The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$115	1-year subscription	$145
2-year subscription	$210	2-year subscription	$270
3-year subscription	$285	3-year subscription	$375
1-year student subscription	$ 85	1-year student subscription	$115
1-year institutional subscription	$345	1-year institutional subscription	$375
2-year institutional subscription	$625	2-year institutional subscription	$685
3-year institutional subscription	$875	3-year institutional subscription	$965
Electronic only		**Electronic only**	
1-year subscription	$ 85	1-year subscription	$ 85
2-year subscription	$155	2-year subscription	$155
3-year subscription	$215	3-year subscription	$215
1-year student subscription	$ 55	1-year student subscription	$ 55

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.


The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, STATA, Stata Press, Mata, MATA, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2015)
15, Number 2, pp. 350–368

A menu-driven facility for sample-size calculation in multiarm, multistage randomized controlled trials with time-to-event outcomes: Update

Daniel J. Bratton
MRC Clinical Trials Unit
University College London
London, UK
d.bratton@ucl.ac.uk

Babak Choodari-Oskooei
MRC Clinical Trials Unit
University College London
London, UK
b.choodari-oskooei@ucl.ac.uk

Patrick Royston
MRC Clinical Trials Unit
University College London
London, UK
j.royston@ucl.ac.uk

Abstract. Barthel, Royston, and Parmar (2009, *Stata Journal* 9: 505–523) presented a menu-driven Stata program for calculating sample size and other design characteristics for a class of multiarm, multistage trials with time-to-event outcomes. In this article, we present several new features for the package. First, we allow hazard ratios greater than 1 to be targeted under the alternative hypothesis to make the design more widely applicable to other outcomes and disease areas. Second, we introduce new subroutines that use simulation to more accurately estimate the correlation between treatment effects at different stages and that calculate the familywise error rate, and we apply them to an example using the original design of the multiarm, multistage trial in prostate cancer. Finally, we present a new design of the dialog menu for `nstage` that improves its usability and incorporates options for calling the new subroutines.

Keywords: st0175_1, nstage, nstagemenu, multiarm, multistage, randomized controlled trial, survival analysis, time-to-event, familywise error rate

1 Introduction

Royston, Parmar, and Qian (2003) introduced a class of novel designs for multiarm clinical trials with time-to-event outcomes aimed at speeding up treatment evaluation. In their design, experimental arms are compared with a common control arm at an interim analysis, at which point recruitment is stopped in arms that fail to show a sufficient amount of additional benefit. Recruitment continues with promising treatments in the final stage of the trial, after which they are compared with the control on the primary definitive outcome (D) of the trial. To increase the speed of treatment selection, one can

perform the interim analysis on an intermediate outcome (I) that is on the causal pathway to the primary outcome of the trial. For instance, failure-free survival—taken to be time to progression or death, whichever occurs first—was chosen as the intermediate outcome in several oncology trials using this design (Raja et al. 2011; Sydes et al. 2009), with overall survival used as the definitive outcome. The requirements for choosing an appropriate intermediate outcome measure are described by Royston et al. (2011).

To further increase efficiency, Royston et al. (2011) extended this design to more than two stages so that treatment arms can be compared with a control on the intermediate outcome at multiple interim analyses. To facilitate the design of trials using this multistage approach, Barthel, Royston, and Parmar (2009) introduced the menu-driven Stata program `nstage` for calculating the required sample sizes, number of events, stage durations, critical values, and pairwise operating characteristics.

Although this class of multiarm, multistage (MAMS) designs helps expedite the evaluation of new treatments (Parmar et al. 2008), the designs also introduce various statistical and practical challenges. Issues that have been addressed include the application of the design to large-scale multiarm trials (Sydes et al. 2009), the addition of new research arms midtrial (Sydes et al. 2012), an assessment of bias in the estimated treatment effects (Choodari-Oskooei et al. 2013), and extensions to binary intermediate and definitive outcomes (Bratton, Phillips, and Parmar 2013).

A previously unresolved problem is the quantification and control of the probability of finding a false positive result at the end of the trial, known as the familywise error rate (FWER). This measure is essential in many multiarm trials, particularly if the trials are confirmatory (Wason et al. 2013), so a means for accurately estimating the FWER is crucial. Another outstanding issue concerns estimating the between-stage correlation structure when the intermediate and definitive outcomes differ (Royston et al. 2011). Because the accuracy of the current calculation is uncertain, a more reliable method is required to better estimate the pairwise operating characteristics of the trial.

Finally, because `nstage` was originally written with the design of cancer trials in mind, it can use only outcomes such as progression-free and overall survival, where events need to be observed more slowly on an experimental arm than on the control to demonstrate superiority. The program will therefore be extended to outcomes where a shorter time represents a more favorable outcome, such as time to healing (for example, Smith et al. [1992]) or time to cure.

In this article, we describe an update of `nstage` to address these issues. We apply a new simulation method for estimating the between-stage correlation to one of the design scenarios for the systemic therapy in advancing or metastatic prostate cancer: the evaluation of drug efficacy (STAMPEDE) trial in prostate cancer (Sydes et al. 2009). We do this to demonstrate its improved accuracy over the existing calculation. We improve the usability of the dialog menus and add options to accommodate the new methods. Throughout, we use the acronym MAMS to refer to the multiarm, multistage design described by Royston et al. (2011).

2 The nstage command

2.1 Syntax

`nstage`, `nstage(#)` `accrue(`*numlist*`)` `alpha(`*numlist*`)` `omega(`*numlist*`)`
`arms(`*numlist*`)` `hr0(# [ # ])` `hr1(# [ # ])` `t(# [ # ])` `[s(# [ # ])`
`aratio(#)` `tunit(#)` `tstop(#)` `probs` `nofwer` `simcorr(#)` `corr(#)`]

Note: the number of values given in each *numlist* must be equal to the number of stages specified in `nstage(#)`.

2.2 Options

`nstage(#)` specifies the number of trial stages, J. `nstage()` is required.

`accrue(`*numlist*`)` specifies the rate per unit time (see `tunit()`) at which patients enter the trial during each stage. The patients are assumed to be allocated in the ratio (control arm:experimental arm: ...) of $1 : A : \cdots : A$, where A is the allocation ratio defined by `aratio()`. `accrue()` is required.

`alpha(`*numlist*`)` specifies the one-sided significance level, α_j, for each comparison at each stage, j. The arms are compared pairwise with control on the intermediate outcome (I) for the first $J-1$ stages, whereas the comparison is on the primary outcome (D) at the Jth stage. Significance levels should decrease with each stage. `alpha()` is required.

`omega(`*numlist*`)` specifies the power (one minus the type 2 error probability), ω_j, for each pairwise comparison at each stage. `omega()` is required.

`arms(`*numlist*`)` specifies the number of arms assumed to be actively recruiting patients at each stage. The number at each stage cannot exceed the number at the previous stage, because arms can only be "dropped" (not added). For example, `arms(4 3 2)` would say that in a three-stage trial of four arms, only three "survived" to the second stage, and only two survived to the final stage. `arms()` is required.

`hr0(# [ # ])` specifies the hazard ratios (HRs) under the null hypothesis for the I outcome and D outcome, respectively. Typically, these values are both 1. `hr0()` is required.

`hr1(# [ # ])` specifies the target HRs under the alternative hypothesis for the I outcome and D outcome, respectively. Typically, the size of the targeted effect is larger for the I outcome than the D outcome. `hr1()` is required.

`t(# [ # ])` defines the times corresponding to the survival probabilities in `s()` for an I event and a D event, respectively. If the default values of 0.5 for `s()` are used, then the required values of `t()` are the median survival times for each type of outcome. The survival distribution for both types of events is assumed to be exponential. `t()` is required.

`s(# [ # ])` defines the survival probability for an I event and a D event, respectively. For example, `s(0.5 0.75)` would say that the survival probability in the relevant interval was 0.5 for I outcomes and 0.75 for D outcomes. The default is `s(0.5 0.5)`.

`aratio(#)` specifies the allocation ratio, A (number of patients allocated to each experimental arm per control arm patient). For example, `aratio(0.5)` specifies that one patient is allocated to each experimental arm for every two patients allocated to control. The default is `aratio(1)` (equal allocation to all arms).

`tunit(#)` defines the code for units of trial time. The codes are `1` = one year, `2` = 6 months, `3` = one quarter (3 months), `4` = one month, `5` = one week, `6` = one day, and `7` = unspecified. `tunit()` has no influence on the computations and is for information only. The default is `tunit(1)` (one year).

`tstop(#)` defines the maximum time at which recruitment is to cease. To be valid and to make sense in the context of the MAMS design, # must be a time that falls within the final stage. If it does not, an error will be reported. The default is `tstop(0)`, meaning no ceasing of recruitment before the end of the final stage.

`probs` reports the probabilities of the numbers of arms passing each stage of the trial under the global null hypothesis (see section 3.5).

`nofwer` suppresses the calculation of the maximum FWER of the trial (probability of making at least one type I error at the end of the trial under any parameter configuration; see section 3.3).

`simcorr(#)` defines the number of replicates in the simulations to estimate the between-stage correlation structure. The estimated correlation structure is used to compute the overall type I error rate and power of the design (see section 3.2). At least 1,000 replicates are recommended. If `simcorr()` is not specified, the program uses the default correlation structure described by Royston et al. (2011). This option does not need to be specified if the I and D outcomes are identical.

`corr(#)` specifies either a) the correlation between HRs on the I and D outcomes at a fixed time point, such as the end of the trial; or b) if `simcorr()` is specified, the correlation between survival times on the I (excluding D) and D outcomes (see section 3.2). If it is the former, the value of # can be estimated by a bootstrap analysis of relevant previous trial data. In both cases, the default is `corr(0.6)` based on I = time to progression or death and D = time to death in cancer. Such a value is not necessarily appropriate in other settings. In the absence of knowledge, we suggest a sensitivity analysis for # in the range $[0.4, 0.8]$. Note that this option affects only the overall type I error rate and power of the design. This option does not need to be specified if the I and D outcomes are identical.

3 Updates to nstage

3.1 Targeting HRs above one under the alternative hypothesis

The decision to continue a treatment arm to the next stage of a MAMS trial is currently based on whether the observed HR for that arm is less favorable than some predetermined "critical" value. These thresholds are calculated using the one-sided significance levels and powers specified for each stage of the trial and the target HRs under the null and alternative hypotheses for each outcome.

In `nstage`, HRs under the null (H_0) and alternative (H_1) hypotheses are specified in options `hr0()` and `hr1()`, respectively, for an experimental arm relative to the control. For outcomes such as overall survival and failure-free survival, which are commonly used in cancer trials, an HR less than one signifies an improvement in the experimental arm. Because the program was initially developed for such settings, only target HRs less than one can currently be entered into `hr1()`. However, other outcomes exist for which an HR greater than one indicates an improvement implying that, under the proportional hazards assumption, events are observed more quickly in the experimental arm than in the control. An example is in tuberculosis, where time-to-culture negativity (a marker for cure) is often an outcome of interest in phase II trials (Phillips, Fielding, and Nunn 2013).

If an HR Δ^1 greater than 1 is to be targeted under H_1, then simply inputting its reciprocal into `nstage` is insufficient, despite the magnitude of the targeted treatment effect remaining the same. This is because the variance of $\log \Delta^1$, which is used in the sample-size calculation, is not equal to the variance of $\log 1/\Delta^1$. Under H_1, Royston et al. (2011) state that

$$V\left(\log \widehat{\Delta}^1\right) \approx \frac{1}{e_0} + \frac{1}{e_1} \tag{1}$$

where e_0 and e_1 are the number of events observed in the control and experimental arms, respectively, for a given pairwise comparison. Because e_1 will be larger under $\Delta^1(>1)$ than under $1/\Delta^1(<1)$, the variances will be unequal for a fixed value of e_0.

To allow a target HR Δ^1 greater than 1 to be selected, we must modify the methodology described in sections 2.3 and 2.4 of Royston et al. (2011). For a J-stage trial with one experimental arm, E, and a control, C, denote the one-sided significance level and power for the comparison at the jth interim analysis ($j = 1, \ldots, J$) by α_j and ω_j; and let Δ_j^0 and Δ_j^1 be the targeted treatment effects for the outcome in the jth stage under H_0 and H_1, respectively. Furthermore, let $\widehat{\Delta}_j$ denote the observed HR on the outcome of interest at the jth interim analysis of the trial, with the HR threshold for continuation being δ_j. By definition, for Δ_j^1 greater than 1,

$$\alpha_j = P\left(\log \widehat{\Delta}_j > \log \delta_j \mid H_0\right)$$

and

$$\omega_j = P\left(\log \widehat{\Delta}_j > \log \delta_j \mid H_1\right)$$

Following a similar calculation to that in section 2.3 of Royston et al. (2011) gives the critical value for $\log \widehat{\Delta}_j$ as

$$\log \delta_j = \log \Delta_j^0 - z_{\alpha_j}\sigma_j^0 = \log \Delta_j^1 - z_{\omega_j}\sigma_j^1$$

where σ_j^0 and σ_j^1 are the standard deviations for the log HR under H_0 and H_1, respectively.

Assuming $\sigma_j^1 = \sigma_j^0$, an initial estimate of the required number of control events, e_{j0}, for the jth interim analysis is given by (see also (4) in Royston et al. [2011])

$$e_{j0} = (1 + A^{-1}) \left(\frac{z_{\alpha_j} - z_{\omega_j}}{\log \Delta_j^0 - \log \Delta_j^1} \right)^2 \tag{2}$$

However, because there will be more events occurring in the trial under H_1 than under H_0, (1) implies that $\sigma_j^1 < \sigma_j^0$. Consequently, the estimate of e_{j0} obtained from (2) that assumes $\sigma_j^1 = \sigma_j^0$ will result in the actual power overshooting the nominal value ω_j. The algorithm described in section 2.4 of Royston et al. (2011) should then be used to calculate the number of events required to achieve the desired level of power, with step 6 modified so that e_{j0} is decreased by 1 at each iteration until ω_j is reached.

3.2 Estimating the between-stage correlation

To determine the pairwise operating characteristics for a MAMS trial, that is, the overall type I error rate and power for a particular research arm, we must obtain an estimate of the correlation between treatment effects at different stages. Royston et al. (2011) derived a formula for the correlation between HRs estimated for a particular outcome at different stages of the trial: if e_{j0} control events have been observed on the outcome of interest by the end of stage $j = 1, \ldots, J$, then the correlation between the HRs estimated in stages j and $k(> j)$ is $R_{jk} = \sqrt{e_{j0}/e_{k0}}$. This formula is not applicable when the outcomes in stages j and k differ, as is the case in the intermediate and final stages in a trial where $I \neq D$. In such a trial, Royston et al. (2011) proposed that the correlation between the observed HRs at each interim stage ($j = 1, \ldots, J-1$) on I and the final stage ($k = J$) on D can be approximated by

$$R_{jJ} \approx c\sqrt{\frac{e_{j0}}{e_{J0}}} \tag{3}$$

where c is a constant approximately equal to the correlation between the log HRs for I and D at a fixed time point, for example, the end of a trial. We can obtain an estimate for c from existing trial data or, if suitable data are unavailable, through expert opinion. In either case, we should perform a sensitivity analysis using various values of c to determine plausible ranges for the pairwise type I error rate and power.

Because (3) is only an approximation likely to be based on guesswork, the subroutine `hrcorrnstage`, which simulates individual patient data, has been created for `nstage` to provide a more reliable estimate of the between-stage correlation when I is a composite

of D and some other outcome I^* that is on the causal pathway to D. For example, in the STAMPEDE trial, the D outcome measure is death from any cause, whereas the I outcome measure is failure-free survival, defined as disease progression (I^*) or death from any cause (Sydes et al. 2009).

Given an estimate of the correlation, ρ, between the survival times on I^* and D as specified in the `corr()` option, `hrcorrnstage` simulates individual patient data under the proposed trial design for one experimental arm and a control. After the number of trials specified in `simcorr()` have been simulated, the correlations between the HRs observed at each intermediate stage and the final stage are estimated. Our empirical investigations showed that using 1,000 replicates provides reliable estimates in practice. The correlation is calculated under H_0 and H_1 for more accurate estimation of the type I error rate and power, respectively. Because ρ is defined as the correlation between I^* and D that is likely to be unobtainable from existing data because of censoring, a sensitivity analysis should be used to determine a plausible range for the true pairwise type I error rate and power, as is the case when specifying estimates of c in (3). Exploring values of ρ between 0.4 and 0.8 is recommended.

3.3 FWER

`nstage` currently outputs the probability of a single experimental arm passing all J stages of the trial under H_0, known as the pairwise type I error rate. However, in a multiarm trial, an estimate of the probability of one or more ineffective arms passing all J stages, known as the FWER, is also useful, particularly if it is to be controlled at some prespecified level. For example, in confirmatory trials, the European Medicines Agency states that it is a "minimal prerequisite" to control the FWER by limiting the maximum probability of making at least one type I error (Committee for Proprietary Medicinal Products 2002).

Calculation of FWER

Wason and Jaki (2012) describe a relatively quick calculation of the FWER for a MAMS design with normally distributed outcomes by simulating the joint distribution of the z test statistics for each arm at each stage, ignoring stopping guidelines. In the group of MAMS designs they consider, the same outcome is used throughout the trial, and the same number of patients is allocated to the control arm in each stage. In the class of MAMS designs considered here, neither of these constraints is likely to be met, particularly the latter. Below we generalize their simulation of the joint distribution to designs where unequal numbers can be allocated to the control arm in each stage and where interim analyses can be conducted on an intermediate outcome that differs from the definitive outcome of the trial.

Consider a $(K+1)$-arm J-stage trial, and denote by Z_{jk} the z test statistic for the kth experimental arm $(k = 1, \ldots, K)$ on the outcome of interest at stage j $(j = 1, \ldots, J)$. For time-to-event outcomes, Z_{jk} may be the z test statistic for the log HR. Ignoring stopping guidelines, we see that the distribution of Z_{jk} is

$$Z_{jk} \sim N\left(\frac{\Delta_{jk} - \Delta_j^0}{\sigma_{jk}}, 1\right)$$

where Δ_{jk} is the underlying effect (for example, true log HR) on the outcome of choice in stage j for arm k; Δ_j^0 is the corresponding effect under H_0; and σ_{jk} is the standard deviation of the observed treatment effects under Δ_{jk}.

Denote the correlation between the test statistics in stages j and j' for experimental arm k by $R_{jj'} = \text{corr}(Z_{jk}, Z_{j'k})$, as calculated using the methods discussed in section 3.2. The correlation between the treatment effects for any two arms k and k' $(k \neq k')$ in the same stage is $\text{corr}(Z_{jk}, Z_{jk'}) = A/(A+1)$ (Dunnett 1955), where A is the allocation ratio specified in `aratio()`.

To simulate the joint distribution of Z_{jk} with the above correlation structure, we first generate standard normally distributed random variables x_{jk} $(j = 1, \ldots, J,\ k = 0, \ldots, K)$ such that the correlation between x_{jk} and $x_{j'k}$ is $R_{jj'}$ for each k. This can be achieved using the `drawnorm` command in Stata. The formula

$$Z_{jk} = \sqrt{\frac{A}{A+1}}x_{j0} + \sqrt{\frac{1}{A+1}}x_{jk} + \frac{\Delta_{jk} - \Delta_j^0}{\sigma_{jk}} \tag{4}$$

then gives JK simulated random variables with the required distribution and between-stage and between-arm correlation structures (proof in the *Appendix*).

Using this technique, we can calculate the FWER under a range of parameter configurations $(\Delta_{jk} : j = 1, \ldots, J,\ k = 1, \ldots K)$. Clearly, the configuration under which the FWER is maximized is of greatest interest, particularly if strong control of the FWER is required. A result by Magirr, Jaki, and Whitehead (2012) states that for MAMS trials with one outcome $(I = D)$, the FWER is maximized under the global null hypothesis, H_G, that is, when $\Delta_{jk} = \Delta_j^0$ for all j and k.

The FWER under H_G is estimated as the proportion of all replicates in which at least one experimental arm passes all J stages of the trial (that is, if, for some k, $Z_{jk} < z_{\alpha_j}$ for all $j = 1, \ldots, J$). Using 250,000 replicates provides a good estimate of the FWER in practice because it ensures the Monte Carlo standard error is no greater than 0.001.

This method for calculating the FWER differs from that used by Wason and Jaki (2012). In their article, the authors first generate the x_{jk} with the appropriate between-arm correlation (rather than between-stage) and then proceed to use an expression similar to (4) to generate the test statistics Z_{jk}, which also have the required between-stage correlation. This seems tractable only if the intermediate and definitive outcomes are identical; otherwise, the between-stage correlation structure becomes too complex for it to be induced using a formula such as (4).

When $I \neq D$, the FWER will be maximized when H_0 is true for all arms on D. However, this maximum value will be achieved only when each arm is infinitely effective on the intermediate outcome—that is, when $\Delta_{jk} = -\infty$ for all $j < J$ and all k (assuming, without loss of generality, that a negative effect is beneficial) and $\Delta_{Jk} = \Delta_J^0$ for all k. Under such a configuration, each experimental arm will always pass all interim analyses, thus making them redundant. Therefore, the pairwise type I error rate for each arm will be equal to the final-stage significance level (because H_0 is true on D for all arms). As the magnitude of the effect on I decreases, the pairwise type I error rate will also decrease because some arms will inevitably be dropped at interim analyses. Consequently, fewer arms will reach the final stage; hence, fewer type I errors will be made. This makes the choice of I particularly important because it should have high specificity for the primary outcome of the trial to limit the inflation of the FWER.

Note that when $I \neq D$, the maximum FWER is equivalent to the FWER for a one-stage design with pairwise type I error rate equal to the final-stage significance level, α_J. In this case, the maximum FWER can be calculated using the Dunnett probability

$$\Phi_K(z_{\alpha_J}, \ldots, z_{\alpha_J}; \mathbf{C})$$

which is computationally simpler and faster than the simulation described above. Here Φ_K is the K-dimensional multivariate standard normal distribution function, and $\mathbf{C}$ is the $K \times K$ between-arm correlation matrix with off-diagonal entries equal to $A/(A+1)$ (Dunnett 1955). When strong control of the FWER is required, this probability should be limited by choosing an appropriate value of α_J.

Subroutine for calculating FWER

To calculate the FWER in `nstage`, we added a subroutine that implements the methods described above. FWER is now calculated by default in `nstage` for all designs. However, we also added an option `nofwer` to circumvent this. The subroutine is applicable to any normally distributed test statistics and can be incorporated into future `nstage`-type programs for other types of outcome (for example, binary or continuous).

3.4 nstage output

Below is an example output given by `nstage` using one of the design scenarios for the original comparisons in the STAMPEDE trial, described in detail by Sydes et al. (2009). The example shows an estimate of the FWER assuming H_0 is true for both I and D for all research arms, with the standard error of the estimate induced by the simulation. The maximum pairwise and FWERs are also provided along with sample sizes, event estimates, and durations for each stage.

For STAMPEDE, the accrual pattern has differed from this scenario and that the number of arms at each stage for the original comparisons has been `arms(6 6 4 4)`.

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

n-stage trial design                  version 3.0.1, 10 Sept 2014
─────────────────────────────────────────────────────────────────

Sample size for a 6-arm 4-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81
─────────────────────────────────────────────────────────────────

Median survival time (I-outcome): 2 time units
Median survival time (D-outcome): 4 time units
Operating characteristics
```

	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
Pairwise**	0.0118	0.833				6.027	
Familywise(SE)**	0.0517	(0.0004)					

Maximum Pairwise Alpha 0.0250 Maximum Familywise Alpha 0.1030

```
 *  Length (duration of each stage) is expressed in one year periods and
    assumes survival times are exponentially distributed
 ** Calculated under global null hypothesis for I and D outcomes
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
```

3.5 Correction to the probs option

The `probs` option in `nstage` reports the probabilities of k out of K experimental arms reaching each stage of the trial under the global null and global alternative hypotheses (Barthel, Royston, and Parmar 2009). The following is an example of the output given by the previous version of `nstage` when specifying the `probs` option for the STAMPEDE-based example described in the previous section:

```
Approx. prob. of k experimental arms reaching stage 2:
------------------------------------------------------------------------------
k (#arms)         0         1         2         3         4         5
------------------------------------------------------------------------------
Under H0      0.031     0.156     0.313     0.313     0.156     0.031
Under H1      0.000     0.000     0.001     0.021     0.204     0.774
------------------------------------------------------------------------------

Approx. prob. of k experimental arms reaching stage 3:
------------------------------------------------------------------------------
k (#arms)         0         1         2         3         4         5
------------------------------------------------------------------------------
Under H0      0.513     0.366     0.105     0.015     0.001     0.000
Under H1      0.000     0.000     0.007     0.069     0.322     0.601
------------------------------------------------------------------------------

Approx. prob. of k experimental arms reaching stage 4:
------------------------------------------------------------------------------
k (#arms)         0         1         2         3         4         5
------------------------------------------------------------------------------
Under H0      0.939     0.059     0.002     0.000     0.000     0.000
Under H1      0.000     0.002     0.021     0.127     0.384     0.466
------------------------------------------------------------------------------
```

The output informs users to adjust their design if the probabilities are not as desired, for instance, if too many ineffective arms are likely to reach the final stage of recruitment. These probabilities were previously calculated using binomial distributions as described by Barthel (2006), which do not account for the correlation between treatment arms and therefore lead to incorrect estimates.

By simulating the joint distribution of Z_{jk} as described in section 3.3, we can obtain more accurate estimates of these probabilities by calculating the proportion of all replicates in which k out of K experimental arms pass each stage. The `probs` option has been amended to use this method. Additionally, the proportion of k experimental arms passing the final stage of the trial is now calculated to give the distribution of type I errors at the end of the trial under H_G. The new output, displayed below for the STAMPEDE-based example, shows the probability of the number of arms passing each stage (rather than reaching each stage). The discrepancy between the estimates from the two calculations is quite large. For example, the binomial method estimates the probability of no ineffective arms reaching the final stage (or equivalently, passing stage 3) to be 0.94, whereas the newly estimated but more accurate value is closer to 0.72.

Probability of k experimental arms passing each stage under global null hypothesis

k(#arms)	0	1	2	3	4	5
Stage 1	0.114	0.178	0.208	0.208	0.178	0.114
Stage 2	0.411	0.279	0.167	0.089	0.041	0.013
Stage 3	0.719	0.194	0.061	0.020	0.005	0.001
Stage 4	0.948	0.046	0.005	0.001	0.000	0.000

Calculation of these probabilities under the global alternative hypothesis is not included because here we want each treatment arm to pass all stages of the trial. Hence, the only probability of real interest is the pairwise power. Therefore, the additional time required to estimate the probabilities under the global alternative is not warranted.

3.6 Update to the dialog menu

Users can access the menu for `nstage` through the **User** menu after entering `nstagemenu on` in the command line. The original dialog box for `nstage` (described in detail by Barthel, Royston, and Parmar [2009]) has been redesigned to improve its usability and to incorporate options for the extensions described in this article. In the first tab, **Design parameters** (see figure 1), users enter general design parameters that are applicable to the trial as a whole, such as number of stages, allocation ratio, units of time, time for stopping accrual, and options for calculating the FWER and the probabilities for the number of ineffective arms passing each stage. In the second tab, **Operating characteristics** (see figure 2), users select the significance levels, powers, accrual rates (per unit of trial time), and number of recruiting arms for each stage of the trial.

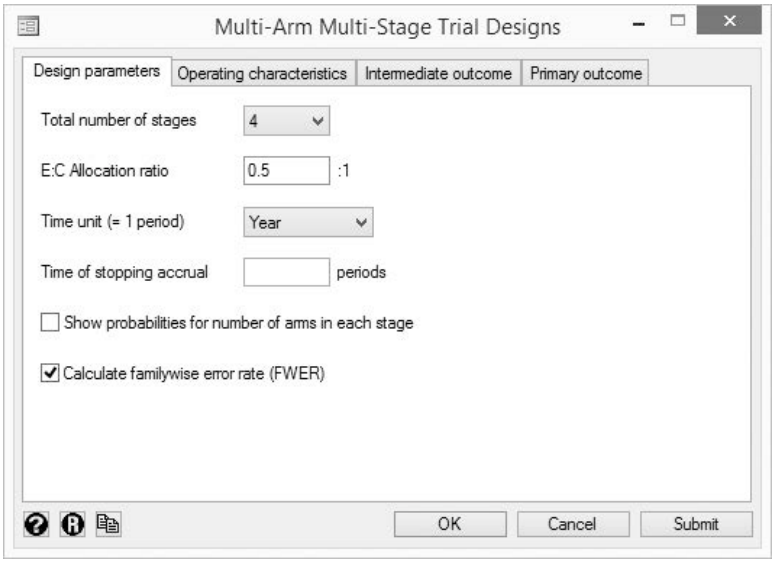

Figure 1. Screenshot of the first tab of the `nstage` dialog box: general trial design parameters

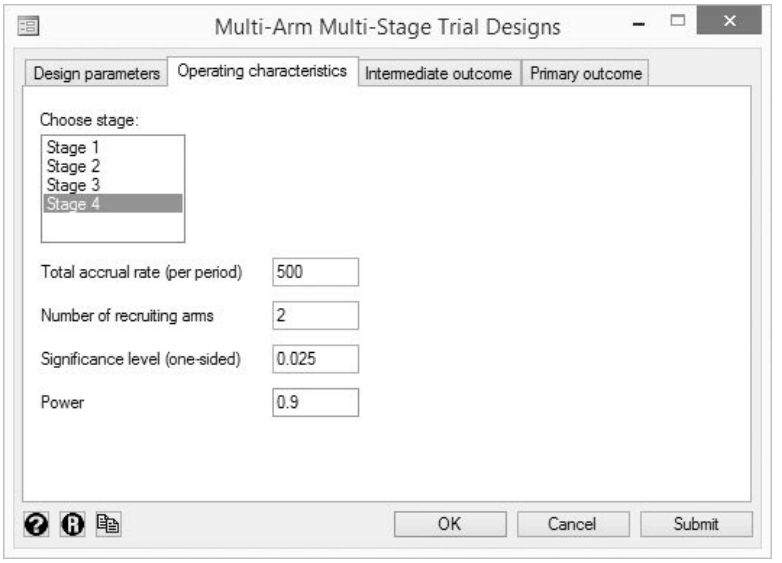

Figure 2. Screenshot of second tab of the `nstage` dialog box: stagewise operating characteristics

If interim assessments are to be performed on an intermediate outcome that is different from the primary outcome, then users enter design parameters for this outcome in the third tab, **Intermediate outcome** (see figure 3). These include the survival time (in the chosen units of time) for the corresponding survival probability at that time (usually 0.5 to represent the median survival time) and the target HRs under H_0 and H_1—the latter can now be either less than or greater than 1. The corresponding parameters for the primary outcome are entered in the final tab, **Primary outcome** (see figure 4).

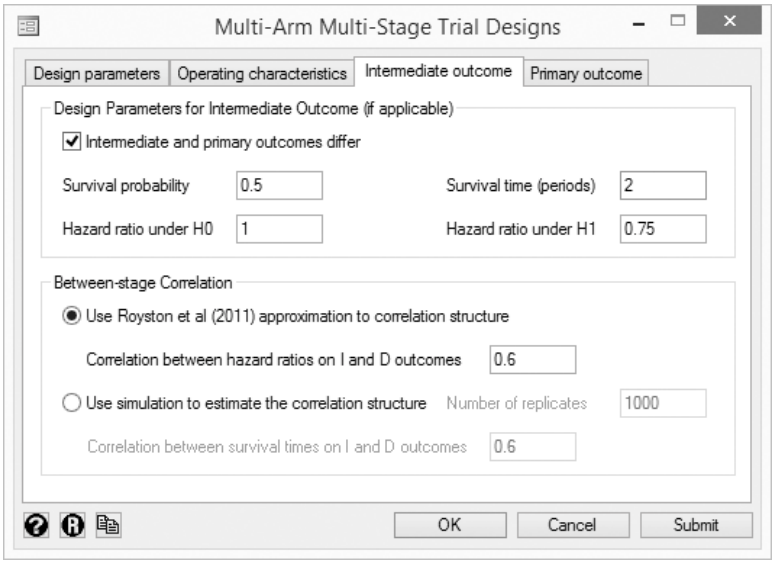

Figure 3. Screenshot of third tab of the `nstage` dialog box: parameters for the intermediate outcome, if applicable, and method for calculating the between-stage correlation

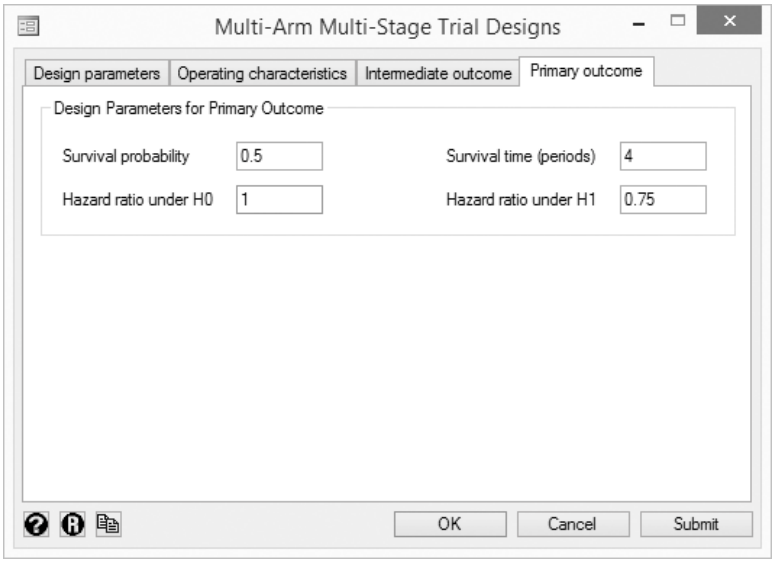

Figure 4. Screenshot of fourth tab of the `nstage` dialog box: parameters for the primary outcome

The method for calculating the between-stage correlation when the intermediate and definitive outcomes differ is also entered on the **Intermediate outcome** tab. Either the algebraic approximation described by Royston et al. (2011) can be chosen, in which case an estimate of c must be specified, or the more accurate yet computationally intensive method of using simulations can be chosen, which requires an estimate of ρ to be entered (see section 3.2).

4 Sensitivity of operating characteristics to c and ρ

Below we investigate the sensitivity of the pairwise and FWERs, under H_0 for I and D, and the pairwise power of the STAMPEDE-based example to the calculation of the between-stage correlation structure described in section 3.2. Pairwise and familywise operating characteristics were calculated using the between-stage correlation structure estimated in `hrcorrnstage` for values of ρ ranging from 0.4 to 0.8 in increments of 0.1. In the simulations, 5,000 replicates were used. For comparison, similar calculations were made by estimating the correlation structure using (3) with c also ranging from 0.4 to 0.8 in increments of 0.1.

The results in table 1 show that α, ω, and the FWER are more sensitive to c than ρ. Using `hrcorrnstage` to estimate the between-stage correlation structure suggests that the FWER under H_0 for I and D is likely to range between 0.043 and 0.056, whereas it is estimated to range between 0.034 and 0.076 when using (3). We recommend using `hrcorrnstage` where possible because it not only more accurately estimates the

between-stage correlation under a particular assumption but also gives more precise estimates of the pairwise and familywise operating characteristics.

Table 1. Estimated pairwise type I error rate (α), power (ω), and FWER for the STAMPEDE-based example shown in section 3.4 using two methods for calculating the between-stage correlation. Values for c and ρ between 0.4 and 0.8 are investigated. Note: The pairwise and familywise type I error rates are calculated assuming H_0 is true for I and D for all experimental arms. Their corresponding maximum values are 0.025 and 0.103, respectively.

Estimating correlation using (3)				Estimating correlation using `hrcorrnstage`			
c	α	ω	FWER	ρ	α	ω	FWER
0.4	0.007	0.823	0.034	0.4	0.010	0.828	0.043
0.5	0.009	0.828	0.043	0.5	0.010	0.828	0.044
0.6	0.012	0.833	0.053	0.6	0.011	0.831	0.048
0.7	0.015	0.839	0.063	0.7	0.012	0.833	0.051
0.8	0.018	0.846	0.076	0.8	0.013	0.835	0.056

5 Conclusion

In this article, we have addressed the need for a quick and accurate calculation of the FWER for the class of MAMS designs described by Royston et al. (2011). Users of `nstage` can now find MAMS designs that control the FWER in the weak or strong sense at any desired level. However, further work is needed to determine how the stagewise design parameters should be specified to maximize the efficiency of such trials.

We have extended the methodology to allow HRs greater than one to be targeted under H_1, thus broadening the areas in which the MAMS design could be used. Furthermore, we have provided a simulation procedure to more accurately calculate the between-stage correlation when the intermediate and definitive outcomes differ. As shown in table 1, this is important if more accurate estimates of the overall operating characteristics of the trial are to be obtained.

6 Acknowledgments

We thank Matthew Sydes from the STAMPEDE trial group and our colleagues at the MRC Clinical Trials Unit for their helpful comments on earlier versions of this article and the `nstage` program.

7 References

Barthel, F. M.-S. 2006. *Issues in the design and analysis of clinical trials with time-to-event outcomes*. PhD thesis, University College London.

Barthel, F. M.-S., P. Royston, and M. K. B. Parmar. 2009. A menu-driven facility for sample-size calculation in novel multiarm, multistage randomized controlled trials with a time-to-event outcome. *Stata Journal* 9: 505–523.

Bratton, D. J., P. P. J. Phillips, and M. K. B. Parmar. 2013. A multi-arm multi-stage clinical trial design for binary outcomes with application to tuberculosis. *BMC Medical Research Methodology* 13: 139.

Choodari-Oskooei, B., M. K. B. Parmar, P. Royston, and J. Bowden. 2013. Impact of lack-of-benefit stopping rules on treatment effect estimates of two-arm multi-stage (TAMS) trials with time to event outcome. *Trials* 14: 23.

Committee for Proprietary Medicinal Products. 2002. Points to consider on multiplicity issues in clinical trials. Technical report, EMEA.

Dunnett, C. W. 1955. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* 50: 1096–1121.

Magirr, D., T. Jaki, and J. Whitehead. 2012. A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika* 99: 494–501.

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.

Phillips, P. P. J., K. Fielding, and A. J. Nunn. 2013. An evaluation of culture results during treatment for tuberculosis as surrogate endpoints for treatment failure and relapse. *PLOS ONE* 8: e63840.

Raja, F. A., C. L. Griffin, W. Qian, H. Hirte, M. K. Parmar, A. M. Swart, and J. A. Ledermann. 2011. Initial toxicity assessment of ICON6: A randomised trial of cediranib plus chemotherapy in platinum-sensitive relapsed ovarian cancer. *British Journal of Cancer* 105: 884–889.

Royston, P., F. M.-S. Barthel, M. K. B. Parmar, B. Choodari-Oskooei, and V. Isham. 2011. Designs for clinical trials with time-to-event outcomes based on stopping guidelines for lack of benefit. *Trials* 12: 81.

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.

Smith, J. M., C. J. Doré, A. Charlett, and J. D. Lewis. 1992. A randomized trial of biofilm dressing for venous leg ulcers. *Phlebology* 7: 108–113.

Sydes, M. R., M. K. B. Parmar, N. D. James, N. W. Clarke, D. P. Dearnaley, M. D. Mason, R. C. Morgan, K. Sanders, and P. Royston. 2009. Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: The MRC STAMPEDE trial. *Trials* 10: 39.

Sydes, M. R., M. K. B. Parmar, M. D. Mason, N. W. Clarke, C. Amos, J. Anderson, J. de Bono, D. P. Dearnaley, J. Dwyer, C. Green, G. Jovic, A. W. S. Ritchie, J. M. Russell, K. Sanders, G. Thalmann, and N. D. James. 2012. Flexible trial design in practice—stopping arms for lack-of-benefit and adding research arms mid-trial in STAMPEDE: A multi-arm multi-stage randomized controlled trial. *Trials* 13: 168.

Wason, J., D. Magirr, M. Law, and T. Jaki. 2013. Some recommendations for multi-arm multi-stage trials. Statistical Methods in Medical Research.

Wason, J. M. S., and T. Jaki. 2012. Optimal design of multi-arm multi-stage trials. *Statistics in Medicine* 31: 4269–4279.

About the authors

Daniel J. Bratton is a medical statistician at the University Hospital, Zurich, and recently completed a PhD investigating design issues in multiarm, multistage clinical trials at the MRC Clinical Trials Unit at University College London.

Babak Choodari-Oskooei is a medical statistician in the Hub for Trials Methodology Research at the MRC Clinical Trials Unit at University College London with a particular interest in clinical trials methodology, model validation, and the applications of predictive ability measures in different settings.

Patrick Royston is a medical statistician with over 30 years of experience and a strong interest in biostatistical methods and in statistical computing and algorithms. He works largely with methodological issues in the design and analysis of clinical trials and observational studies. He is currently focusing on alternative outcome measures in trials with a time-to-event outcome; on problems of model building and validation with survival data, including prognostic factor studies and treatment-covariate interactions; on parametric modeling of survival data; and on novel clinical trial designs.

Appendix

Below is a proof that the test statistics, Z_{jk}, generated using (4) have the required distribution

$$Z_{jk} \sim N\left(\frac{\Delta_{jk} - \Delta_j^0}{\sigma_{jk}}, 1\right)$$

between-stage correlation

$$\text{corr}(Z_{jk}, Z_{j'k}) = R_{jj'}$$

and between-arm correlation

$$\text{corr}(Z_{jk}, Z_{jk'}) = \frac{A}{A+1} \text{ if } k \neq k'$$

1. Expectation

$$\begin{aligned} E(Z_{jk}) &= \sqrt{\frac{A}{A+1}}E(x_{j0}) + \sqrt{\frac{1}{A+1}}E(x_{jk}) + \frac{\Delta_{jk} - \Delta_j^0}{\sigma_{jk}} \\ &= \frac{\Delta_{jk} - \Delta_j^0}{\sigma_{jk}} \end{aligned}$$

because $E(x_{jk}) = 0$ for all $j = 1, \ldots, J$ and $k = 0, \ldots, K$.

2. Variance

$$\begin{aligned} V(Z_{jk}) &= \frac{A}{A+1}V(x_{j0}) + \frac{1}{A+1}V(x_{jk}) \\ &= \frac{A}{A+1} + \frac{1}{A+1} \\ &= 1 \end{aligned}$$

because $V(x_{jk}) = 1$ for all $j = 1, \ldots, J$ and $k = 0, \ldots, K$, and $\text{cov}(x_{jk}, x_{jk}) = 0$ for $k \neq k$.

3. Between-stage correlation

$$\begin{aligned} \text{corr}(Z_{jk}, Z_{j'k}) &= \text{cov}(Z_{jk}, Z_{j'k}) \\ &= \frac{A}{A+1}\text{cov}(x_{j0}, x_{j'0}) + \frac{1}{A+1}\text{cov}(x_{jk}, x_{j'k}) \\ &= \frac{A}{A+1}R_{jj'} + \frac{1}{A+1}R_{jj'} \\ &= R_{jj'} \end{aligned}$$

because $\text{cov}(x_{jk}, x_{jk'}) = 0$ for $k \neq k'$ and $\text{cov}(x_{jk}, x_{j'k}) = R_{jj'}$ for all $k = 0, \ldots, K$.

4. Between-arm correlation

$$\begin{aligned} \text{corr}(Z_{jk}, Z_{jk'}) &= \frac{A}{A+1}\text{cov}(x_{j0}, x_{j0}) \\ &= \frac{A}{A+1} \end{aligned}$$