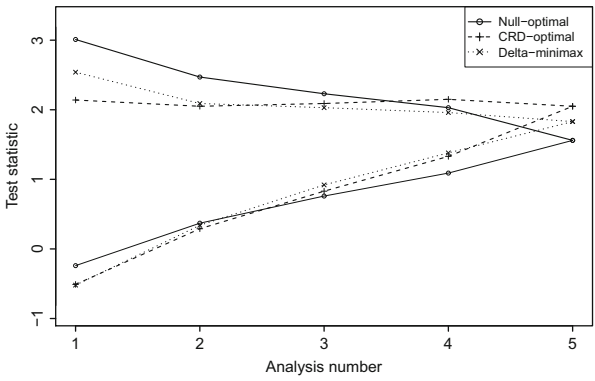

Fig. 6.1 Futility and efficacy stopping boundaries, in terms of test statistics, of the null-optimal, CRD-optimal, and δ-minimax design for $\alpha = 0.05, \beta = 0.1, \sigma = 3, \delta_0 = 0, \delta_1 = 1$

by it being slightly less likely to stop once the trial is at a later stage. The maximum sample sizes are similar, but differ between the designs for some values of J (see Table 6.1).

The δ-minimax design has desirable properties in comparison to the other two optimal designs. By definition it has the lowest maximum expected sample size of the three designs, but it also has low expected sample sizes across the range of treatment effects considered. When the treatment effect is close to δ_0, its expected sample size is only slightly higher than that of the null-optimal design; similarly its expected sample size is only slightly higher than that of the CRD-optimal design when δ is close to δ_1. The optimal designs perform well when δ is close to the treatment effect for which they are optimal, but poorly when δ is different. As one would expect, the expected sample size curves shifts downwards as J increases, indicating that including more stages results in lower expected sample sizes at each value of δ. The relative shapes of the curves change slightly, especially as δ increases past δ_1.

Minimising the expected sample size is an important objective in trials, but it is also of interest to control the maximum potential sample size. A design which yields a small improvement in expected sample size at a cost of a large increase in maximum sample size is unlikely to be preferred in practice. Table 6.1 shows that the δ-minimax and triangular designs generally have larger maximum sample sizes compared to null-optimal and CRD-optimal designs. All the optimal designs have maximum sample sizes noticeably larger than the sample size required for the one-stage design (155).

6.5 Admissible Designs

Optimal designs tend to have large maximum sample sizes, which can be problematic for planning individual trials. In addition, they may perform poorly with respect to other criteria of interest. For example, the null-optimal design has

a relatively high maximum expected sample size. Admissible designs have been proposed in order to balance over more than one criteria of interest.

The first work on admissible designs was in the context of two-stage trial designs with binary outcomes and only futility stopping allowed. These designs have been well studied in the literature due to their relative simplicity and the fact that all possible designs can be enumerated (as sample size and stopping boundary parameters are all integers). Simon [26] discussed and recommended two designs for this type of trial. The first was the 'optimal' design (in the terminology of Sect. 6.4, the null-optimal design). The second was the 'minimax' design, which chooses the design with the lowest expected sample size at the null out of all designs that have the lowest maximum sample size. Jung et al. [14] noted that the optimal design has a relatively large maximum sample size, and the minimax design has a relatively large expected sample size. These observations motivated investigation of 'admissible' designs, which would balance the two criteria.

To do this, the authors specified a loss function as the weighted sum of the expected sample size under the null treatment effect and the maximum sample size: $\omega\mathbb{E}(N|H_0) + (1-\omega)\max(N)$, for $\omega \in [0,1]$. Admissible designs are feasible designs that minimise the loss function for some value of w. Additional information is available in Jung et al. [14] about how this corresponds to admissible decision rules in Bayesian decision theory. The optimal and minimax designs are admissible (for $\omega = 1$ and 0 respectively), but other admissible designs also exist which balance the two quantities in different ways. Admissible designs exist that show very small increases in expected sample size compared with the optimal design, but large decreases in the maximum sample size. In practice, such a design may be preferable to the optimal design, as a small maximum sample size is desirable.

Mander et al. [17] extend the ideas in Jung et al. to phase II trials with binary outcomes allowing early stopping for efficacy. When stopping for efficacy is allowed, the expected sample sizes at treatment effects other than the null are also of interest. Designs that are admissible with respect to the expected sample size at the null, the expected sample size at the CRD, and the maximum sample size are evaluated.

When considering normally distributed endpoints, finding admissible designs is more challenging. This is because the stopping boundary parameters are non-integer and so infinitely many feasible designs exist. This is as opposed to the binary outcome case where the stopping boundary parameters are integers, and so all designs can be enumerated. Instead, in Wason et al. [31], it was argued that the maximum expected sample size could be used as a surrogate for all expected sample sizes of interest. The loss function in this case is:

$$\omega\mathbb{E}(N|\tilde{\delta}) + (1-\omega)\max(N)\,. \tag{6.3}$$

The advantage of just considering the two criteria in (6.3) is that it is computationally feasible to find all admissible designs. For each possible maximum sample size, the futility and efficacy parameters can be chosen so that the maximum expected sample size is minimised. Any other design with that maximum sample

Table 6.3 Properties of admissible designs for $J = 3$; $\max(N) =$ maximum sample size per arm, ω interval gives the values of ω that would lead to that design being the admissible design of choice

$\max(N)$	$\mathbb{E}(N\|\delta_0)$	$\mathbb{E}(N\|\delta_1)$	$\mathbb{E}(N\|\tilde{\delta})$	ω interval
156	117.29	124.73	139.02	[0,0.426)
159	107.34	121.43	134.97	[0.426,0.539)
165	102.05	114.78	129.83	[0.539,0.713)
168	101.44	112.55	128.62	[0.713,0.820)
171	100.21	111.23	127.96	[0.820,0.843)
177	98.19	111.14	126.84	[0.843,0.921)
186	98.74	109.19	126.07	[0.921,0.981)
189	98.20	109.88	126.01	[0.981,1]

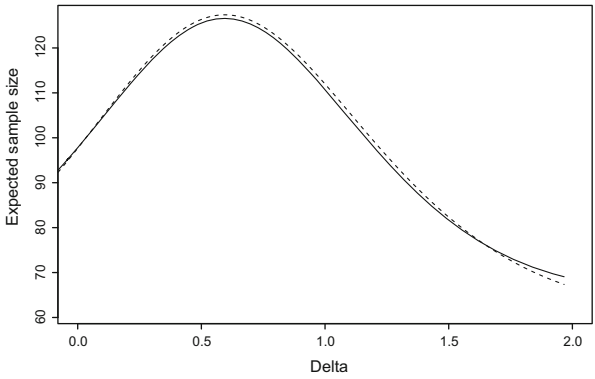

Fig. 6.2 Expected sample sizes of δ-minimax design (*solid line*) and admissible design with $N = 171$ in Table 6.3 (*dashed line*)

size cannot be admissible because loss function (6.3) will always be higher (unless $\omega = 0$). No design with maximum sample size greater than that of the δ-minimax design can be admissible, as such a design would have both a higher maximum sample size and a higher maximum expected sample size.

As an illustration, Table 6.3 displays the properties of the possible admissible designs for $\delta_0 = 0, \delta_1 = 1, \sigma = 3, \alpha = 0.05, 1 - \beta = 0.9$.

From Table 6.3, using the value of $\max(\mathbb{E}(N))$ as an admissibility criterion is a good surrogate for jointly considering $\mathbb{E}(N|\delta_0)$ and $\mathbb{E}(N|\delta_1)$, since the two latter quantities generally decrease as the former does. The table includes the range of ω's (i.e. the weighting put on the maximum expected sample size) for which each design is best. For instance, if the two quantities are each given equal weight ($\omega = 0.5$), the second design in the table is the best one to pick. The choice of ω may depend on several factors. For instance, if the trial is being carried out in an area with limited patient numbers, ω might be chosen to be low, since it would be desirable to reduce the maximum sample size. In other situations, a higher value of ω may be preferred, since on average the number of patients required is reduced.

Figure 6.2 shows the expected sample size curve of the δ-minimax design for a range of values of δ. Also included is the expected sample size curve for the admissible design from Table 6.3 with $\max(N) = 171$. The difference in the

expected sample size curves is very small, but there is a 9.5 % reduction in the maximum sample size. This indicates that by relaxing the requirement for optimality very slightly, a big improvement in other characteristics of interest is possible.

6.6 Unknown Variance

A common assumption made in the design of group-sequential trials is that the variance of the treatment response, σ^2 is known for each arm. In practice this is unlikely to be the case, and if the postulated value is incorrect, then the operating characteristics of the trial can be strongly affected.

Various techniques to allow for unknown variance have been proposed in the literature. Shao and Feng [24] suggest using Monte-Carlo simulation to choose an appropriate critical value. Although this technique would be too computationally intensive to be used in conjunction with a search for optimal designs, it could be used to modify the final design's stopping boundaries. Jennison and Turnbull [12] show how one can convert boundaries for the known variance case to the unknown variance case using a recursive algorithm.

Jennison and Turnbull [13] propose a method for converting the stopping boundaries that is simpler than the recursive algorithm and less computationally intensive than simulation. Recall that f_j and e_j are the stopping boundaries for analysis j, and n_j is the number of patients per arm that are randomised by the time of the analysis. Then the thresholds for stopping in terms of p-values are attained from the quantile of the normal distribution, i.e. $1 - \Phi(e_j)$ and $1 - \Phi(f_j)$ respectively. With unknown variance, when $\delta = 0$, the test-statistics would be marginally distributed as a Student's t-distribution with $2n_j - 2$ degrees of freedom. Therefore by substituting in new stopping boundaries $f'_j = T_{2n_j-2}(1 - \Phi(f_j))$ and $e'_j = T_{2n_j-2}(1 - \Phi(e_j))$, where T_p is the cumulative distribution function of Student's t-distribution with p degrees of freedom, the design will marginally have the correct stopping characteristics (under the null) at each stage. The overall type-I error rate of the trial will still differ from its nominal value because the assumed correlation between test-statistics when the variance is known will differ from the actual correlation when it is unknown (the size of the difference is investigated later on in this section).

Table 6.4, taken from Wason et al. [31], shows the type-I error rate and power for the five-stage δ-minimax design for $\delta_1 = 1, \sigma = 3, \alpha = 0.05, \beta = 0.1$ as the true value of σ differs from 3. Three scenarios are considered: (1) no modification is made, (2) t-tests are used with the known-variance stopping boundaries, (3) t-tests are used with the stopping boundaries modified using the quantile-substitution method. The type-I error rate and power are estimated from 250, 000 independent replicates each.

The simulated type-I error rates show that methods (2) and (3) both work well. The type-I error rates are very close to the required level of 0.05, with quantile-substitution working slightly better. The power is not controlled as the value of

Table 6.4 Type-I error rate and power estimates as the true standard deviation varies from the assumed value of 3

	Type I error			Power		
σ	Z-test	T-test	T-test with modified boundaries	Z-test	T-test	T-test with modified boundaries
1	0.000	0.051	0.050	1.000	1.000	1.000
1.5	0.000	0.052	0.050	0.998	1.000	1.000
2	0.000	0.051	0.050	0.984	0.995	0.995
2.5	0.021	0.052	0.050	0.95	0.965	0.965
3	0.050	0.051	0.050	0.900	0.900	0.899
3.5	0.086	0.052	0.050	0.851	0.810	0.809
4	0.124	0.052	0.051	0.807	0.714	0.712
4.5	0.158	0.052	0.051	0.768	0.626	0.623
5	0.189	0.051	0.050	0.737	0.550	0.547

Table 6.5 Type-I error rate and power estimates as the true standard deviation varies from the assumed value of 1

	Type I error			Power		
σ	Z-test	T-test	T-test with modified boundaries	Z-test	T-test	T-test with modified boundaries
0.25	0.000	0.070	0.054	1.000	1.000	1.000
0.5	0.000	0.069	0.052	0.997	1.000	1.000
0.75	0.011	0.069	0.053	0.964	0.986	0.985
1	0.050	0.069	0.052	0.900	0.902	0.893
1.25	0.102	0.068	0.052	0.832	0.768	0.750
1.5	0.154	0.069	0.052	0.775	0.64	0.613
1.75	0.201	0.069	0.052	0.726	0.533	0.503
2	0.236	0.069	0.052	0.691	0.455	0.424

σ increases however. To overcome this, an adaptive design would be required in which the sample size of the rest of the trial is chosen depending on the estimated variance; an example of this is given in Whitehead et al. [34]. The good performance of both methods (2) and (3) could be due to the large group-size resulting in the degrees of freedom of the t-distribution being sufficiently high to allow the standard normal to be a good approximation. To see what happens when the group-size is lower, results are shown for the five-stage δ-minimax design with $\sigma = 1$. This results in a group-size of 4, $f = (-0.914, -0.026, 0.698, 1.177, 1.761)$, and $e = (2.980, 2.308, 2.048, 1.976, 1.761)$. It is clear that the type-I error rate is less well controlled in this case (Table 6.5), although the T-test in conjunction with the quantile-substitution method controls the type-I error rate fairly well.

Thus it seems that quantile substitution is a straightforward, but effective method to control the type-I error rate when the variance is unknown.

6.7 OptGS: An R Package for Optimal and Admissible Group-Sequential Designs

The consideration of optimal and admissible group-sequential designs has been motivated in the previous sections. All the theory to implement finding such designs is available in the literature, but it takes a lot of work to implement from scratch. There are some existing software packages that implement group-sequential designs, summarised by Wassmer and Vandemeulebroecke [32]. The IML module in SAS® contains routines that allow calculation of stopping boundaries that give a specified type-I error rate. In R, the package gsDesign [1] allows the user to find boundaries and group-size required for several group-sequential designs, including O'Brien-Flemming and Pocock. Commercially available stand-alone programs that implement group-sequential designs include ADDPLAN, East, PASS, and PEST. However, none of these software packages include a function that searches for optimal or admissible designs.

FORTRAN code that implements searching for optimal designs using dynamic programming, as described in Barber and Jennison [3] is available from Stuart Barber's webpage (http://www1.maths.leeds.ac.uk/~stuart/Research/Software/0118.tar). Compilation of the code requires some technical computing knowledge, as it requires installation of the GNU Scientific Library. The code would also not be extendable to all optimality criterion, for example the maximum expected sample size. In this section, the R package OptGS [28] is described, which is freely available from the author's website (http://sites.google.com/site/jmswason). The package allows quick searching for designs that are near-optimal, or admissible with respect to four optimality criteria. Instead of simulated annealing, an extension of the Power-family is used. This extension allows a wide range of stopping boundary shapes, but considerably reduces the time taken to search. A quick method for searching is desirable so that investigators may explore many possible admissible designs in a short time.

6.7.1 Two-Parameter Power Family

The power family of group-sequential tests was first proposed by Emerson and Flemming [9] for symmetric designs (i.e. $\alpha = \beta$). Pampallona and Tsiatis [21] extended the family to allow non-symmetric designs ($\alpha \neq \beta$). In this section we consider the formulation of Pampallona and Tsiatis. The family is indexed by a parameter Δ, which determines the shape of the stopping boundaries. The power-family stopping boundaries are:

$$e_j = C_e(J, \alpha, \beta, \Delta)(j/J)^{\Delta - 0.5}$$

$$f_j = \delta_1 \sqrt{\mathscr{I}_j} - C_f(J, \alpha, \beta, \Delta)(j/J)^{\Delta - 0.5},$$

where $\mathscr{I}_j = 2n_j/\sigma^2$.

To meet the required constraint $e_J = f_J$, the value of $\mathscr{I}_J$ is set to:

$$\mathscr{I}_J = 2n_J/\sigma^2 = \frac{\left\{C_e(J,\alpha,\beta,\Delta) + C_f(J,\alpha,\beta,\Delta)\right\}^2}{\delta^2} . \tag{6.4}$$

For a specific value of Δ, $C_f(J,\alpha,\beta,\Delta)$ and $C_e(J,\alpha,\beta,\Delta)$ take values such that the design has correct type-I error rate and power. Varying Δ changes the shape of the boundaries, and thus the operating characteristics of the design, with higher values generally giving designs with lower expected sample sizes, but higher maximum sample sizes.

Although the power-family provides a flexible range of stopping boundary shapes, it does not provide enough flexibility to include optimal designs. For optimal designs, the shape of the efficacy stopping boundaries will differ from the shape of the futility stopping boundaries.

OptGS uses a straightforward extension to the power family: introducing two shape parameters Δ_f and Δ_e, allowing the shape of the futility and efficacy boundaries to differ, and thus allowing greater flexibility in shape. The stopping boundaries are:

$$\begin{aligned} e_j &= C_e(J,\alpha,\beta,\Delta)(j/J)^{\Delta_e - 0.5} \\ f_j &= \delta_1\sqrt{\mathscr{I}_j} - C_f(J,\alpha,\beta,\Delta)(j/J)^{\Delta_f - 0.5} . \end{aligned} \tag{6.5}$$

Note that Eq. (6.4) still ensures $e_J = f_J$.

Given values of $(J, \Delta_f, \Delta_e, C_f, C_e)$, the group-size and stopping boundaries are determined from (6.4) and (6.5). As in Pampallona and Tsiatis [21], for each value of (Δ_f, Δ_e), values of C_f and C_e exist so that the design has desired type-I error rate, α, and power, $1-\beta$. These values can be found by searching for the values of (C_f, C_e) that minimise the following function:

$$(\alpha^*(J,\Delta_f,\Delta_e,C_f,C_e) - \alpha)^2 + (\beta^*(J,\Delta_f,\Delta_e,C_f,C_e,\delta) - \beta)^2 , \tag{6.6}$$

where $\alpha^*(\cdot)$ and $\beta^*(\cdot)$ are the type-I and type-II error rate of the design given by $(J,\Delta_f,\Delta_e,C_f,C_e)$. The value of (6.6) is 0 if and only if the type-I error rate and power of the design are equal to the required values. In OptGS, this minimisation is performed using the Nelder-Mead algorithm [19].

The Nelder-Mead algorithm is also used to search over values of (Δ_f, Δ_e) in order to find an optimal design. Almost surely, the optimal value of (Δ_f, Δ_e) will imply a non-integer group size. To get a final design with integer group-size, two additional optimisations are run. The first with the constraint that the final group-size is equal to the group-size implied by the optimal (Δ_f, Δ_e) rounded up. The second instead rounding down. Of the designs found, the one that is closer to optimal is picked as the final design. Additional details are provided in Wason [28].

OptGS allows the user to find a design that balances the three optimality criteria discussed in Sect. 6.4 as well as the maximum sample size. A vector of weights, $(\omega_1, \omega_2, \omega_3, \omega_4)$, is specified by the user such that all are non-negative. Then the feasible design is found that minimises the following function:

$$\omega_1 \mathbb{E}(N|\delta = \delta_0) + \omega_2 \mathbb{E}(N|\delta = \delta_1) + \omega_3 \max \mathbb{E}(N) + \omega_4 J n_1 \ . \tag{6.7}$$

This design balances the three optimality criteria together with the maximum sample size. Note that one of ω_1, ω_2, and ω_3 must be strictly positive, because an infinite number of designs will exist with the lowest maximum sample size.

6.7.2 Comparison of OptGS and Simulated Annealing

Table 6.6, taken from Wason [28], shows the time taken to find J-stage null-optimal designs using SA and using OptGS. A single M5000 SPARC 2.4 GHz processor was used to carry out all computation. Ten independent simulated annealing (SA) searches were carried out for each value of J because SA is a stochastic process, and results may vary between runs. The average and minimum expected sample size under the null over the ten processes are shown in the table.

For several values of J, the optimal design found from OptGS is actually better than that found from the best of 10 runs of SA. This is despite the shape constraint imposed by use of the extended power-family. Only for $J = 5$ does SA show some improvement over OptGS. OptGS is substantially faster than even one SA run. Clearly, using OptGS has substantial advantages over using simulated annealing.

Table 6.7, also taken from Wason [28], shows the optimal values of $\Delta_f, \Delta_e, C_f, C_e$ for the three types of optimal design implemented in OptGS as well as the $(1, 1, 1, 1)$-admissible design, i.e. the admissible design that puts equal weight on all four operating characteristics. The results show that allowing Δ_f to differ from Δ_e is necessary to allow optimal designs to be found – the null-optimal and CRD-optimal designs have Δ_f and Δ_e designs with opposite signs. Interestingly,

Table 6.6 Comparison of run-time and expected sample size at $\delta = \delta_0$ of designs found from simulated annealing (SA) and OptGS

	$\mathbb{E}(N\delta_0)$			Time taken	
J	Average from 10 SA runs	Minimum from 10 SA runs	OptGS	Average SA run (s)	OptGS (s)
2	108.2	107.9	107.5	18.2	0.27
3	95.2	94.8	94.8	193.5	9.91
4	89.9	89.6	89.0	373.7	13.7
5	85.6	85.7	85.8	573.6	25.5

Table 6.7 Optimal design parameters $(\Delta_f, \Delta_e, C_f, C_e)$ for various optimality criteria and number of stages. The rows labelled $(1, 1, 1, 1)$ correspond to the $(1, 1, 1, 1)$-admissible design. Note that the expected and maximum sample sizes shown are for both treatment arms

Design	J	Δ_f	Δ_e	C_f	C_e	$\mathbb{E}(2N\delta_0)$	$\mathbb{E}(2N\delta_1)$	$\max \mathbb{E}(2N)$	$\max(2N)$
Null-optimal	2	0.45	−0.34	1.50	1.57	215.0	285.1	293.4	340
	3	0.52	−0.55	1.66	1.52	189.6	276.0	283.2	366
	4	0.52	−0.41	1.74	1.53	178.0	261.4	272.1	384
	5	0.53	−0.37	1.81	1.52	171.6	256.2	267.8	400
CRD-optimal	2	−0.18	0.46	1.25	1.84	241.0	234.6	276.9	344
	3	−0.15	0.48	1.26	1.96	231.1	214.8	265.6	372
	4	−0.13	0.49	1.27	2.03	222.7	205.5	259.0	392
	5	−0.01	0.48	1.31	2.06	207.3	200.0	250.6	410
δ-minimax	2	0.30	0.33	1.40	1.74	221.4	238.7	266.5	356
	3	0.33	0.33	1.48	1.79	196.8	219.4	251.9	384
	4	0.32	0.32	1.51	1.82	185.8	210.0	244.2	400
	5	0.32	0.32	1.53	1.84	179.7	204.2	239.4	410
$(1, 1, 1, 1)$	2	−0.01	0.08	1.32	1.68	226.3	245.7	272.9	324
	3	0.06	0.05	1.37	1.68	206.1	233.1	259.0	336
	4	0.12	0.12	1.41	1.71	194.3	220.2	248.8	352
	5	0.08	0.04	1.42	1.70	191.3	219.2	246.4	350

the δ-minimax and $(1, 1, 1, 1)$-admissible designs would be well approximated by the original one-parameter power-family, as Δ_f and Δ_e are very close in value.

6.7.3 Tutorial on Use of OptGS

OptGS contains a single function `optgs()`. The arguments taken by `optgs` are documented in the help file. The default arguments will produce a two-stage design with $\delta_0 = 0, \delta_1 = 1, \sigma = 3, \alpha = 0.05, 1 - \beta = 0.9$, and $(\omega_1, \omega_2, \omega_3, \omega_4) = (0.95, 0, 0, 0.05)$. The entries of ω imply that the design of interest is the admissible design that puts 0.95 weight on the expected sample size at the δ_0, and 0.05 weight on the maximum sample size. The output is as follows:

```
> optgs()

Groupsize:  84
Futility boundaries  0.5781 1.5776
Efficacy boundaries  2.9559 1.5776
ESS at null:     107.522
ESS at CRD:      145.325
Maximum ESS:     148.302
Max sample-size: 168
```

The output shows the required group-size (i.e. patients to be recruited per arm per stage); the futility and efficacy boundaries; and the operating characteristics of the design. Note that the expected sample sizes and maximum sample size are per arm. If the user wanted a design with three stages, then they could change the `J` argument:

```
> optgs(J=3)

Groupsize:  60
Futility boundaries  0.1388 0.9458 1.5551
Efficacy boundaries  3.9195 2.1874 1.5551
ESS at null:     94.935
ESS at CRD:      132.496
Maximum ESS:     137.018
Max sample-size: 180
```

Note that the futility and efficacy boundaries now have three entries. The expected sample sizes have all fallen, and the maximum sample size has risen, as one would expect. The above designs put weight on the expected sample size at the null, so will tend to have high expected sample sizes at the CRD, and also high maximum sample sizes. If the user wanted to put some of the weight on the expected sample size at the CRD, they could change the `weights` argument as follows:

```
> optgs(J=3,weights=c(0.5,0.45,0,0.05))

Groupsize:  62
Futility boundaries  -0.0062 1.0382 1.77
Efficacy boundaries  2.2247 1.9258 1.77
ESS at null:     98.945
ESS at CRD:      110.062
Maximum ESS:     126.107
Max sample-size: 186
```

Note that the resulting design has a somewhat higher expected sample size at the null, but considerably reduced expected sample size at the CRD (and also a reduced maximum expected sample size and an increased maximum sample size despite the respective weights not having changed).

As discussed in Sect. 6.6, in practice the assumption of known variance is not reasonable. OptGS uses the quantile-substitution method to convert the known-variance stopping boundaries to unknown-variance stopping boundaries. Setting the `sd.known` 0 argument to F will return unknown-variance stopping boundaries:

```
> optgs(J=3,weights=c(0.5,0.45,0,0.05),sd.known=F)

Groupsize:  62
Futility boundaries  -0.0062 1.0404 1.7749
Efficacy boundaries  2.2522 1.9351 1.7749
```

```
ESS at null:      98.945
ESS at CRD:       110.062
Maximum ESS:      126.107
Max sample-size: 186
```

Notice that in this case the stopping boundaries do not differ considerably to previously. This is because the group-size is fairly large. If the group-size was smaller, there would be a more noticeable difference between the two.

6.8 Multi-arm Multi-stage Clinical Trials

In this section, we briefly discuss recent work that extends group-sequential design methodology to allow testing of multiple experimental treatments against a control treatment. If more than one experimental treatment is available for testing, then testing all within a multi-arm trial is more efficient than separate randomised trials of each. That is because only one control group is needed instead of one control group per treatment. Applying group-sequential methodology to a multi-arm trial gives a multi-arm multi-stage (MAMS) clinical trial. At each interim analysis, treatments may be dropped for futility, or the whole trial may be stopped if an effective treatment is found.

6.8.1 Notation

Consider a MAMS trial with J stages and K experimental treatments and one control treatment. At each stage n patients are allocated to each remaining treatment. The treatment response of patient i on treatment k ($k = 0$ represents the control group), X_{ik}, is assumed to be distributed as $N(\mu_k, \sigma_k^2)$. The parameters of interest are $(\delta^{(1)}, \ldots, \delta^{(K)})$, where $\delta^{(k)} = \mu_k - \mu_0$. There are K null hypotheses being tested in the trial; the kth is $H_0^{(k)} : \delta^{(k)} \leq 0$.

At a given interim analysis j, the z-statistic for testing $H_0^{(k)}$, $Z_j^{(k)}$, is calculated:

$$Z_j^{(k)} = \sqrt{\frac{jn}{\sigma_k^2 + \sigma_0^2}} \frac{\sum_{i=1}^{jn} X_{ik} - \sum_{i=1}^{jn} X_{i0}}{jn} . \tag{6.8}$$

If $Z_j^{(k)} \leq f_j$, arm k is dropped for futility. If $Z_j^{(k)} > e_j$, then the trial stops for efficacy, and $H_0^{(k)}$ is rejected.

6.8.2 Designing a MAMS Trial

As in the group-sequential case, designing a MAMS trial involves choosing the group-size, futility boundaries and efficacy boundaries so that the type-I error and power are as required. The type-I error is more complicated than previously as there are multiple hypotheses. Magirr et al. [16] explain that it is sufficient to consider the probability of rejecting any null hypothesis when $\delta^{(1)} = \delta^{(2)} = \ldots = \delta^{(K)} = 0$, because this strongly controls the family-wise error rate. In other words, the probability of rejecting any true null hypothesis is maximised when $\delta^{(1)} = \delta^{(2)} = \ldots = \delta^{(K)} = 0$. The authors derive an analytic formula for this probability.

The power is also more complicated. Magirr et al. recommend powering the trial at the least favourable configuration (LFC) of Dunnett [6]. This is the probability of rejecting $H_0^{(1)}$ when $\delta^{(1)} = \delta_1$ and $\delta^{(2)} = \delta^{(3)} = \ldots = \delta^{(K)} = \delta_0$. Here, δ_1 is the clinically relevant difference, and δ_0 is the threshold such that if $\delta^{(k)}$ is below δ_0, treatment k is considered uninteresting. A suitable value of δ_0 could be 0, with higher values requiring a larger sample size but making it more likely that the best treatment will be picked.

Magirr et al. show how to apply traditional stopping boundaries to MAMS trials, for example those of Pocock. However, the same ideas of optimal and admissible designs discussed previously can be applied. Wason and Jaki [29] discuss considerations for searching for optimal designs in the case of a MAMS trial.

6.8.3 Future Work for Design of MAMS

MAMS trials are a very broad class of designs, with the ones considered above being relatively straightforward. In practice, MAMS trials have been used when the endpoints considered differ at each interim analysis, such as in the MRC STAMPEDE trial [27]. The methodology for this is described in Royston et al. [23], and consists of powering each individual stage separately. Efficiency could be gained by considering the whole trial at once, as Magirr et al. do, but this becomes difficult when the endpoint differs at each stage. Currently this area is an important priority for research.

6.9 Discussion

There are strong ethical and efficiency arguments for the use of group-sequential designs in practice. They reduce the average number of patients used in a trial, and therefore allow more trials to be run using the same limited population of patients. Statistical research in group-sequential designs has been ongoing since the 1970s, and shows no sign of slowing down. Greater computational power

has allowed considerable progress in areas such as searching for optimal group-sequential designs and group-sequential multi-arm multi-stage trial designs. This chapter has provided a summary of some of the recent research on group-sequential designs.

We have just considered normally distributed endpoints with known variance. Although this may at first seem highly restrictive, in fact asymptotically normally distributed test statistics are used for binary and survival endpoints. Thus, with some modification, methods discussed in this chapter can be used for other types of endpoints. The known variance assumption can be overcome with methods discussed in Sect. 6.6.

In practice, analyses may not take place when the planned number of patients have been assessed. Some patients may have dropped out of the trial, or practical considerations may have determined that the interim analysis must be at a certain time. In a time-to-event trial, it is particularly hard to ensure the planned number of events have taken place. As long as the total number of analyses is not varied this does not cause a problem as the stopping boundaries can be modified. Jennison and Turnbull [13] describe a method to adapt stopping boundaries from the one-parameter power family to allow different numbers of patients at each analysis. Additionally, fixed stopping boundaries from an optimal or admissible group-sequential design can be interpolated using an error spending function, as described by Kittelson and Emerson [15]. Both of these approaches control the overall type-I error, but not necessarily the power.

Group-sequential designs are less useful when the endpoint takes a long time to observe, such as in a time-to-event trial. In this case, one cannot pause recruitment until a group of patients have had the effect of treatment fully observed. Although group-sequential designs will not be able to reduce the expected number of patients recruited, they can still be useful in order to determine if a trial should be stopped early. Hampson and Jennison [11] propose group-sequential methods for when treatment responses are delayed. A Bayesian approach could also be used to incorporate early information to improve decision making at interim analyses, as discussed in chapter 5 of Berry et al. [4].

References

1. Anderson, K.: *gsDesign: Group Sequential Design* (2012). URL http://CRAN.R-project.org/package=gsDesign. R package version 2.6-04
2. Armitage, P., McPherson, C.K., Rowe, B.C.: Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society A* **132**, 235–244 (1969)
3. Barber, S., Jennison, C.: Optimal asymmetric one-sided group sequential tests. *Biometrika* **89**, 49–60 (2002)
4. Berry, S.M., Carlin, B.P., Lee, J.J., Muller, P.: *Bayesian adaptive methods for clinical trials*. CRC Press (2010)
5. Cook, R.J., Farewell, V.T.: Guidelines for monitoring efficacy and toxicity responses in clinical trials. *Biometrics* **50**, 1146–1152 (1994)

6. Dunnett, C.W.: A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* **50**, 1096–1121 (1955)
7. Eales, J.D., Jennison, C.: An improved method for deriving optimal one-sided group sequential tests. *Biometrika* **79**, 13–24 (1992)
8. Eales, J.D., Jennison, C.: Optimal two-sided group sequential tests. *Sequential Analysis* **14**, 273–286 (1995)
9. Emerson, S.S., Flemming, T.R.: Symmetric group sequential designs. *Biometrics* **45**, 905–923 (1989)
10. Genz, A., Bretz, F.: Methods for the computation of multivariate t-probabilities. *Journal of Computational and Graphical Statistics* **11**, 950–971 (2002)
11. Hampson, L.V., Jennison, C.: Group sequential tests for delayed responses. *Journal of the Royal Statistical Society B* **75**, 1–37 (2013)
12. Jennison, C., Turnbull, B.W.: Exact calculations for sequential t, χ^2 and f tests. *Biometrika* **78**, 133–141 (1991)
13. Jennison, C., Turnbull, B.W.: *Group sequential methods with applications to clinical trials*. Chapman and Hall (2000)
14. Jung, S.H., Lee, T., Kim, K., George, S.L.: Admissible two-stage designs for phase II cancer clinical trials. *Statistics in Medicine* **23**, 561–569 (2004)
15. Kittelson, J.M., Emerson, S.: A unifying family of group sequential test designs. *Biometrics* **55**, 874–882 (1999)
16. Magirr, D., Jaki, T., Whitehead, J.: A generalized Dunnett test for multiarm-multistage clincial studies with treatment selection. *Accepted by Biometrika* **99**, 494–501 (2012)
17. Mander, A.P., Wason, J.M.S., Sweeting, M.J., Thompson, S.G.: Admissible two-stage designs for phase ii cancer clinical trials that incorporate the expected sample size under the alternative hypothesis. *Pharmaceutical Statistics (in press)* **11**, 91–96 (2012)
18. McPherson, K., Armitage, P.: Repeated significance tests on accumulating data when the null hypothesis is not true. *Journal of the Royal Statistical Society A* **134**, 15–25 (1971)
19. Nelder, J.A., Mead, R.: A simplex method for function minimization. *The Computer Journal* **7**, 308–313 (1965)
20. O'Brien, P.C., Flemming, T.R.: A multiple-testing procedure for clinical trials. *Biometrics* **35**, 549–556 (1979)
21. Pampallona, S., Tsiatis, A.A.: Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *J. Statist. Planning and Inference* **42**, 19–35 (1994)
22. Pocock, S.J.: Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191–199 (1977)
23. Royston, P., Parmar, M.K.B., Qian, W.: Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Statistics in Medicine* **22**, 2239–2256 (2003)
24. Shao, J., Feng, H.: Group sequential t-tests for clinical trials with small sample sizes across stages. *Contemporary Clinical Trials* **28**, 563–571 (2007)
25. Shuster, J.: Optimal two-stage designs for single-arm phase II cancer trials. *Journal of Biopharmaceutical Statistics* **22**, 39–51 (2002)
26. Simon, R.: Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* **10**, 1–10 (1989)
27. Sydes, M.R., Parmar, M.K.B., James, N., et al.: Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the MRC STAMPEDE trial. *Trials* **10** (2009)
28. Wason, J.M.S.: OptGS: An R package for finding near-optimal group-sequential designs. *Journal of Statistical Software* **Accepted** (2013)
29. Wason, J.M.S., Jaki, T.: Optimal design of multi-arm multi-stage trials. *Statistics in Medicine* **31**, 4269–4279 (2012)
30. Wason, J.M.S., Mander, A.P.: Minimising the maximum expected sample size in two-stage phase II clinical trials with continuous outcomes. *Journal of Biopharmaceutical Statistics, in press* **22**, 836–852 (2012)

31. Wason, J.M.S., Mander, A.P., Thompson, S.: Optimal multi-stage designs for randomised clinical trials with continuous outcomes. *Statistics in Medicine* **31**, 301–312 (2012)
32. Wassmer, G., Vandemeulebroecke, M.: A brief review on software developments for group sequential and adaptive designs. *Biometrical Journal* **48**, 732–737 (2006)
33. Whitehead, J., Stratton, I.: Group sequential clinical trials with triangular continuation regions. *Biometrics* **39**, 227–236 (1983)
34. Whitehead, J., Valdes-Marquez, E., Lissmats, A.: A simple two-stage design for quantitative responses with application to a study in diabetic neuropathic pain. *Pharmaceutical statistics* **8**, 125–135 (2009)