Sample Size Reassessment and Hypothesis Testing in Adaptive Survival Trials

Dominic Magirr[1,*], Thomas Jaki[2], Franz Koenig[1], Martin Posch[1]

1 Section of Medical Statistics, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria
2 Medical and Pharmaceutical Statistics Research Unit, Lancaster University, Lancaster, UK

*** d.magirr@gmail.com**

Abstract

Mid-study design modifications are becoming increasingly accepted in confirmatory clinical trials, so long as appropriate methods are applied such that error rates are controlled. It is therefore unfortunate that the important case of time-to-event endpoints is not easily handled by the standard theory. We analyze current methods that allow design modifications to be based on the full interim data, i.e., not only the observed event times but also secondary endpoint and safety data from patients who are yet to have an event. We show that the final test statistic may ignore a substantial subset of the observed event times. An alternative test incorporating all event times is found, where a conservative assumption must be made in order to guarantee type I error control. We examine the power of this approach using the example of a clinical trial comparing two cancer therapies.

1 Introduction

There are often strong ethical and economic arguments for conducting interim analyses [1] of an ongoing clinical trial and for making changes to the design if warranted by the accumulating data. One may decide, for example, to increase the sample size on the basis of promising interim results. Or perhaps one might wish to drop a treatment from a multi-arm study on the basis of unsatisfactory safety data. Owing to the complexity of clinical drug development, it is not always possible to anticipate the need for such modifications, and therefore not all contingencies can be dealt with in the statistical design.

Unforeseen interim modifications complicate the frequentist statistical analysis of the trial considerably. Over recent decades many authors have investigated so-called "adaptive designs" in an effort to maintain the concept of type I error control [2–6]. While the theory behind these methods is now well understood if responses are observed immediately, subtle problems arise when responses are delayed, e.g., in survival trials.

[7] proposed adaptive survival tests that are constructed using the independent increments property of logrank test statistics [8–10]. However, as pointed out by [11], these methods only work if interim decision making is based solely on the interim logrank test statistics and any secondary endpoint data from patients who have already had an event. In other words, investigators must remain blind to the data from patients

who are censored at the interim analysis. [12] argue that decisions regarding interim design modifications should be as substantiated as possible, and propose a test procedure that allows investigators to use the full interim data. This methodology, similar to that of [13], does not require any assumptions regarding the joint distribution of survival times and short-term secondary endpoints, as do, e.g., the methods proposed by [14], [15, 16] and [17].

In this article we analyze the proposals of [13] and [12] and show that they are both based on weighted inverse-normal test statistics [18], with the common disadvantage that the final test statistic may ignore a substantial subset of the observed survival times. This is a serious limitation, as disregarding part of the observed data is generally considered inappropriate even if statistical error probabilities are controlled – see, for example, the discussion on overrunning in group sequential trials [17]. We quantify the potential inflation of the type I error rate if all observed data were used in these approaches. By adjusting the critical boundaries for the least favourable scenario we derive an alternative testing procedure which allows both, sample size reassessment and the use of all observed data.

The article is organized as follows. In Section 2 we review standard adaptive design theory and the recent methods of [13] and [12], as well as calculating the maximum type I error rate if the ignored data is naively reincorporated into the test statistic. In addition we construct a full-data guaranteed level-α test. In Section 3 we illustrate the procedures in clinical trial example and discuss the efficiency of the considered testing procedures. We present our conclusions in Section 4.

2 Methods

2.1 Adaptive Designs

Comprehensive accounts of adaptive design methodology can be found in [6, 19]. For testing a null hypothesis, $H_0 : \theta = 0$, against the one-sided alternative, $H_a : \theta > 0$, the two-stage adaptive test statistic is of the form $f_1(p_1) + f_2(p_2)$, where p_1 is the p-value based on first-stage data, p_2 is the p-value based on second-stage data, and f_1 and f_2 are prespecified monotonically decreasing functions. Consider the simplest case that no early rejection of the null hypothesis is possible at the end of the first stage. We will restrict attention to the weighted inverse-normal test statistic [18],

$$Z = w_1\Phi^{-1}(1-p_1) + w_2\Phi^{-1}(1-p_2), \tag{1}$$

where Φ denotes the standard normal distribution function and w_1 and w_2 are prespecified weights such that $w_1^2 + w_2^2 = 1$. If $Z > \Phi^{-1}(1-\alpha)$, then H_0 may be rejected at level α. The assumptions required to make this a valid level-α test are as follows [20].

Assumption 1

Let X_1^{int} denote the data available at the interim analysis, where $X_1^{\text{int}} \in \mathbb{R}^n$ with distribution function $G(x_1^{\text{int}}; \theta)$. In general, X_1^{int} will contain information not only concerning the primary endpoint, but also measurements on secondary endpoints and safety data. It is assumed that the first-stage p-value function $p_1 : \mathbb{R}^n \to [0, 1]$ satisfies

$$\int_{\mathbb{R}^n} \mathbf{1}\{p_1(x_1^{\text{int}}) \le u\}\, \mathrm{d}G(x_1^{\text{int}}; 0) \le u \text{ for all } u \in [0, 1]\,.$$

Assumption 2

At the interim analysis, a second-stage design d is chosen. The second-stage design is allowed to depend on the unblinded first-stage data without prespecifying an adaptation rule. Denote the second-stage data by Y, where $Y \in \mathbb{R}^m$. It is assumed that the

distribution function of Y, denoted by $F_{\delta,x_1^{\text{int}}}(y,\theta)$, is known for all possible second stage designs, δ, and all first-stage outcomes, x_1^{int}.

Assumption 3

The second-stage p-value function $p_2 : \mathbb{R}^m \to [0,1]$ satisfies $\int_{\mathbb{R}^m} \mathbf{1}\{p_2(y) \leq u\}\, \mathrm{d}F_{\delta,x_1^{\text{int}}}(y;0) \leq u$ for all $u \in [0,1]$.

Immediate responses The aforementioned assumptions are easy to justify when primary endpoint responses are observed more-or-less immediately. In this case X_1^{int} contains the responses of all patients recruited prior to the interim analysis. A second-stage design δ can subsequently be chosen with the responses from a new cohort of patients contributing to Y.

Delayed responses and the independent increments assumption An interim analysis may take place whilst some patients have entered the study but have yet to provide a data point on the primary outcome measure. Most approaches to this problem [7,8,10] attempt to take advantage of the well known independent increments structure of score statistics in group sequential designs [21]. As pictured in Figure 1, X_1^{int} will generally include responses on short-term secondary endpoints and safety data from patients who are yet to provide a primary outcome measure, while Y consists of some delayed responses from patients recruited prior to the interim analysis, mixed together with responses from a new cohort of patients.

Figure 1. Schematic of a two-stage adaptive trial design with delayed response using the independent increments assumption.

Let $S(X_1^{\text{int}})$ and $\mathcal{I}(X_1^{\text{int}})$ denote the score statistic and Fisher's information for θ, calculated from primary endpoint responses in X_1^{int}. Assuming suitable regularity conditions, the asymptotic null distribution of $S(X_1^{\text{int}})$ is Gaussian with mean zero and variance $\mathcal{I}(X_1^{\text{int}})$ [22]. The independent increments assumption is that for all first-stage outcomes x_1^{int} and second-stage designs δ, the null distribution of Y is such that

$$S(x_1^{\text{int}}, Y) - S(x_1^{\text{int}}) \sim \mathcal{N}\left\{0, \mathcal{I}(x_1^{\text{int}}, Y) - \mathcal{I}(x_1^{\text{int}})\right\}, \tag{2}$$

at least approximately, where $S(X_1^{\text{int}}, Y)$ and $\mathcal{I}(X_1^{\text{int}}, Y)$ denote the score statistic and Fisher's information for θ, calculated from primary endpoint responses in (X_1^{int}, Y).

Unfortunately, (2) is seldom realistic in an adaptive setting. [11] show that if the adaptive strategy at the interim analysis is dependent on short-term outcomes in X_1^{int} that are correlated with primary endpoint outcomes in Y, i.e., from the same patient, then a naive appeal to the independent increments assumption can lead to very large type I error inflation.

Delayed responses with patient-wise separation An alternative approach, which we call "patient-wise separation", redefines the first-stage p-value, $p_1 : \mathbb{R}^p \to [0,1]$, to be a function of X_1, where X_1 denotes all the data from patients recruited prior to the interim analysis at calendar time T^{int}, followed-up until a pre-fixed maximum calendar time $T^{\max}$. In this case p_1 may not be observable at the time the second-stage design δ is chosen. This is not a problem, as long as no early rejection at the end of the first stage is foreseen. Any interim decisions, such as increasing the sample size, do not require knowledge of p_1. It is assumed that Y consists of responses from a new cohort of patients, such that x_1^{int} could be formally replaced with x_1 in the aforementioned adaptive design assumptions. We call this patient-wise separation because data from the same patient cannot contribute to both p_1 and p_2.

[23] and [24] apply this approach when a patient's primary outcome can be measured after a fixed period of follow-up, e.g., 4 months. However, one must take additional care with a time-to-event endpoint, as one is typically not prepared to wait for all first-stage patients – those patients recruited prior to T^{int} – to have an event. Rather, p_1 is defined as the p-value from a statistical test applied to the data from first-stage patients followed up until time T^{end}, for some $T^{\text{end}} \leq T^{\max}$. In this case it is vital that T^{end} be fixed at the start of the trial, either explicitly or implicitly [12, 13]. Otherwise, if T^{end} were to depend on the adaptive strategy at the interim analysis, this would impact the distribution of p_1 and could lead to type I error inflation.

The situation is represented pictorially in Figure 2. An unfortunate consequence of pre-fixing T^{end} is that this will not, in all likelihood, correspond to the end of follow-up for second-stage patients. All events from first-stage patients that occur after T^{end} make no contribution to the statistic (1).

Figure 2. Schematic of a two-stage adaptive trial design with delayed response using patient-wise separation.

2.2 Adaptive Survival Studies

Consider a randomized clinical trial comparing survival times on an experimental treatment, E, with those on a control treatment, C. For simplicity, we will focus on the logrank statistic for testing the null hypothesis $H_0 : \theta = 0$ against the one-sided alternative $H_a : \theta > 0$, where θ is the log hazard ratio, assuming proportional hazards. Similar arguments could be applied to the Cox model. Let $D_1(t)$ and $S_1(t)$ denote the number of uncensored events and the usual logrank score statistic, respectively, based on the data from first-stage patients – those patients recruited prior to the interim analysis – followed up until calendar time t, $t \in [0, T^{\max}]$. Under the null hypothesis, assuming equal allocation and a large number of events, the variance of $S_1(t)$ is approximately equal to $D_1(t)/4$ [25]. The first-stage p-value must be calculated at a prefixed time point T^{end}:

$$p_1 = 1 - \Phi\left[S_1(T^{\text{end}}) / \left\{D_1(T^{\text{end}})/4\right\}^{1/2}\right]. \tag{3}$$

The number of events can be prefixed at d_1, say, with T^{end} chosen implicitly

$$T^{\text{end}} := \min\left\{t : D_1(t) = d_1\right\}. \tag{4}$$

Jenkins et al., method [13] describe a "patient-wise separation" adaptive survival trial, with test statistic (1), first-stage p-value (3) and T^{end} defined as in (4). While their focus is on subgroup selection, we will appropriate their method for the simpler situation of a single comparison, where at the interim analysis one has the possibility to adapt the pre-planned number of events from second-stage patients – i.e., those patients recruited post T^{int}. The weights in (1) are pre-fixed in proportion to the pre-planned number of events to be contributed from each stage, i.e., $w_1^2 = d_1/(d_1 + d_2)$, where $d_1 + d_2$ is the total originally required number of events. The second-stage p-value corresponds to a logrank test based on second-stage patients, i.e.,

$$p_2 = 1 - \Phi\left[S_2(T_2^*) / \left\{D_2(T_2^*)/4\right\}^{1/2}\right],$$

where $T_2^* := \min\left\{t : D_2(t) = d_2^*\right\}$ with $S_2(t)$ and $D_2(t)$ defined analogously to $S_1(t)$ and $D_1(t)$, and where d_2^* is specified at the interim analysis.

Irle and Schäfer method Instead of explicitly combining stage-wise p-values, [12] employ the closely related "conditional error" approach [3, 4, 26].

They begin by prespecifying a level-α test with decision function, φ, taking values in $\{0, 1\}$ corresponding to non-rejection and rejection of H_0, respectively. For a survival trial, this entails specifying the sample size, duration of follow-up, test statistic, recruitment rate, etc. Then, at some not necessarily prespecified timepoint, T^{int}, an interim analysis is performed. The timing of the interim analysis induces a partition of the trial data, (X_1, X_2), where X_1 and X_2 denote the data from patients recruited prior- T^{int} and post- T^{int}, respectively, followed-up until time $T^{\max}$. For a standard log-rank test, the decision function is

$$\varphi(X_1, X_2) = \mathbf{1}\left[S(T^{\text{end}})/\left\{D(T^{\text{end}})/4\right\}^{1/2} > \Phi^{-1}(1-\alpha)\right], \tag{5}$$

where $D(T^{\text{end}})$ and $S(T^{\text{end}})$ denote the number of uncensored events and the usual logrank score statistic, respectively, based on data from all patients followed-up until time $T^{\text{end}} := \min\{t : D(t) = d\}$ for some prespecified number of events d.

At the interim analysis, the general idea is to use the unblinded first-stage data x_1^{int} to define a second-stage design, δ, without the need for a prespecified adaptation strategy. Again, the definition of δ includes factors such as sample size, follow-up period, recruitment rate, etc., in addition to a second-stage decision function $\psi : \mathbb{R}^m \to \{0, 1\}$ based on second-stage data $Y \in \mathbb{R}^m$. Ideally, one would like to choose ψ such that $E_{H_0}(\psi \mid X_1^{\text{int}} = x_1^{\text{int}}) = E_{H_0}(\varphi \mid X_1^{\text{int}} = x_1^{\text{int}})$, as this would ensure that

$$E_{H_0}(\psi) = E_{H_0}\left\{E_{H_0}\left(\psi \mid X_1^{\text{int}}\right)\right\} = E_{H_0}\left\{E_{H_0}\left(\varphi \mid X_1^{\text{int}}\right)\right\} = E_{H_0}(\varphi) = \alpha, \tag{6}$$

i.e., the overall procedure controls the type I error rate at level α. Unfortunately, this approach is not directly applicable in a survival trial where X_1^{int} contains short-term data from first-stage patients surviving beyond T^{int}. This is because it is impossible to calculate $E_{H_0}(\varphi \mid X_1^{\text{int}} = x_1^{\text{int}})$ and $E_{H_0}(\psi \mid X_1^{\text{int}} = x_1^{\text{int}})$, owing to the unknown joint distribution of survival times and the secondary endpoints already observed at the interim analysis. One may, however, condition on X_1 rather than on X_1^{int} and choose ψ such that $E_{H_0}(\psi \mid X_1 = x_1) = E_{H_0}(\varphi \mid X_1 = x_1)$, thus ensuring type I error control following the same argument as (6). For example, it is possible to extend patient follow-up and use the second-stage decision function

$$\psi(X_2) = \mathbf{1}\left[S(T^*)/\left\{D(T^*)/4\right\}^{1/2} \geq b^*\right], \tag{7}$$

where $T^* := \min\{t : D(t) = d^*\}$, $d^* \geq d$ is chosen at the interim analysis, and b^* is a cutoff value that must be determined. [12] show that, asymptotically,

$$E_{H_0}\{\varphi \mid X_1 = x_1\} = E_{H_0}\left\{\varphi \mid S_1(T^{\text{end}}) = s_1\right\}$$

and

$$E_{H_0}\{\psi \mid X_1 = x_1\} = E_{H_0}\{\psi \mid S_1(T^*) = s_1^*\}.$$

In each case, calculation of the right-hand-side is facilitated by the asymptotic result that, assuming equal allocation under the null hypothesis, for $t \in [0, T^{\max}]$,

$$\begin{pmatrix} S_1(t) \\ S(t) - S_1(t) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} D_1(t)/4 & 0 \\ 0 & \{D(t) - D_1(t)\}/4 \end{pmatrix}\right). \tag{8}$$

One remaining subtlety is that $E_{H_0}\{\psi \mid S_1(T^*) = s_1^*\}$ can only calculated at calendar time T^*, where $T^* > T^{\text{int}}$. Determination of b^* must therefore be postponed until this later time.

Using result (8), it is straightforward to show that $\psi = 1$ if and only if $Z > \Phi^{-1}(1-\alpha)$, where Z is defined as in (1) with p_1 defined as in (3), the second-stage p-value function defined as

$$p_2(Y) = 1 - \Phi\left(\left\{S(T^*) - S_1(T^*)\right\} / \left[\left\{D(T^*) - D_1(T^*)\right\}/4\right]^{1/2}\right), \tag{9}$$

and the specific choice of weighting $w_1^2 = D_1(T^{\text{end}})/D(T^{\text{end}})$. Full details are provided in the 4.

Remark 1. The Irle and Schäfer method uses the same test statistic as the Jenkins et al. method, with a clever way of implicitly defining the weights and the end of first-stage follow-up, T^{end}. It has two potential advantages. Firstly, the timing of the interim analysis need not be prespecified – in theory, one is permitted to monitor the accumulating data and at any moment decide that design changes are necessary. Secondly, if no changes to the design are necessary, i.e., the trial completes as planned at calendar time T^{end}, then the original test (5) is performed. In this special case, no events are ignored in the final test statistic.

Remark 2. From first glance at (7), it may appear that the events from first-stage patients, occurring after T^{end}, always make a contribution to the final test statistic. However, this data is still effectively ignored. We have shown in the online supplement that the procedure is equivalent to a p-value combination approach where p_1 depends only on data available at time T^{end}. In addition, the distribution of p_2 is asymptotically independent of the data from first-stage patients: note that $S(T^*) - S_1(T^*)$ and $S_2(T^*)$ are asymptotically equivalent [12]. The procedure therefore fits our description of a "patient-wise separation" design, and the picture is the same as in Figure 2. The first-stage patients have in effect been censored at T^{end}, despite having been followed-up for longer. This fact has important implications for the choice of d^*. If one chooses d^* based on conditional power arguments, one should be aware that the effective sample size has not increased by $d^* - d$. Rather, it has increased by $d^* - d - \left\{D_1(T^*) - D_1(T^{\text{end}})\right\}$, which could be very much smaller.

Remark 3. A potential disadvantage of the Irle and Schäfer method compared to the Jenkins et al. method is that one is not permitted to adapt any aspect of the recruitment process prior to time T^{end}. Contrary to what is claimed in [12], it is not valid to extend the recruitment period (or speed up recruitment as in the example they give) to reach an increased number of events d^* within the originally planned trial duration. This is because T^{end} is defined implicitly as $T^{\text{end}} := \min\{t : D(t) = d\}$ under the assumptions of the original design. Therefore T^{end} is unobservable if the recruitment process is changed in response to the interim data. [27] discuss this issue further.

2.3 Hypothesis tests based on all available follow-up data

Suppose that the trial continues until calendar time $T^* \in (T^{\text{end}}, T^{\max})$. Data from first-stage patients – those patients recruited prior to T^{int} – accumulating between times T^{end} and T^* should be ignored. In this section we will investigate what happens, in a worst case scenario, if this illegitimate data is naively incorporated into the adaptive test statistic (1). Specifically, we find the maximum type I error associated with the test statistic

$$Z^* = w_1 S_1(T^*) / \left\{D_1(T^*)/4\right\}^{1/2} + w_2 \Phi^{-1}(1-p_2). \tag{10}$$

Since T^* depends on the interim data in a complicated way, the null distribution of (10) is unknown. One can, however, consider properties of the stochastic process

$$Z(t) = w_1 S_1(t) / \left\{D_1(t)/4\right\}^{1/2} + w_2 \Phi^{-1}(1-p_2), \qquad t \in \left[T^{\text{end}}, T^{\max}\right].$$

In other words, we consider continuous monitoring of the logrank statistic based on first-stage patient data. The worst-case scenario assumption is that the responses on short-term secondary endpoints, available at the interim analysis, can be used to predict the exact calendar time the process $Z(t)$ reaches its maximum. In this case, one could attempt to engineer the second stage design such that T^* coincides with this timepoint, and the worst-case type I error rate is therefore

$$P_{H_0}\left\{\max_{T^{\text{end}}\leq t\leq T^{\max}} Z(t) > \Phi^{-1}(1-\alpha)\right\}. \tag{11}$$

Although the worst-case scenario assumption is clearly unrealistic, (11) serves as an upper bound on the type I error rate. It can be found approximately via standard Brownian motion results. Let $u := D_1(t)/D_1(T^{\max})$ denote the information time at calendar time t, and let $S_1(u)$ denote the logrank score statistic based on first-stage patients, followed-up until information time u. It can be shown that $B(u) := S_1(u)/\left\{D_1(T^{\max})/4\right\}^{1/2}$ behaves asymptotically like a Brownian motion with drift $\xi := \theta\left\{D_1(T^{\max})/4\right\}^{1/2}$ [28]. We wish to calculate

$$\begin{aligned} &P_{\theta=0}\left\{\max_{T^{\text{end}}\leq t\leq T^{\max}} Z(t) > \Phi^{-1}(1-\alpha)\right\} \\ =&\int_0^1 P_{\theta=0}\left[\max_{u_1\leq u\leq 1} B(u) > u^{1/2}w_1^{-1}\left\{\Phi^{-1}(1-\alpha) - w_2\Phi^{-1}(1-p_2)\right\}\right] \mathrm{d}p_2, \end{aligned} \tag{12}$$

where $u_1 = D_1(T^{\text{end}})/D_1(T^{\max})$. While the integrand on the right-hand-side is difficult to evaluate exactly, it can be found to any required degree of accuracy by replacing the square root stopping boundary with a piecewise linear boundary [29].

The two parameters that govern the size of (11) are w_1 and u_1. Larger values of w_1 reflect an increased weighting of the first-stage data, which increases the potential inflation. In addition, a low value for u_1 increases the window of opportunity for stopping on a random high. Table 1 shows that for a nominal $\alpha = 0.025$ level test, the worst-case type I error can be up to 15% when $u_1 = 0.1$ and $w_1 = 0.9$. As $u_1 \to 0$ the worst-case type I error rate tends to 1 for any value of $w_1 > 0$ [30].

Table 1. Worst case type I error for various choices of weights and information fractions. Nominal level $\alpha = 0.025$ one-sided.

	u_1								
w_1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.052	0.047	0.044	0.041	0.039	0.037	0.035	0.033	0.030
0.2	0.067	0.059	0.054	0.050	0.046	0.043	0.039	0.036	0.032
0.3	0.081	0.070	0.062	0.057	0.052	0.047	0.043	0.039	0.034
0.4	0.094	0.080	0.071	0.063	0.057	0.052	0.046	0.041	0.036
0.5	0.106	0.089	0.078	0.069	0.062	0.056	0.050	0.044	0.037
0.6	0.119	0.098	0.085	0.075	0.067	0.059	0.053	0.046	0.038
0.7	0.131	0.107	0.092	0.081	0.072	0.063	0.055	0.048	0.040
0.8	0.143	0.116	0.100	0.087	0.076	0.067	0.058	0.050	0.041
0.9	0.155	0.125	0.106	0.092	0.081	0.070	0.061	0.052	0.042

A full-data guaranteed level-α test If one is unprepared to give up the guarantee of type I error control, an alternative test can be found by increasing the cut-off value for Z^* from $\Phi^{-1}(1-\alpha)$ to k^* such that

$$\int_0^1 P_{\theta=0}\left[\max_{u=u_1}^{1} B(u) > u^{1/2}w_1^{-1}\left\{k^* - w_2\Phi^{-1}(1-p_2)\right\}\right] \mathrm{d}p_2 = \alpha.$$

3 Results

3.1 Clinical trial example

The upper bound on the type I error rate varies substantially across w_1 and u_1. To give an indication of what can be expected in practice, consider a simplified version of the trial described in [12]. A randomized trial is set up to compare chemotherapy (C) with a combination of radiotherapy and chemotherapy (E). The anticipated median survival time on C is 14 months. If E were to increase the median survival time to 20 months then this would be considered a clinically relevant improvement. Assuming exponential survival times, this gives anticipated hazard rates $\lambda_C = 0.050$ and $\lambda_E = 0.035$, and a target log hazard ratio of $\theta_R = -\log(\lambda_E/\lambda_C) \approx 0.36$. If the error rates for testing $H_0 : \theta = 0$ against $H_a : \theta = \theta_R$ are $\alpha = 0.025$ (one-sided) and $\beta = 0.2$, the required number of deaths (assuming equal allocation) is

$$d = 4\left[\left\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\right\}/\theta_R\right]^2 \approx 248.$$

The relationship between the required number of events and the sample size depends on the recruitment pattern, and we will consider two scenarios. In our "slow recruitment" scenario, patients are recruited uniformly at a rate of 8 per month for a maximum of 60 months with an interim analysis performed at 23 months. In our"fast recruitment" scenario, patients are recruited uniformly at a rate of 50 per month for a maximum of 18 months with an interim analysis after 8 months. In both cases, the only adaptation we allow at the interim analysis is to increase the number of events. Recruitment must continue as planned but the follow-up period may be extended. The maximum duration of the trial is restricted to 100 months in the first case and 30 months in the second case.

Figure 3 shows the expected number of events as a function of time for both scenarios assuming exponentially distributed survival times with hazards equal to the planned values.

Figure 3. Expected total number of events as a function of time based on exponential survival with hazard rates $\lambda_C = 0.05$ and $\lambda_E = 0.035$. Slow recruitment: 8 patients per month for a maximum of 60 months. Fast recruitment: 50 patients per month for a maximum of 18 months. Vertical lines are at T^{int}, T^{end} and $T^{\max}$.

The maximum type I error inflation, determined via w_1 and u_1, will depend on the observed number of events from first- and second-stage patients at calendar times T^{int} and T^{end}. However, the expected pattern of events in Figure 3 provide some indication. In the slow recruitment scenario, we expect to recruit 179 patients by the time of the interim analysis. We also expect 149 of the first 248 events to come from patients recruited prior to the interim analysis. These numbers would give $w_1 = (149/248)^{1/2}$, $u_1 = 149/179$ and, according to equation (12), $\max \alpha = 0.044$. For the fast recruitment scenarios the respective quantities are $w_1 = (169/248)^{1/2}$, $u_1 = 169/264$ and $\max \alpha = 0.060$.

On the efficiency of the full-data level-α test Consider the full-data guaranteed level-α test defined above. Recall that this test has the advantage of allowing interim decision making to be based on all available data whilst using a final test statistic that takes account of all observed event times. Unfortunately, this advantage is likely to be outweighed by the loss in power resulting from the increased cut-off value, as can be seen in Figure 4. The difference between the noncentrality parameters of $Z(T^*)$ and $Z(T^{\text{end}})$ is plotted against the time extension $T^* - T^{\text{end}}$ for various choices of θ. In the slow recruitment scenario the increase in the noncentrality parameter is outweighed by

the increase in the cut-off value, even when the log-hazard ratio is as large as was expected in the planning phase. In the fast recruitment setting, it is possible for the increase in the noncentrality parameter to exceed the increase in the cut-off value when the trial is extended substantially. However, the trial would typically only need to be increased substantially if the true effect size were lower than planned. And in this case ($\theta \leq 0.66\theta_R$) one can see that the increased cut-off value still dominates.

Figure 4. Difference between the noncentrality parameters of the adaptive test statistics $Z(T^*)$ and $Z(T^{\text{end}})$ as a function of the time extension $T^* - T^{\text{end}} \in \left[0, T^{\max} - T^{\text{end}}\right]$. Horizontal lines are drawn at $k^* - \Phi^{-1}(0.975)$, where k^* denotes the cut-off value of the full-data guaranteed level-α test, and Φ denotes the standard normal distribution function.

4 Discussion

Unblinded sample-size recalculation has been criticized for its lack of efficiency relative to classical group sequential designs [32, 33]. If the recalculation is made on the basis of an early estimate of treatment effect, the final sample size is likely to have high variability [34], and, in addition, the test decision is based on a non-sufficient statistic. [35] show how, for a given re-estimation rule, a classical group sequential design can be found with an essentially identical power function but lower expected sample size.

In response to these arguments [36] emphasize that "the real benefit of the adaptive approach arises through the ability to invest sample resources into the trial in stages". An efficient group sequential trial, on the other hand, requires a large up-front sample size commitment and aggressive early stopping boundaries. From the point of view of the trial sponsor, the added flexibility may in some circumstances outweigh the loss of efficiency.

In this paper we have shown that when the primary endpoint is time-to-event, a fully unblinded sample-size recalculation – i.e., a decision based on all available efficacy and safety data – has additional drawbacks not considered in the aforementioned literature. Recently proposed methods [12, 13] share the common disadvantage that some patients' event times are ignored in the final test statistic. This is usually deemed unacceptable by regulators. Furthermore, it is the long-term data of patients recruited prior to the interim analysis that is ignored, such that more emphasis is put on early events in the final decision making. This neglect becomes more serious, therefore, if the hazard rates differ substantially only at large survival times. Note, however, that a standard logrank test would already be inefficient in this scenario [37].

The relative benefit of the Irle and Schäfer method [12], in comparison with that of Jenkins et al. [13], is that the timing of the interim analysis need not be pre-specified and, in addition, the method is efficient if no design changes are necessary. On the other hand, the Irle and Schäfer method has the serious practical flaw that it is not permissible to change any aspect of the recruitment process in response to the interim data.

Confirmatory clinical trials with time-to-event endpoints appear to be one of the most important fields of application of adaptive methods [38]. It is therefore especially important that investigators considering an unblinded sample size re-estimation in this context are aware of the additional issues involved. We have shown that all considered procedures will require giving up an important statistical property – a situation summarized succinctly in Table 2.

Table 2. Trade-off involved in choosing between methods when extending the follow-up period of a survival trial. Methods considered: (A), data is combined assuming independent stage-wise increments; (B), patient-wise separation with pre-fixed end of first-stage follow-up; (C), naive patient-wise separation without pre-fixed end of first-stage follow-up; and (D), patient-wise separation using the full-data guaranteed level-α test.

	Strict type I error control	All data available for interim decisions	All events included in test statistic	Relative power
(A) Ind. Increments	✓	×	✓	✓
(B) $Z(T^{\text{end}}) > z_{1-\alpha}$	✓	✓	×	✓
(C) $Z(T^*) > z_{1-\alpha}$	×	✓	✓	✓
(D) $Z(T^*) > k^*$	✓	✓	✓	×

The relevance of these issues is highlighted by the recently published VALOR trial in acute myeloid leukaemia [39]. Treatment effect estimates from phase II data suggested that 375 events might be sufficient to confirm efficacy. However, there is always uncertainty surrounding such an estimate. A smaller effect size - corresponding to upwards of 500 required events - would still be clinically meaningful, but funding such a trial was beyond the resources of the study sponsor. The solution was to initiate the trial with the smaller sample size but plan an interim analysis, whereby promising results would trigger additional investment. In this case, the interim decision rules were pre-specified and, upon observing a promising hazard ratio after 173 events, the total required number of events was increased to 562. The final analysis was based on a weighted combination of log-rank statistics, corresponding to method (A) in Table 2. It is important to emphasize that the validity of this approach relies on the second-stage sample size being a function of the interim hazard ratio. Had other information – e.g., disease progressions – played a part in the interim decision making, then the type I error rate could have been compromised as described in this paper.

While statistical theory can be developed to control the type I error rate given certain model assumptions, there is always the potential for "operational bias" to enter an adaptive trial. FDA draft guidance [40] emphasizes the need to shield investigators as much as possible from knowledge of the adaptive changes. The very knowledge that sample size has been increased – implying a "promising" interim effect estimate – could lead to changes of behavior in terms of treating, managing, and evaluating study participants. As a minimum, the European Medicines Agency requires that the primary analysis "be stratified according to whether patients were randomized before or after the protocol amendment" [41]. Aside from the regulatory importance, it is also in the sponsor's interest to minimize operational bias when trial outcomes will influence significant investment decisions [42]. For a further discussion on the regulatory and logistical challenges sponsors may face we refer to [6, 19].

We have focussed our attention on the type I error control and power of the various procedures. Estimation of the treatment effect size following an adaptive survival trial is also an important topic. Current available methods can be found in [8], [43] and [44].

Supporting Information

S1 Appendix

Connection between conditional error and combination test. The cut-off b^*

satisfies

$$\begin{aligned}
&E_{H_0}\left\{\varphi \mid S_1(T^{\text{end}}) = s_1\right\} \\
&= P_{H_0}\left\{S(T^*)/(d^*/4)^{1/2} \geq b^* \mid S_1(T^*) = s_1^*\right\} \\
&= P_{H_0}\left(\left\{S(T^*) - S_1(T^*)\right\}/\left[\left\{d^* - D_1(T^*)\right\}/4\right]^{1/2} \geq c^* \mid S_1(T^*) = s_1^*\right],
\end{aligned}$$

which implies that $c^* = \Phi^{-1}\left[1 - E_{H_0}\left\{\varphi \mid S_1(T) = s_1\right\}\right]$. Therefore,

$$\begin{aligned}
\psi = 1 &\Leftrightarrow S(T^*)/(d^*/4)^{1/2} \geq b^* \\
&\Leftrightarrow \left\{S(T^*) - S_1(T^*)\right\}/\left[\left\{d^* - D_1(T^*)\right\}/4\right]^{1/2} \geq c^* \\
&\Leftrightarrow \Phi^{-1}(1-p_2) \geq \Phi^{-1}\left[1 - E_{H_0}\left\{\varphi \mid S_1(T^{\text{end}}) = s_1\right\}\right] \\
&\Leftrightarrow p_2 \leq E_{H_0}\left\{\varphi \mid S_1(T^{\text{end}}) = s_1\right\}.
\end{aligned}$$

The conditional error probability, $E_{H_0}\left\{\varphi \mid S_1(T^{\text{end}}) = s_1\right\}$, can be found from the joint distribution of $S_1(T^{\text{end}})$ and $S(T^{\text{end}})$. Omitting the argument T^{end} from S_1, S, D_1 and D:

$$\begin{aligned}
E_{H_0}\left\{\varphi \mid S_1 = s_1\right\} &= P_{H_0}\left\{S/(D/4)^{1/2} > \Phi^{-1}(1-\alpha) \mid S_1 = s_1\right\} \\
&= P_{H_0}\left[2(S - S_1)/(D - D_1)^{1/2} > \Phi^{-1}(1-\alpha)\left\{D/(D - D_1)\right\}^{1/2}\right. \\
&\qquad \left. -2S_1/(D - D_1)^{1/2} \mid S_1 = s_1\right] \\
&= 1 - \Phi\left[\Phi^{-1}(1-\alpha)\left\{D/(D - D_1)\right\}^{1/2} - \Phi^{-1}(1-p_1)\left\{D_1/(D - D_1)\right\}^{1/2}\right]
\end{aligned}$$

and therefore $p_2 \leq E_{H_0}\left\{\varphi \mid S_1(T) = s_1\right\}$ if and only if

$$\left\{D_1(T^{\text{end}})/d\right\}^{1/2}\Phi^{-1}(1-p_1) + \left[\left\{d - D_1(T^{\text{end}})\right\}/d\right]^{1/2}\Phi^{-1}(1-p_2) \geq \Phi^{-1}(1-\alpha).$$

Acknowledgments

DM was funded by the Medical Research Council (MR/J004979/1) and the Austrian Science Fund (FWF P23167), TJ was supported by the National Institute for Health Research (NIHR-CDF-2010-03-32). FK was supported by European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 602552 (IDEAL). MP was supported by the EU FP7 HEALTH.2013.4.2-3 project Asterix (Grant Number 603160). The views expressed in this publication are those of the authors and should not be attributed to any of the funding institutions, or organisations with which the authors are affiliated

References

1. Jiang Z1, Wang L, Li C, Xia J, Jia H. A practical simulation method to calculate sample size of group sequential trials for time-to-event data under exponential and Weibull distribution. PLoS One. 2012;7(9):e44013.

2. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. Biometrics. 1994;50:1029–1041. Correction: *Biometrics* 1996; 52:380.

3. Proschan MA, Hunsberger SA. Designed Extension of Studies Based on Conditional Power. Biometrics. 1995;51:1315–1324. Available from: `http://www.jstor.org/stable/2533262`.

4. Müller HH, Schäfer H. Adaptive Group Sequential Designs for Clinical Trials: Combining the Advantages of Adaptive and of Classical Group Sequential Approaches. Biometrics. 2001;57:886–891.

5. Hommel G. Adaptive modifications of hypotheses after an interim analysis. Biometrical Journal. 2001;43(5):581–589.

6. Bauer P, Bretz F, Dragalin V, König F, Wassmer G. Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. Statistics in Medicine. 2015 (Early View). *DOI: 10.1002/sim.6472* .

7. Schäfer H, Müller HH. Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. Statistics in medicine. 2001;20(24):3741–3751.

8. Wassmer G. Planning and analyzing adaptive group sequential survival trials. Biometrical Journal. 2006;48(4):714–729.

9. Desseaux K, Porcher R. Flexible two-stage design with sample size reassessment for survival trials. Statistics in medicine. 2007;26(27):5002–5013.

10. Jahn-Eimermacher A, Ingel K. Adaptive trial design: A general methodology for censored time to event data. Contemporary clinical trials. 2009;30(2):171–177.

11. Bauer P, Posch M. Letter to the editor. Statistics in Medicine. 2004;23:1333–1334.

12. Irle S, Schäfer H. Interim design modifications in time-to-event studies. Journal of the American Statistical Association. 2012;107:341–348.

13. Jenkins M, Stone A, Jennison C. An Adaptive Seamless Phase II/III Design for Oncology Trials with Subpopulation Selection Using Correlated Survival Endpoints. Pharmaceutical Statistics. 2011;10:347–356.

14. Stallard N. A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. Statistics in medicine. 2010;29(9):959–971.

15. Friede T, Parsons N, Stallard N, Todd S, Valdes Marquez E, Chataway J, et al. Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: An application in multiple sclerosis. Statistics in medicine. 2011;30(13):1528–1540.

16. Friede T, Parsons N, Stallard N. A conditional error function approach for subgroup selection in adaptive clinical trials. Statistics in Medicine. 2012;31(30):4309–4320.

17. Hampson LV, Jennison C. Group sequential tests for delayed responses (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2013;75(1):3–54.

18. Lehmacher W, Wassmer G. Adaptive Sample Size Calculations in Group Sequential Trials. Biometrics. 1999;55(4):pp. 1286–1290. Available from: `http://www.jstor.org/stable/2533757`.

19. Bretz F, Koenig F, Brannath W, Glimm E, Posch M. Adaptive Designs for Confirmatory Clinical Trials. Statistics in Medicine. 2009;28:1181–1217.

20. Brannath W, Gutjahr G, Bauer P. Probabilistic Foundation of Confirmatory Adaptive Designs. Journal of the American Statistical Association. 2012;107:824–832.

21. Jennison C, Turnbull BW. Group Sequential Methods with Applications to Clinical Trials. Boca Raton, FL: Chapman and Hall; 2000.

22. Cox DR, Hinkley DV. Theoretical statistics. CRC Press; 1979.

23. Liu Q, Pledger GW. Phase 2 and 3 Combination Designs to Accelerate Drug Development. Journal of the American Statistical Association. 2005;100(470):493–502. Available from: http://amstat.tandfonline.com/doi/abs/10.1198/016214504000001790.

24. Schmidli H, Bretz F, Racine-Poon A. Bayesian predictive power for interim adaptation in seamless phase II/III trials where the endpoint is survival up to some specified timepoint. Statistics in medicine. 2007;26(27):4925–4938.

25. Whitehead J. The Design and Analysis of Sequential Clinical Trials. Chichester: Wiley; 1997.

26. Posch M, Bauer P. Adaptive Two Stage Designs and the Conditional Error Function. Biometrical Journal. 1999;41:689–696.

27. Mehta C, Schäfer H, Daniel H, Irle S. Biomarker driven population enrichment for adaptive oncology trials with time to event endpoints. Statistics in Medicine. 2014;33(26):4515–4531.

28. Proschan MA, Lan KKG, Wittes JT. Statistical Monitoring of Clinical Trials. New York: Springer; 2006.

29. Wang L, Pötzelberger K. Boundary crossing probability for Brownian motion and general boundaries. Journal of Applied Probability. 1997;34:54–65.

30. Proschan MA, Follmann DA, Waclawiw MA. Effects of Assumption Violations on Type I Error Rate in Group Sequential Monitoring. Biometrics. 1992;48:1131–1143.

31. Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical trials: conditional or predictive power? Controlled clinical trials. 1986;7(1):8–17.

32. Tsiatis AA, Mehta C. On the Inefficiency of Adaptive Design for Monitoring Clinical Trials. Biometrika. 2003;90:367–378.

33. Jennison C, Turnbull BW. Adaptive and Nonadaptive Group Sequential Tests. Biometrika. 2006;93:1–21.

34. Bauer P, Koenig F. The reassessment of trial perspectives from interim data—a critical view. Statistics in Medicine. 2006;25(1):23–36.

35. Jennison C, Turnbull BW. Mid-course sample size modification in clinical trials based on the observed treatment effect. Statistics in Medicine. 2003;22(6):971–993.

36. Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: A practical guide with examples. Statistics in medicine. 2011;30(28):3267–3284.

37. Lagakos S, Schoenfeld D. Properties of proportional-hazards score tests under misspecified regression models. Biometrics. 1984;p. 1037–1048.

38. Elsäßer A, Regnstrom J, Vetter T, Koenig F, Hemmings RJ, Greco M, et al. Adaptive clinical trial designs for European marketing authorization: a survey of scientific advice letters from the European Medicines Agency. Trials. 2014;15(1):383.

39. Ravandi F, Ritchie EK, Sayar H, Lancet JE, Craig MD, Vey N, et al. Vosaroxin plus cytarabine versus placebo plus cytarabine in patients with first relapsed or refractory acute myeloid leukaemia (VALOR): a randomised, controlled, double-blind, multinational, phase 3 study. The Lancet Oncology. 2015;16(9):1025–1036.

40. Food and Drug Administration. Draft Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics. 2010. Available from: http://www.fda.gov/downloads/Drugs/Guidances/ucm201790.pdf.

41. European Medicines Agency Committee for Medicinal Products for Human Use. Reflection Paper on Methodological Issues in Confirmatory Clinical Trials Planned with an Adaptive Design. 2007. Doc. Ref. CHMP/EWP/2459/02. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003616.pdf.

42. Cuffe RL, Lawrence D, Stone A, Vandemeulebroecke M When is a seamless study desirable? Case studies from different pharmaceutical sponsors. Pharmaceutical statistics. 2014;13(4):229–237.

43. Brannath W, Zuber E, Branson M, Bretz F, Gallo P, Posch M, et al. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. Statistics in Medicine. 2009;28(10):1445–1463.

44. Carreras M, Gutjahr G, Brannath W. Adaptive seamless designs with interim treatment selection: a case study in oncology. Statistics in Medicine. 2015;34(8);1317–1333.