The Stata Journal (2009)
9, Number 4, pp. 505–523

A menu-driven facility for sample-size calculation in novel multiarm, multistage randomized controlled trials with a time-to-event outcome

Friederike M.-S. Barthel
Oncology R&D, GlaxoSmithKline
London, United Kingdom
sophie@fm-sbarthel.de

Patrick Royston
MRC Clinical Trials Unit
London, United Kingdom
pr@ctu.mrc.ac.uk

Mahesh K. B. Parmar
MRC Clinical Trials Unit
London, United Kingdom
mp@ctu.mrc.ac.uk

Abstract. We present menu- and command-driven Stata programs for the calculation of sample size, number of events, and trial duration for a novel type of clinical trial design with a time-to-event outcome and two or more experimental arms. The approach is based on terminating accrual of patients to inferior experimental treatment arms at an early stage in the trial, allowing through to the next stage only treatments that show a predefined degree of advantage against the control treatment. The first stage of testing uses an intermediate outcome measure for the definitive (primary) outcome rather than with the primary outcome itself. The experimental arms are compared pairwise with the control arm according to the intermediate outcome measure. Arms that survive the comparison enter the next stage of patient accrual, culminating in comparisons against control on the primary outcome measure.

The features supported include unequal patient allocation, target hazard ratios that may differ from 1 under the null hypothesis, and the ability to stop patient recruitment at a specified time after trial initiation. The computations of sample size and power are based on the asymptotic mean and variance of the log hazard-ratio under the null and alternative hypotheses. The overall operating characteristics are computed from the intermediate and final stage significance levels and power, and the correlation between the log hazard-ratios on the intermediate and primary outcome measures at the different stages. We illustrate the approach with the design of a United Kingdom Medical Research Council six-arm trial in prostate cancer in which the intermediate outcome is failure-free survival and the primary outcome is overall survival.

Keywords: st0175, nstage, nstagemenu, multiple arms, randomized controlled trial, survival analysis, surrogate marker, multistage trial

1 Introduction

With the increasing pace of drug development in cancer and other diseases, it is not unusual for several promising treatment regimens to be sufficiently mature for simultaneous testing in large-scale, randomized Phase II/III trials. Limiting factors, such as the time needed to transfer research results to clinical practice and a narrow "window of opportunity", may make it infeasible to perform trials to test such regimens sequentially or in parallel against a control treatment in a traditional two-arm, parallel group design. In this article, we illustrate an approach to designing trials with multiple experimental arms and a single control arm. Further details of the background and the methodology are given by Royston, Parmar, and Qian (2003), with extension to more than two stages in Royston et al. (n.d.). The latter article also examines the robustness of the methods through simulation studies. We present menu-driven Stata software to assist with the sample-size calculations. The approach is based on stopping accrual to inferior arms at an intermediate stage of testing, allowing through to a final stage only treatments that show a predefined degree of advantage against the control treatment. The intermediate stages use an intermediate outcome measure for the primary outcome of interest. The intermediate outcome must be on the causal path to the definitive outcome but does not have to be a true "surrogate" for the latter in the strict sense (see Buyse and Molenberghs [1998] for definitions of surrogacy). Such intermediate endpoints are characterized by a high negative predictive value but not necessarily a high positive predictive value for screening new therapies. Negative predictive value is defined such that if there is no effect of treatment on the intermediate outcome measure (i.e., the null hypothesis is true), then there will be no effect on the primary outcome measure. An example of an intermediate endpoint in cancer studies is disease progression (progression-free survival time). The primary outcome in such studies is typically (time to) death.

The experimental arms that survive comparison with the control arm on the intermediate outcome measure enter a final stage of patient accrual, culminating in comparisons against control on the outcome measure of primary interest. In practice, such a design may be realized by considering hypothetically distinct trials at each stage, each with its own operating characteristics. The overall operating characteristics are computed from the intermediate and final stage significance level and power, and the correlation between the treatment effects (log hazard-ratios) on the intermediate and primary outcome measures at the different stages, assuming a multivariate normal distribution. The correlation may be estimated by bootstrap analysis of individual patient data from previous trials in the same disease.

We outline the basics of the design, present the menu, and work through its associated dialog box, discussing the concepts and parameters. This is followed by an example in prostate cancer and some conclusions. Some details of the algorithm used in the underlying ado-files are given in the technical appendix in section 6 of this article.

2 Outline of the multiarm, multistage (MAMS) trial design

The general idea of the MAMS design is straightforward. Assume that the principal outcome measure in a clinical trial is time to a disease-related event, D, commonly death. We also require a time-related intermediate outcome, I, for example, disease progression. Taking $I = D$ is also a valid option.

Suppose that $k \geq 1$ experimental treatments are to be compared with a control treatment, C. The design is realized through pairwise comparisons between each experimental arm and the control arm. Let E denote an experimental arm, and let Δ_i be the hazard ratio (HR) between E and C on I at the ith stage ($i < s$). Also, let Δ_s be the HR between E and C on D at the sth stage. The null hypothesis and alternative hypothesis for each E/C comparison are as follows:

$$\begin{aligned} H_{0i} &: \Delta_i = \Delta_i^0 \text{ for all } i = 1, \ldots, s \\ H_{1i} &: \Delta_i = \Delta_i^1 \text{ for all } i = 1, \ldots, s \end{aligned}$$

An E that is superior to C will have $\Delta_i < \Delta_i^0$. Typically, $\Delta_i^0 = 1$ and $\Delta_i^1 = \Delta^1$ (i.e., a constant value) for all $i = 1, \ldots, s$. In cancer trials, a characteristic value of Δ^1 is 0.75. The trial proceeds in a maximum of s stages, as follows.

For each stage $i = 1, \ldots, s-1$ and for a given experimental arm E:

1. Determine a critical HR, δ_i, for rejecting H_{0i} at stage i, and determine a threshold number, e_i, of I events in the control arm.
2. Randomize patients between E and C according to a fixed allocation ratio. Continue randomization until e_i I events have been observed in the control arm. e_i is cumulative and therefore includes events in control-arm patients recruited at stages before i.
3. At the end of stage i—that is, when e_i events have been observed in the control arm—compute the HR $\widehat{\Delta}_i$ on the accumulated data. If $\widehat{\Delta}_i < \delta_i$, then recruitment to E continues to the next stage. If $\widehat{\Delta}_i \geq \delta_i$, then recruitment to E ceases (but patients in arm E still continue to be followed up).

If no E survives the final test at stage $s-1$, the trial is terminated. Otherwise, the trial proceeds with the remaining Es to the final stage:

1. Determine a critical HR, δ_s, for rejecting H_0 on the D outcome, and determine a threshold number, e_s, of D events in the control arm.
2. Randomize patients between each remaining E and C according to the fixed allocation ratio. Continue randomization until e_s D events have been observed in the control arm. e_s is cumulative and therefore includes D events in control-arm patients recruited at earlier stages before i.

3. When e_s events have been observed in the control arm, compute the HR $\widehat{\Delta}_s$ on the D outcome with the accumulated data. If $\widehat{\Delta}_s < \delta_s$, the null hypothesis H_{0s} is rejected.

To limit the total number of patients in the trial, an option is to stop recruitment at a predefined time, t_{stop}, during the final stage. Stopping recruitment early increases the length of the final stage.

The above scheme illustrates the basic principles of how an MAMS trial would be conducted. Details of the calculations are given by Royston, Parmar, and Qian (2003) and Royston et al. (n.d.). In the following section, we describe the inputs required by the dialog box, with comments on the design concepts and parameters as necessary.

3 Design of menu and dialog box

The **n-stage trial** menu is initiated by entering `nstagemenu on` in the Stata command window. A new item, **n-stage trial**, appears on the user menu. The menu is turned off by entering `nstagemenu off`. At present, the **n-stage trial** menu comprises just one choice, **Multi-Stage Trial Designs**, which is used to design trials whose intermediate and primary outcome measures are both based on time-to-event outcomes. Designs such as binary/survival in which the intermediate outcome is binary, e.g., biological response of a tumor to treatment, may be added to the software later.

All the features are available through the *Multi-Stage Trial Designs* dialog box. When the computations are complete, Stata displays in the Review window the command line that generated the results. The computations are performed by an ado-file called `nstage`, which is provided with this article. By recalling the command from the Review window, editing it, and re-executing it, the menu system may also be used as a tutor for the command-driven approach using `nstage`. This command-driven method is important for the documentation and reproducibility of the design and its associated parameters by using Stata do-files and log files. We suggest that the user open a log file before executing the commands via the dialog box, which will hence save the command line. This log file can then be edited to produce a do-file to repeat the calculations if desired.

The *Multi-Stage Trial Designs* dialog box, with the input choices presented for illustration purposes only, is shown in figure 1. Table 1 shows all the parameters required to be input into the program by the user. The program outputs are summarized in table 2.

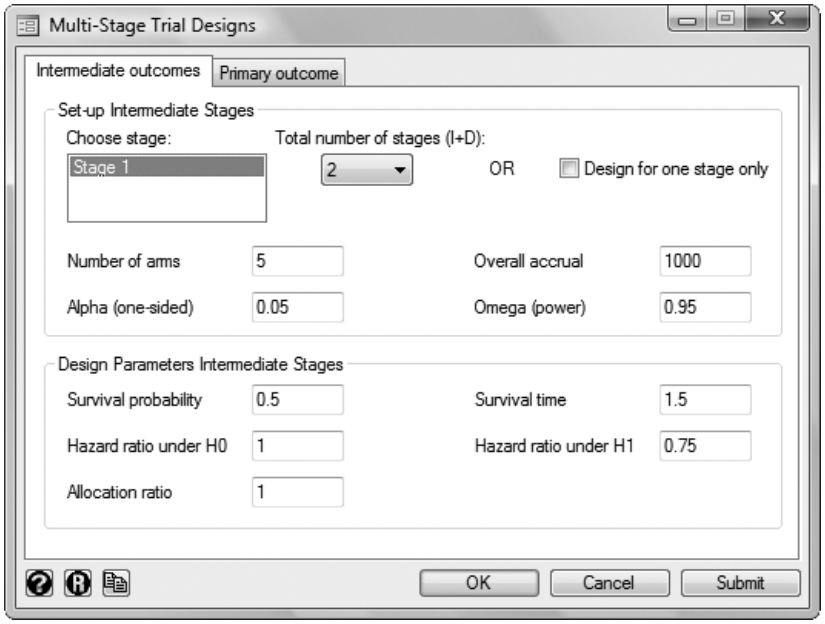

Figure 1. Screenshot of *Multi-Stage Trial Designs* dialog box

Table 1. User inputs required by the *Multi-Stage Trial Designs* dialog box and its associated ado-file, `nstage`. Single default values apply to all stages; pairs of values apply to stages 1 and 2, respectively. The maximum number of stages permitted by the dialog box is five.

Parameter	Default
Total number of stages (I and D together)	2
Number of arms in each of the stages	$5, 2$
Total accrual rate (all arms) per unit time	1000
Allocation ratio	1
Survival probability	0.5
Survival time	$1.5, 3$
HR under H_0	1
HR under H_1	0.75
One-sided significance level (alpha) for each stage	$0.05, 0.025$
Power (omega) for each stage	$0.95, 0.9$
Time of stopping accrual, t_{stop} (if required)	optional
Time units	1 (= one year)
Correlation between log HRs	0.6

Table 2. Results output by the `nstage` program

Result	Notation*
Critical HR at ith stage	δ_i
Time to the end of ith stage	t_i
Number of control arm events required at ith stage	e_i
Number of patients in each arm at ith stage and total	–
Overall significance level	α
Overall power	$1-\beta$

*See the technical appendix.

We describe the required user inputs in turn. The inputs for the **Intermediate outcomes** and **Primary outcome** parts of the dialog box window are similar.

3.1 Number of arms

The design assumes a single control arm to which patients are recruited throughout the intermediate stages and, provided that at least one experimental arm shows a sufficient and consistent effect for the trial to continue, the final stage. Choosing the number of experimental arms at stage 1 is a fundamental decision for the trial designer and depends on the number of novel treatments available. The number of experimental arms progressing to the next stages in reality depends on the effectiveness of the new treatments compared with the control, but this number needs to be specified to allow the sample size to be calculated. We suggest trying a number of scenarios to see the implications for the sample size, power, and duration of the trial.

The software expects the *total* number of arms (C + all E) to be input for each stage. For example, a design with four experimental arms would be entered as 5 at stage 1.

3.2 Overall accrual rate

Patients are assumed to be recruited to the trial at a steady (uniform) rate in each stage. The rate of overall accrual is the number of patients entered per unit time, typically, per year. The program requires the total rate across all arms (C + all E). The accrual rate per arm is assumed to be the same for each arm, but this can be varied through the *Allocation ratio* option. The program allows the user to enter a different accrual rate for each stage; for example, a trial may recruit at a higher rate once it has passed the first stage(s).

For example, with five arms ($C + 4E$) and $1,000$ patients recruited per year at stage 1, 200 new patients would enter each arm each year for the duration of stage 1. With two arms (C + 1 remaining E) and $1,000$ patients recruited per year at stage 2 (final stage), 500 additional patients per year would enter each of the two active arms

for the duration of stage 2. Follow-up would continue on all patients until the point of data analysis at the end of stage 2.

3.3 Alpha (one-sided)

These inputs define the one-sided significance level for each pairwise comparison with control at each stage. One-sided type I errors are natural at the intermediate stages because we have no interest in experimental arms that are worse than control except on ethical and safety grounds. The intermediate stages act only as a filter for unsuccessful experimental treatments, allowing resources to be concentrated on the better treatments at the following intermediate and final stages. Hence, the program expects one-sided error probabilities at each stage. The MAMS design stages may be considered as a succession of independent smaller trials. Hence, the significance level and power must be specified at each stage. For a detailed discussion of this approach, please refer to section 2.3 of Royston, Parmar, and Qian (2003).

The overall significance level is computed by the program, based on the significance level specified at each stage and the correlation structure (see section 3.10 below).

For example, the default option is 0.05 (one-sided) at stage 1 and 0.025 (one-sided) at stage 2. The latter is equivalent to a conventional two-sided type I error probability of 0.05 at stage 2.

3.4 Omega (power)

These inputs define omega (the power) required for each pairwise comparison with control at each stage. The power at the intermediate stages should be high enough to ensure that a successful experimental treatment has only a small probability of failing to progress to the next stage. For example, a power of 0.80 is too small for the intermediate stage(s); we suggest 0.95.

For example, the default option is a power of 0.95 at stage 1 and 0.9 at stage 2. With target HRs at stage 1 of 1 and 0.75 under H_0 and H_1, respectively, the threshold HR beneath which a treatment progresses to stage 2 would be reported by the program as a `Crit. HR` of 0.869.

3.5 Allocation ratio

The allocation ratio is the number of patients allocated to each experimental arm for each patient allocated to the control arm. The ratio must be the same at all stages. By default, the allocation of patients is equal for each of the treatment arms, i.e., an allocation ratio of 1.

(*Continued on next page*)

For example, an allocation ratio of 0.5 means that each experimental arm would receive half as many patients as the control arm. If the overall accrual rate was 1,000 per year and the total number of arms was 4, then C would receive 400 and each E would receive 200 patients per year.

3.6 Survival probability

The survival probability is the probability that a control arm patient survives to a given time specified by the *Survival time* inputs. The default survival probability is 0.5, in which case the survival times are the medians. Two values are required, one for the intermediate outcome and one for the primary outcome.

3.7 Survival time

The survival time is the benchmark survival time of a patient in the control group corresponding to the value chosen in *Survival probability*. This may be estimated from previous trials. For the MAMS design to be maximally effective, the median survival time for intermediate outcome events should be substantially shorter than for primary outcome events (unless, of course, $I = D$, i.e., the same outcome is used at all stages). Over a given period, there will then be more events for the intermediate than for the primary outcome, shortening the lead-time for discarding unsuccessful treatments. The underlying survival distribution is assumed to be exponential. The survival probability and survival time values are translated into hazard values according to the exponential assumption. Time must be expressed in the same units as the accrual rate (by default, years).

For example, in advanced ovarian cancer, typical median survival times may be approximately 1.5 and 3 years for events for progression-free survival and overall survival, respectively (McGuire et al. 1996).

3.8 HR under H_0

This is the target HR for each experimental arm relative to the control arm, under the null hypothesis of no advantage of an experimental treatment. An HR < 1 represents benefit of an experimental arm relative to control. The HR for the intermediate stages is for events on the intermediate outcome; for the final stage, the HR is for events on the primary outcome. Although typically the null hypothesis is HR $= 1$ for each stage, the possibility of specifying HR < 1 for the intermediate stages might be entertained. This would reflect a situation in which an experimental treatment gave a temporary advantage over control not maintained on the primary outcome. See Royston, Parmar, and Qian (2003) for further comments on this aspect.

3.9 HR under H_1

This is the target HR for each experimental arm relative to the control arm, under the alternative hypothesis that an experimental treatment is better than control (i.e., has HR < 1). All comparisons are considered pairwise with control, and the same target HR is assumed for each experimental arm. The HRs entered here for the intermediate and primary outcome are crucial to the design and should be considered carefully.

For example, in common cancers, where modest improvements in survival are usually all that can be hoped for, a typical HR under H_1 is 0.75. For reasons noted above, it may be appropriate to reduce the target HR at earlier stages, e.g., from 0.75 to 0.70. However, this will tend to reduce the overall sample size and the duration of the trial, so it should be used with caution and only if supported by good evidence.

3.10 Correlation

The correlation relevant to MAMS designs is not encountered in conventional trials. It measures the strength of association between the treatment effects on the I and D outcomes at a fixed time (e.g., the end of follow-up). The correlation can be estimated by applying bootstrap analysis to trial data similar to that expected in the new trial (Royston, Parmar, and Qian 2003). We suggest a default value of 0.6 for this parameter. If you have no idea of the value, we suggest a sensitivity analysis in the range (0.4, 0.8). The correlation value affects only the overall significance level and power of the design. Further suggestions as to plausible determination of the correlation are given by Royston et al. (n.d.). If you have only one outcome type (i.e., $I = D$), the correlation is optional because the program knows how to calculate the necessary correlation structure.

3.11 Time of stopping accrual

The time of stopping accrual (t_{stop}) is the point at which patient accrual is to cease, in the same units used to define the accrual rate. The time scale runs from the start of the trial. If t_{stop} is omitted (by leaving the edit box empty), accrual is assumed to continue until the target number, e_s, of D events has accumulated in the control arm. If t_{stop} is specified, recruitment stops at that point, but follow-up and accumulation of events continues until e_s events have been observed, and the duration of the final stage is adjusted accordingly. If t_{stop} is set too early, an error message is given, stating that the specified period is too short to accumulate the required events.

For example, a trial in advanced cancer might recruit patients for four years and then continue follow-up awaiting the required events.

(*Continued on next page*)

3.12 Show probabilities for number of arms in each stage

Checking this box displays a table of approximate probabilities of $0, 1, \ldots$ experimental arms progressing to the next stage, under the null and alternative hypotheses. As already noted, to progress, an arm must have an HR on the intermediate outcome less than a critical value (reported by the program as `Crit. HR`). The chance of one or more arms passing under H_0 is equal to one minus the reported probability of 0 arms passing.

4 Example. Systemic therapy in advancing or metastatic prostate cancer: Evaluation of drug efficacy (STAMPEDE) trial in prostate cancer

4.1 Design

STAMPEDE is a MAMS trial for men with prostate cancer conducted from the MRC Clinical Trials Unit. The aim of the trial is to assess drugs from three different classes for men starting androgen suppression, the standard treatment for high-risk, hormone-sensitive disease. Five experimental arms are compared with a control of androgen suppression alone in five stages. A randomized pilot phase is carried out prior to the efficacy stages to confirm feasibility and safety of treatments when used in combination with androgen suppression. Hence in terms of the MAMS design and its calculations, we are dealing only with four stages (so that $s = 4$). Stages 1 to 3 are a randomized comparison of compounds shown to be safe using the intermediate outcome measure of failure free survival. The final analysis is then carried out in stage 4 as a comparison of all those arms still recruiting after stage 3 with the control based on overall survival as the primary outcome measure.

The basic design parameters of this trial are set out in table 3.

Table 3. Design parameters for STAMPEDE. The median survival times were assumed to be 2 years and 4 years for failure-free and overall survival, respectively. The HRs at all stages under the null hypothesis were 1.0 and under the alternative hypothesis were 0.75. The allocation ratio was 0.5, i.e., 0.5 patients are allocated to each experimental arm for every 1 patient to the control arm.

Stage i	Critical HR δ_i	Sig. level α_i	Power $1 - \beta_i$
1	1.000	0.5	0.95
2	0.923	0.25	0.95
3	0.885	0.1	0.95
4	0.844	0.025	0.9

The default correlation of 0.6 was used in the calculations shown above. As is apparent in table 3, high values of the significance levels α_i were chosen for stages 1 to 3. The aim here is to avoid rejecting a potentially promising treatment arm too early in the trial while at the same time rejecting any treatments for which the HRs exceed the critical values δ_i. Because of the parameter values chosen, a treatment should therefore pass from stage 1 to stage 2 if it shows any beneficial effect in comparison with the control arm. A higher significance level early in the trial also means that we will not have to wait too long for the first comparisons while maintaining a reasonable power. With the correlation of 0.6, the overall pairwise significance level (for each comparison of an experimental arm with control) is calculated to be 0.0118 and overall pairwise power is 0.833.

The filled-out dialog boxes in figures 2 and 3 illustrate how the design is realized.

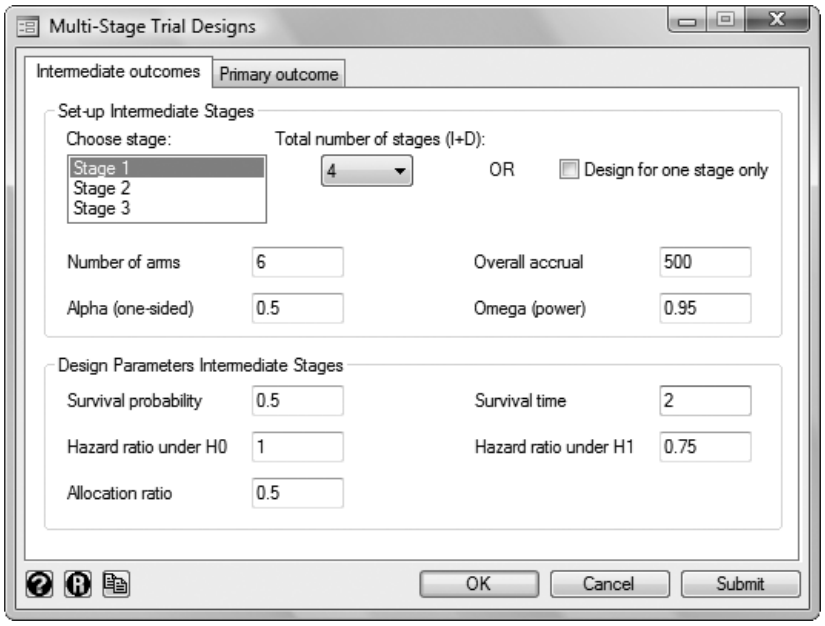

Figure 2. STAMPEDE inputs for stage 1 (other intermediate stages are similar)

(*Continued on next page*)

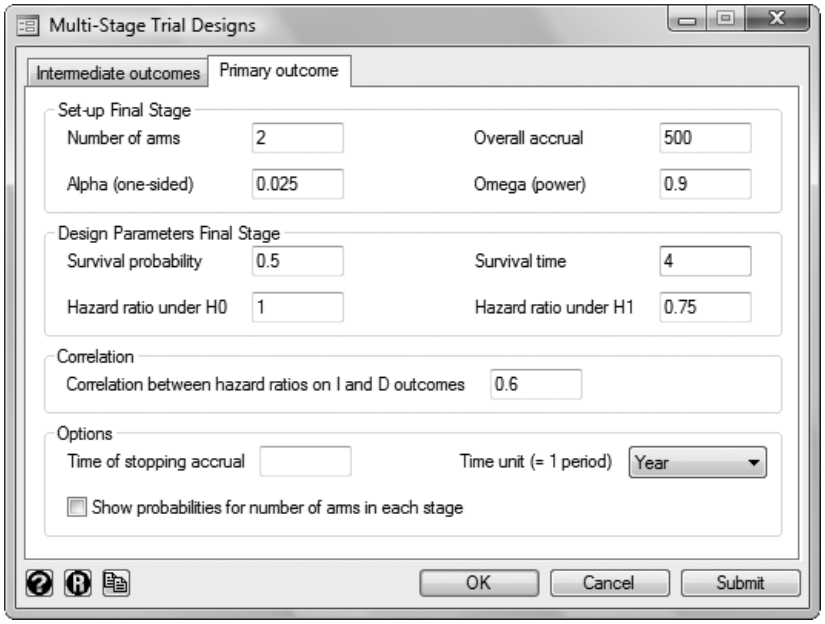

Figure 3. STAMPEDE inputs for final stage

We will take the specification as given and explore the resulting numbers of patients and events, and the likely duration of the trial.

4.2 Results

On pressing the **Submit** button with the design parameters shown in table 3 as well as, for example, 6, 5, 3, and 2 arms in the four stages respectively, the following output is obtained:

```
N-STAGE TRIAL DESIGN                          version 2.1.0, 28 August 2009
-------------------------------------------------------------------------------
A sample size program for n-stage trial designs by Friederike Barthel &
Patrick Royston, based on Royston, Barthel, Parmar and Oskooei (2009)
-------------------------------------------------------------------------------

OPERATING CHARACTERISTICS

DESIGN FOR 4 STAGES

MEDIAN SURVIVAL TIME (I-OUTCOME): 2 TIME UNITS
MEDIAN SURVIVAL TIME (D-OUTCOME): 4 TIME UNITS
```

	Alpha(1S)	Power	HR|H0	HR|H1	Crit. HR	Length*	Time
STAGE 1	0.5000	0.950	1.000	0.750	1.000	2.436	2.436
STAGE 2	0.2500	0.951	1.000	0.750	0.924	1.078	3.514
STAGE 3	0.1000	0.950	1.000	0.750	0.886	0.919	4.433
STAGE 4	0.0250	0.900	1.000	0.750	0.845	1.594	6.027
Overall**	0.0118	0.833				6.027	
Lowest	0.0020	0.809					
Highest	0.0250	0.900					
I-stages	0.0799	0.899					

```
 *  Length (duration of each stage) is expressed in one year periods
 ** Correlations between hazard ratios estimated internally by the program
    assuming corhr(), correlation between hazard ratios on I & D, is 0.60

SAMPLE SIZE AND NUMBER OF EVENTS
```

	STAGE 1			STAGE 2		
	Overall	Control	Exper.	Overall	Control	Exper.
Arms	6	1	5	5	1	4
Acc. rate	500	143	357	500	167	333
Patients*	1218	348	870	1757	528	1229
Events**	343	113	230	572	216	356

	STAGE 3			STAGE 4		
	Overall	Control	Exper.	Overall	Control	Exper.
Arms	3	1	2	2	1	1
Acc. rate	500	250	250	500	333	167
Patients*	2216	757	1459	3014	1289	1725
Events**	612	334	278	568	405	163

```
0.5 patients allocated to each E arm for every 1 to control arm.
 *  Patients are cumulative across stages
 ** Events are cumulative across stages, but are only displayed
    for those arms to which patients are still being recruited
 ** Events are for I-outcome at stages 1 to 3, D-outcome at stage 4
```

This design, as illustrated in the output for stage 4, requires $3,014$ patients with 405 D events (deaths) in the control arm. However, because we cannot know in advance how many arms will pass each stage, all possible scenarios would need to be considered. The calculation should be run with 2, 3, 4, and 5 arms in each of stages 2–4 to get an adequate idea of the sample size required, depending on how many arms pass each stage. For STAMPEDE, the total sample size if 4 stages are conducted turns out to be in the range $2,800$ to $3,600$. We suggest doing such calculations while the trial protocol is being

developed, to plan for adequate resources in all circumstances. Similarly, sensitivity analyses should be performed to gauge the effects of variations in recruitment rates and other key inputs.

The `nstage` command to generate the above output is as follows:

```
. nstage, accrue(500 500 500 500) arms(6 5 3 2) hr0(1 1) hr1(0.75 0.75)
> alpha(0.5 0.25 0.1 0.025) omega(0.95 0.95 0.95 0.9) t(2 4) s(0.5 0.5)
> aratio(0.5) nstage(4) tunit(1) corhr(0.6)
```

One important point concerning the output from `nstage` is that the number of events reported in the experimental arms at each stage reflects the number of arms "surviving" to that stage. In the above example, 2 experimental arms survive to stage 3 and are expected to have accrued a total of 278 I events by that time point (4.4 years). At the end of the trial (6.0 years), in stage 4, only 163 D events are expected in the one surviving arm. The program does not report the events in the four "dropped" arms.

4.3 Recruiting up to a fixed time point

In table 4, we show the effect on trial duration, total sample size, and number of D events by stopping patient recruitment at progressively earlier time points, t_{stop}. We use the STAMPEDE design as an example.

Table 4. Effect on the duration of the STAMPEDE trial and its sample size of varying t_{stop}, the time of stopping patient recruitment

t_{stop} (year)	t_s (duration, year)	Total patients	D events	
			Total	Control arm
4.5	6.9	2,250	569	403
5.0	6.3	2,500	568	404
5.5	6.1	2,750	568	405
6.0	6.0	3,000	568	405

All other trial design parameters were fixed at the values in table 3. Ceasing recruitment after 5 years may be a good option, because the overall duration is only increased by about 0.3 years compared with a policy of continuous recruitment until the end of stage 4. There is a worthwhile reduction of 500 patients. Stopping earlier than 4.43 years (the end of stage 3) is not considered feasible, and the software reports this as an error.

5 Conclusion

In an age of increasing numbers of potentially effective treatments requiring rapid evaluation and a restricted patient population available for trials, it is important that the design of a randomized controlled clinical trial be efficient. The software provided here

will allow researchers to explore the new MAMS designs proposed by Royston, Parmar, and Qian (2003) and Royston et al. (n.d.) in a flexible and user-friendly manner. As we saw in the example, a worthwhile reduction in patient numbers may be available by a careful choice of the time of stopping recruitment. In due course, we plan to extend the present software to accommodate designs in which either stage is based on a binary outcome. An example in cancer is tumor response to chemotherapy as the intermediate outcome.

With MAMS designs, particular care needs to be taken in terms of specifying the rate of accrual in all stages and the number of arms. As has become apparent during the conduct of GOG0182-ICON5 (Bookman, M. A., for the Gynecologic Cancer InterGroup (GCIG) 2006; International Collaborative Ovarian Neoplasm (ICON) Group 2002), specifying a lower accrual rate in stage 1 than occurs in reality means that the stage 1 analysis may become infeasible. If patients are recruited "too quickly", accrual for stage 2 will start before the necessary events for a stage 1 analysis have accumulated. This may happen because accrual is not stopped while the analysis at each stage is conducted to ease the operational burden. In the event that the trial is stopped at the stage 1 analysis, the design becomes inefficient because more patients than necessary have entered the analysis.

Practical considerations surrounding these designs are discussed by Parmar et al. (2008). Recent experience has shown that although the designs seem complex at first acquaintance, with further information and experience, patients, clinicians, and industry partners alike appreciate the merits of this type of design in evaluating new agents.

The authors have carried out a series of case studies, the results of which are presented by Barthel, Parmar, and Royston (2009). Data from completed cancer trials conducted at the MRC Clinical Trials Unit were reanalyzed in a counterfactual manner as though they were MAMS designs. The results were positive in terms of reduction in trial time and acceptability of type 1 and type 2 error rates.

6 Technical appendix

In this section, we note some details of the mathematics behind the computations and the algorithms used by `nstage.ado`, the program that underlies the `n-stage Trial Design` dialog box, to compute sample size, number of events, and duration of the trial. Further details are given in appendix A of Royston et al. (n.d.).

We continue the notation of section 2; see table 5 for a summary of the notation.

(*Continued on next page*)

Table 5. Notation for quantities used in MAMS designs

Symbol	Meaning
δ_i	Critical HR at stage i, $i = 1, ..., s$
Δ_j^0	Target HR under H_0 for outcome of type j, $j = I, D$
σ_i^0	Standard error of estimated log HR at stage i under H_0
z_{α_i}	Normal equivalent deviate of one-sided significance level α_i
Δ_j^1	Target HR under H_1 for outcome of type j, $j = I, D$
σ_i^1	Standard error of estimated log HR at stage i under H_1
$z_{1-\beta_i}$	Normal equivalent deviate of power $1 - \beta_i$
e_i	No. of events in control arm required to terminate stage i
e_i^*	No. of events in an experimental arm under H_1 with e_i events in control arm
λ_j	Hazard of an event for outcome j
r_i	Rate of recruitment to the control arm in stage i
t_i	Calendar time at the end of stage i

Assuming that we have specified the significance level, power, and target HRs in all stages, we need to calculate the cutoff δ_i as well as e_i, the number of control arm events needed in all stages. Let Φ^{-1} denote the inverse standard normal distribution function. By definition, for all stages $i < s$, we have

$$\begin{aligned} z_{\alpha_i} &= \frac{\ln \delta_i - \ln \Delta_I^0}{\sigma_i^0} \\ &= \Phi^{-1}(\alpha_i) \end{aligned}$$

and

$$z_{\alpha_s} = \frac{\ln \delta_s - \ln \Delta_D^0}{\sigma_s^0} = \Phi^{-1}(\alpha_s)$$

Similarly, under H_1, we have

$$\begin{aligned} z_{1-\beta_i} &= \frac{\ln \delta_i - \ln \Delta_i^1}{\sigma_i^1} \\ &= \Phi^{-1}(1 - \beta_i) \end{aligned}$$

and

$$z_{1-\beta_s} = \frac{\ln \delta_s - \ln \Delta_D^1}{\sigma_s^1} = \Phi^{-1}(1 - \beta_s)$$

According to Tsiatis (1981), assuming an allocation ratio of unity between the control arm and each experimental arm, the variance of an observed HR at stage i under H_0 may be approximated using the following formula:

$$(\sigma_i^0)^2 = (\sigma_i^1)^2 = \frac{2}{e_i}$$

It follows that for $i = 1, \ldots, s-1$,

$$e_i = \frac{2(z_{\alpha_i} - z_{1-\beta_i})^2}{(\ln \Delta_I^1 - \ln \Delta_I^0)^2} \tag{1}$$

and

$$e_s = \frac{2(z_{\alpha_s} - z_{1-\beta_s})^2}{(\ln \Delta_D^1 - \ln \Delta_D^0)^2} \tag{2}$$

(1) and (2) for the number of control arm events are based on an estimate of the variance under H_0. Because fewer events are expected under H_1, multiplying the events by the number of arms in the trial underestimates the sample size needed to achieve power $1 - \beta_i$. The following algorithm included in `nstage` corrects the underestimation:

1. Calculate e_i, the required number of I events in the control arm under H_0.
2. Calculate the critical log HR $\ln \delta_i = \ln(\Delta_i^0) + z_{\alpha_i}\sigma_i^0$.
3. Calculate the time, t_i, needed to run the trial until the end of stage i.
4. Calculate e_i^*, the number of events in the experimental arm(s) under H_1 by the end of stage i (assuming an exponential survival distribution).
5. Calculate the power for stage i conferred by e_i and e_i^*.
    a. If power is less than needed, increment e_i by 1 and return to step 2.
    b. If power is as desired, terminate the algorithm.

Regarding the assumption of exponential survival, Royston, Parmar, and Qian (2003) provide evidence that the duration of the intermediate stages is robust to departures from exponentiality. A "reasonable" estimate of the average hazard, λ_I, is needed, preferably based on the early part of the distribution of time to the intermediate outcome, because this will affect stage 1 the most. Such an estimate may be obtained by fitting an exponential distribution to individual data for patients similar to those expected in the control arm in the new trial. If individual patient data are unavailable, λ_I may be crudely estimated from a published survival curve or from the survival probability, $S(t)$, at a given time, t, by using the formula $\lambda_I = -\log S(t)/t$. When t is the median survival time, then $S(t) = 0.5$.

For the exponential distribution, Royston, Parmar, and Qian (2003) show that the number of events e_i accrued by time t_i is given by

$$e_i = r_i \left(t_i - \frac{1 - e^{-\lambda_I t_i}}{\lambda_I} \right) \tag{3}$$

Step 3 of the algorithm requires t_i to be determined from the current value of e_i, which necessitates inversion of (3) for stage 1 and the extensions to this formula for more than two stages described in Royston et al. (n.d.). The inversion is implemented

in the supplied ado-file `timetoevn4` and uses successive numerical approximation by Newton's method. The supplied ado-file `evfromtin4` performs the calculation in the other direction, i.e., computes the number of events from the time and the other two parameters.

7 Acknowledgment

The authors would like to thank Matthew Sydes for his helpful input on the STAMPEDE section of the article.

8 References

Barthel, F. M.-S., M. K. B. Parmar, and P. Royston. 2009. How do multi-stage, multi-arm trials compare to the traditional two-arm parallel group design—a reanalysis of 4 trials. *Trials* 10: 21–31.

Bookman, M. A., for the Gynecologic Cancer InterGroup (GCIG). 2006. GOG0182-ICON5: 5-arm phase III randomized trial of paclitaxel (P) and carboplatin (C) vs combinations with gemcitabine (G), PEG-lipososomal doxorubicin (D), or topotecan (T) in patients (pts) with advanced-stage epithelial ovarian (EOC) or primary peritoneal (PPC) carcinoma. 2006 ASCO Annual Meeting Proceedings Part I. *Journal of Clinical Oncology* 24(18S): 5002.

Buyse, M., and G. Molenberghs. 1998. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* 54: 1014–1029.

International Collaborative Ovarian Neoplasm (ICON) Group. 2002. Paclitaxel plus carboplatin versus standard chemotherapy with either single-agent carboplatin or cyclophosphamide, doxorubicin, and cisplatin in women with ovarian cancer: The ICON3 randomised trial. *Lancet* 360: 505–515.

McGuire, W. P., W. J. Hoskins, M. F. Brady, P. R. Kucera, E. E. Partridge, K. Y. Look, D. L. Clarke-Pearson, and M. Davidson. 1996. Cyclophosphamide and cisplatin compared with paclitaxel and cisplatin in patients with stage III and stage IV ovarian cancer. *New England Journal of Medicine* 334: 1–6.

Parmar, M. K. B., F. M.-S. Barthel, M. Sydes, R. Langley, R. Kaplan, E. Eisenhauer, M. Brady, N. James, M. A. Bookman, A.-M. Swart, W. Qian, and P. Royston. 2008. Speeding up the evaluation of new agents in cancer. *Journal of the National Cancer Institute* 100: 1204–1214.

Royston, P., F. M.-S. Barthel, M. K. B. Parmar, and B. C. Choodari-Oskooei. n.d. Designs for clinical trials with time-to-event outcomes based on stopping guidelines for lack of efficacy. Forthcoming.

Royston, P., M. K. B. Parmar, and W. Qian. 2003. Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Statistics in Medicine* 22: 2239–2256.

Tsiatis, A. A. 1981. The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika* 68: 311–315.

About the authors

Friederike Barthel is a senior statistician in Oncology R&D at GlaxoSmithKline. Previously, she worked at the MRC Clinical Trials Unit and Institute of Psychiatry. Her current research interests include sample-size issues, particularly concerning multistage, multiarm trials, microarray study analyses, and competing risks. Friederike has taught undergraduate courses in statistics at the University of Westminster and Kingston University.

Patrick Royston is a medical statistician with 30 years of experience, with a strong interest in biostatistical methods and in statistical computing and algorithms. He now works in cancer clinical trials and related research issues. Currently, he is focusing on problems of model building and validation with survival data, including prognostic factor studies; on parametric modeling of survival data; on multiple imputation of missing values; and on novel clinical trial designs.

Mahesh Parmar is the director designate of the MRC Clinical Trials Unit. He has wide-ranging interests in the design, running, analysis, and execution of randomized controlled clinical trials, as well as being involved in statistical methodology, including prognostic research and biomarkers. Mahesh and other colleagues originally proposed the concept of the MAMS trial realized in the present software.