Published in final edited form as:

Clin Trials. 2017 June ; 14(3): 237–245. doi:10.1177/1740774517694130.

Controlling the family-wise error rate in multi-arm, multi-stage trials

Luis A. Crouch¹, Lori E. Dodd², and Michael A. Proschan²

¹University of Washington, Seattle, WA, USA

²National Institute of Allergy and Infectious Diseases, Bethesda, MD, USA

Abstract

Background and Aims—Multi-arm, multi-stage trials have recently gained attention as a means to improve the efficiency of the clinical trials process. Many designs have been proposed, but few explicitly consider the inherent issue of multiplicity and the associated type I error rate inflation. It is our aim to propose a straightforward design that controls family-wise error rate while still providing improved efficiency.

Methods—In this paper we provide an analytical method for calculating the family-wise error rate for a multi-arm, multi-stage trial and highlight the potential for considerable error rate inflation in uncontrolled designs. We propose a simple method to control the error rate that also allows for computation of power and expected sample size.

Results—Family-wise error rate can be controlled in a variety of multi-arm, multi-stage trial designs using our method. Additionally, our design can substantially decrease the expected sample size of a study while maintaining adequate power.

Conclusions—Multi-arm, multi-stage designs have the potential to reduce the time and other resources spent on clinical trials. Our relatively simple design allows this to be achieved while weakly controlling family-wise error rate and without sacrificing much power.

Keywords

multi-arm; multi-stage; clinical trials; multiple testing; adaptive design

Introduction

Randomized clinical trials comparing a single experimental arm to a control arm have long been the gold standard for demonstrating treatment safety and efficacy. This high standard requires lengthy studies along with substantial financial and scientific resources. For this reason, there has been considerable interest in modifying the standard design to be more

Reprints and permission: sagepub.co.uk/journalsPermissions.nav

Corresponding author: Luis Crouch, University of Washington Department of Biostatistics, Seattle, Washington 98105, USA. crouch@uw.edu.

Conflict of interest
None declared.

efficient. One such modification is the multi-arm, multi-stage trial. Multiarm designates the inclusion of multiple treatment arms all compared to the same control arm. Multi-stage means that the trial is conducted in multiple stages. At the end of each stage, an analysis is carried out to determine which, if any, of the treatment arms are performing well enough (relative to control) to merit further study by continuing enrollment in the subsequent stage.

The multi-arm, multi-stage design is more efficient than the standard design in two ways. First, only one control arm is necessary to evaluate multiple novel treatments, reducing sample size as well as administrative overhead for conducting multiple two-arm studies. Second, dropping poorly performing arms early in the study can reduce the expected sample size.

The Systemic Therapy in Advancing or Metastatic Prostate cancer: Evaluation of Drug Efficacy (STAMPEDE) trial was perhaps the first large trial to employ a multi-arm, multi-stage design. Subjects were randomized to one of six treatment arms, with interim looks scheduled to drop poorly performing arms.¹ More recently, the multi-arm, multi-stage design has been recommended in the setting of tuberculosis treatment trials;² tuberculosis treatments require multi-drug regimens, so any new drug must be evaluated in combination with others, resulting in numerous combinations. While each combination could be evaluated in a separate trial, it seems more sensible to consider different combinations in a single study with a common control regimen.

Explicit recommendations about adjustment for the multiple comparisons inherent in the multi-arm, multi-stage design are lacking. Many papers consider a single treatment arm over multiple stages, claiming generalizability to multiple treatment arms. Others consider multiple treatment arms but a single stage. Some have argued that the multi-arm, multi-stage design is best suited for Phase II trials where concerns about type I error rate inflation or bias in effect size estimation are less of a concern than in Phase III. Phase II trials often do not control the type I error rate at the same level as a Phase III trial, but adequate description of the error rate remains important and cannot be dismissed. In a regulatory setting, control of the type I error rate at a level similar to a phase III trial may be relevant. For instance, companies may wish to support Phase III findings with results from Phase II trials. For example, Novartis combined results from the RELAX-AHF trial with those from the (Phase II) Pre-RELAX AHF trial when presenting on REASANZ™ (Serelaxin) to the Food and Drug Administration's Cardiovascular and Renal Drugs Advisory Committee.³ In such a situation, the results from Phase II cannot be considered without addressing concerns of multiplicity and its attendant family-wise error rate (FWER) inflation and bias.

Evaluations of the operating characteristics of multi-arm, multi-stage designs have been almost exclusively based on simulation studies of specific settings, such as Royston et al.,⁴ who explored timing of interim looks, power and significance levels. This work was later adapted to studies with binary outcomes by Bratton et al.⁵ Wason and Jaki⁶ focused on methods for determining optimal multi-arm, multi-stage designs with regard to stopping boundaries and sample size via simulation. Proschan and Dodd⁷ derived FWERs for different thresholds for dropping arms in a simplified setting with multiple arms without interim evaluations. It is worth noting that this work focused on a multi-arm, single-stage

setting with intention of future work to generalize to a multi-arm, multi-stage setting. In particular, Proschan and Dodd⁷ used the idea of dropping poorly performing arms as a way to reduce the threshold for significance in a single stage, not to (directly) reduce expected sample size. Magirr et al.⁸ adopted a very different approach. They used a clever generalization of the Dunnett testing procedure in the multi-arm, multi-stage setting and provided methods for both calculating and controlling the FWER. We use their approach and simplify the numerical integration to make it amenable to existing software for monitoring clinical trials while building on Proschan and Dodd's⁷ basic methodology.

In this paper, we characterize the properties of a particular multi-arm, multi-stage design with respect to FWER inflation and bias and make corresponding recommendations about trial design and analysis. We begin by extending the method from Proschan and Dodd⁷ to include multiple stages, demonstrating the potential for FWER inflation. Next, we derive a method to calculate the FWER, which leads to adjusted thresholds for significance. We then demonstrate the expected sample size, power, and bias of the multi-arm, multi-stage design with control of the FWER via simulation, with further demonstration via example. We end with a discussion including directions for future work and possible limitations.

Extension of previous work

We extend the method described by Proschan and Dodd,⁷ which considered m treatment arms ($m + 1$ arms including the control arm) but restricted to a single stage ($J = 1$). Each treatment arm is compared to the control arm based on the relevant summary statistic (such as the sample mean). Any treatment arm failing to meet a pre-specified threshold of performance, based on the estimated effect comparing treatment to control, as measured by the z -score, is dropped. Without multiple stages, Proschan and Dodd⁷ found that, when the criteria for dropping treatment arms was a z -score < 0 , the FWER was nearly controlled without the need to adjust α beyond using α/k , where α is the pre-specified nominal type I error rate and k is the number of arms meeting the performance threshold at the final stage. The straightforward extension considers multiple stages ($J > 1$), in which each treatment arm is compared to the control arm at the end of each stage. Poorly performing treatment arms are dropped, and the study continues to the subsequent stage with all remaining treatment arms. Hypothesis testing is not carried out until the end of the last stage of the study. At that point, we conduct tests with a significance level of α/k . Through simulation, we investigated whether this nearly controlled FWER when considering multiple stages ($J > 1$).

In our simulations, we generated control and treatment arm sample means for 1,000,000 hypothetical data sets from a standard normal distribution. Sample size was assumed to be large enough to treat standard deviation as known. We considered designs with $J = 1, 2$, or 3 stages, corresponding to equally spaced intervals of recruitment time, and $m = 2, 5$, or 10 treatment arms. Thresholds were held constant at each stage, corresponding to one of the following z -scores: $-0.5, 0$, or 0.5 . For instance, a Z -score threshold of 0 means that arms whose Z -scores with control are below 0 are dropped. A nominal α of 0.025 was used to determine significance at the end of the simulated trial.

The choice of sampling ratio is also relevant. Many multi-arm designs suggest a $\sqrt{M}:1$ control arm sample size to treatment arm sample size as the most efficient sampling ratio. In practice, this ratio is problematic for many reasons. First, ratios higher than 1:1:...1 may necessitate more adjustment to the nominal α , reducing or negating the efficiency gained, as mentioned in Proschan and Dodd.⁷ Second, comparisons among active arms may also be of interest, but would have reduced power if the total sample size is fixed and the control sample size is increased. Third, patients may prefer to have a reasonably high probability of being assigned to a new therapy, as discussed in Halpern et al.⁹ Finally, recent work from Wason et al.¹⁰ has shown that in multi-arm, multi-stage trials, the optimal ratio does not differ substantially from 1:1:...1 and the gains in efficiency comparing the optimal ratio to the equal assignment ratio are relatively small. For these reasons, we consider only 1:1:...1 randomization ratios.

Table 1 demonstrates that FWER inflation can be substantial for certain study designs. FWER tends to increase both as the threshold for continuation of an arm becomes more stringent and as the number of stages increases. For example, when $m = 10$, $J = 3$, and thresholds of $Z > 0$ are used, the simulated FWER of 0.0477 is nearly double the target of 0.025. While the number of treatment arms clearly has an effect on FWER, the trend is not as clear as it is for number of stages. Interestingly, for less stringent thresholds and larger number of arms, we observe FWER deflation (that is, an actual FWER lower than the nominal α). This is primarily due to the conservative nature of the Bonferroni correction for multiple testing on which this method is based, particularly when the test statistics are correlated (as they are here). The possibility of performing no hypothesis tests (if all arms are dropped) may contribute as well, though this is less likely as the number of arms increases.

Note that while the simulation results presented above assume a large enough sample to treat the standard deviation as known and to have asymptotically normally distributed sample means in each arm, additional simulations (data not shown) conducted using small samples ($n = 10$ observation per arm per stage) and a variety of heavy-tailed and/or heavily-skewed distributions had similar results.

Based on our simulation results, it is clear that, particularly for designs with multiple stages, FWER control must be considered more carefully and further adjustment of α is necessary.

Method

Inherent in the multi-arm, multi-stage design described in the previous section is the issue of multiplicity. At the end of the trial, k different tests are carried out. As in Proschan and Dodd,⁷ we see that while testing with a significance level of α/k reduces the probability of type I errors, it is not sufficient to control FWER. This is particularly true in a design with multiple stages. We must, therefore, determine a method for calculating an adjusted significance level α' . While simulation allows us to estimate the true FWER associated with a particular α , an analytic method is preferred for more precise calculation of the true FWER.

We adopt the Magirr et al.⁸ idea of first conditioning on all of the control data, as is done when determining Dunnett critical values. Once we condition on the control data, the Z -scores comparing the active arms with control are independent, greatly facilitating the calculation of the Type I error rate. The end result is that the FWER can be computed according to the following multiple integral, subject to certain assumptions (refer to Appendix for derivation),

$$FWER = \int \cdots \int_{\mathbb{R}^J} \sum_{k=0}^n \binom{m}{k} \left\{ p_Z^k - (p_A - p_B)^k \right\} \times (1 - p_A)^{m-k} \phi(x^{(1)}) \cdots \phi(x^{(J)}) dx^{(1)} \cdots dx^{(J)},$$

(1)

where p_A and p_B are defined by equations 2 and 3, respectively, in the Appendix. In words, p_A is the conditional probability, given the control means at the different stages, that a given arm lasts until the end of the trial, while p_B is the conditional probability, given the control means at the different stages, that a given arm lasts until the end of the trial and is significant at level α/k . Importantly, both p_A and p_B are probabilities evaluated using existing monitoring techniques.¹¹ These techniques address the more complex mathematics and allow us to express the FWER as a relatively simple summation and multiple integral.

As previously mentioned, to control FWER at a particular value (e.g., 0.025) we must replace the nominal α with an adjusted α' . This α' can be solved for iteratively using either equation (1) or simulation. Table 2 provides these α' s for target FWERs of 0.025 and 0.05 with $m = 2, \dots, 10$ treatment arms, $J = 1, 2$ or 3 stages, and a Z -score threshold of $Z > 0$. For all values of m , α' decreases as the number of stages increases. In other words, the increased FWER inflation from more stages must be countered by a more stringent criterion for determining significance. The trend is less clear for number of treatment arms. With 2 or more stages, the degree of stringency increases, up to some number of arms, but then decreases. With $J = 2$ stages, the most adjustment is necessary for $m = 3$ treatment arms. With $J = 3$ stages, the most adjustment is necessary for $m = 6$ treatment arms.

The FWER control we have demonstrated in the appendix is weak control: control of FWER under the global null. Using the same reasoning as Proschan and Dodd,⁷ it can be shown that we in fact have strong control when no treatment arm is harmful. While a similar proof is unavailable when allowing for harmful treatment arms, we have not observed lack of strong FWER control in any of our simulations, though they were not designed to investigate this in particular.

Expected Sample Size, Power, and Bias

Having calculated α' for a variety of design parameters, we conducted further simulations to evaluate power, expected sample size, and bias. These simulations demonstrate the advantages of the multi-arm, multi-stage design as well as the performance of our method. We generated normally distributed observations representing control and treatment arm

sample means for 1,000,000 hypothetical data sets. We considered designs with $m = 2, 5$, or 10 treatment arms, $J = 2$ or 3 stages, and a constant threshold of $Z > 0$. To determine significance at the end of the simulated trial we used α' corresponding to a FWER of 0.025, as provided in Table 2.

When calculating expected sample sizes we considered three settings: the “global null” setting with all treatment arm means equal to the control arm mean, a “single effective” setting in which only one treatment arm mean differed from the control arm mean, and an “all effective” in which all treatment arm means differed from the control arm mean (and were equal to each other). The effect size was specified so that a two-armed trial with the same FWER would have 95% power: $\delta = (z_{1-\alpha} + z_{0.95}) (2\sigma^2/n)^{-1/2}$.

Table 3 shows relative expected sample sizes (i.e. the ratio of the expected sample size of a study using our multi-arm, multi-stage design to the sample size of a study without dropping treatment arms) for several different combinations of parameters and data settings. The true structure of the data plays a big role: relative expected sample size decreases as the number of treatment arms for which the alternative hypothesis holds decreases. This is anticipated, since treatment arms that have true means higher than the mean of the control arm are unlikely to be dropped. Indeed, when all treatment arms are more effective than the control arm we see virtually no reduction in expected sample size as compared to a trial that does not allow for dropping arms. The number of stages also influences expected sample size, with additional stages being associated with lower expected sample sizes. This is also anticipated, as additional stages offer more opportunities to drop poorly performing treatment arms.

When evaluating power and bias we considered only the single effective and all effective settings. In a study with multiple arms, power can be defined in several ways. Here we define it as the probability that we reject the null hypothesis of equality between a *particular* treatment arm and the control arm, when the parameter for that particular treatment arm truly differs from the parameter for the control arm. This is in contrast to an alternative method for evaluating power that computes the probability of rejecting *at least one* of the treatment arms that truly differ from the control arm. We calculated power using the method described above, comparing it to a Dunnett method in which no arms were dropped. Both were applied using this “basic” method as well as a sequentially rejective method adapting the Holm sequential method.¹² The Hochberg¹³ sequential method, which is known not to control the FWER in all settings, was considered but not employed due to further inflation of the FWER when applied to our method. The Holm sequential method showed no FWER inflation (data not shown).

Table 4 shows simulated powers for several different combinations of parameters and data settings. Under the single effective setting, there is no appreciable difference between the basic and sequential methods. This is as expected; we do not expect to reject the other null hypotheses because they are truly null. In this setting our method outperforms the Dunnett method, especially as the number of arms and stages increase. There is a difference between the basic and sequential methods in the all effective setting. Compared to the single effective setting, the basic version of our method has reduced power in the all effective setting (this is

not true for the Dunnett method). Most of this lost power can be regained by using sequential rejection methods. Indeed, in the all effective setting, the Dunnett sequential method has better power than it did in the single effective setting. Notice that Dunnett has a slight power advantage over our method in the all effective setting. Nonetheless, it is likely a rare occurrence for all arms to be substantially superior to the control.

Some broader trends are also apparent from Table 4. Power tends to decrease slightly with additional stages. The decrease in power is likely due to incorrectly dropping treatment arms that may truly be superior to the control arm. For both methods, power decreases as the number of arms increases. This is as expected, since a larger m means a smaller α/m and likewise for k and α'/k .

As noted in Proschan and Dodd,⁷ bias must be carefully defined in a multi-arm trial because it depends not only on the parameter values in the arm considered, but on parameter values in other arms. Therefore, we evaluate bias conditional only on the arm of interest making it

to the end of the study. Table 5 shows simulated percent relative biases $\left(\frac{E(\hat{\theta}) - \theta}{\theta} \times 100\% \right)$ for several different combinations of parameters and data settings.

All percent relative biases shown in Table 5 are quite small - less than 1%. Bias was also minimal under the global null setting (data not shown). This is partly because there is no early stopping for efficacy. Still, bias clearly increases with increasing number of stages and, by extension given our set-up, earlier initial interim analyses. This may not be of particular concern in the settings we considered, but bias can be much larger when more stringent thresholds are used or when the study is underpowered. For example, using the same simulation settings as above but with a power of only 50%, relative bias for $J=3$ stages can be as high as 9.8%. This is because, in a low power setting, the signal to noise ratio in the data is lower, so the decision to keep or drop an arm is more dependent on chance and therefore more prone to introducing bias as we define it. While we would not intentionally design a study with such low power, it is possible to find ourselves in such a setting when the true effect size is lower than expected. Further exploration of bias in a similar setting and with similar results can be found in Choodari-Oskooei et al.¹⁴

As explained above, we have focused on bias for arms making it to the end of the study. It is worth noting that bias will also exist in arms that are dropped at an interim analysis. The estimates for these arms will be biased in the direction of poor performance relative to control. This is not particularly concerning given that the implication of dropping treatment arms is that we are no longer interested in them, but knowing that the estimates are biased downwards may be important if new studies may be designed based on the results from a study using this design.

Example

Our example is loosely based on the PanACEA multi-arm, multi-stage tuberculosis trial.¹⁵ Suppose we want to conduct a study evaluating $m=4$ potential new treatments for tuberculosis. For a study with $J=2$ stages, both with performance thresholds of $Z>0$, using

$\alpha = 0.05$ would give us a FWER of 0.0615. Instead, for a desired one-sided FWER of 0.05, we use $\alpha' = 0.0410$, as shown in Table 2. Determining sample size is less straightforward both because the number of treatment arms that will perform sufficiently well relative to the control arm (i.e., k) is unknown *a priori* and because of our sequential rejection method. Because of this, the significance threshold necessary to power a study is unknown. One can choose to err on the side of being overpowered by using $\alpha'/m = \alpha'/4$. Since this is likely to be a bit conservative, we can balance it to some degree by choosing a relatively low power of 0.8. Assuming an effect size of $\delta = \alpha/2$, this yields a per arm sample size of $n = 80$ (40 per arm per stage), with a maximum total sample of 400.

Through simulation, we can estimate expected sample sizes, powers, and percent relative biases for this particular study design. Under the global null, we calculate a relative expected sample size of 0.78 with minimal bias (percent relative bias does not apply). Using the “single effective” setting from before, we get a relative expected sample size of 0.85, a power of 0.89, and percent relative bias of 0.79%. Finally, in the “all effective” setting, we obtain a relative expected sample size of 0.99, a power of 0.88, and a percent relative bias of 0.76%. Note that both powers are calculated using our sequentially rejective method.

Discussion

In this paper, we propose a simple multi-arm, multi-stage design that allows for FWER calculation to be reframed as a relatively simple extension of existing monitoring techniques, in turn allowing control of the FWER based on adjustments to the significance threshold. We have shown that, depending on the design parameters, FWER inflation can present substantial problems in the multi-arm, multi-stage setting, without appropriate adjustments. The proposed method can alleviate these issues, while still providing large efficiency gains by reducing expected sample size.

One aspect of our multi-arm, multi-stage design bears further examination: the inclusion of “terminal dropping.” That is, at the end of the final stage, treatment arms performing relatively poorly compared to control are “dropped” immediately before the final analysis is performed. The appropriateness of this is debatable. There are no logistical or financial gains, as no further recruitment into any of the treatment arms would take place, regardless of their performance. On the other hand, we are able to increase our power for testing the other arms since reducing k , the number of treatment arms continued to the end of the study, increases α'/k . While one might argue this is statistical trickery achieved by previewing the data before conducting the primary analysis, it is important to note two things. First, any dropped arms would not be able to meet the significance thresholds. Second, any possible type I error rate inflation that may accompany this method is adjusted for by the use of α' . For these reasons, and because “terminal dropping” allows us to be consistent across all interim analyses, we chose to apply it.

We should note that the use of “non-binding” futility stopping boundaries does not fit well with our design at a conceptual level. Fundamentally, we are considering the multi-arm, multi-stage framework as a way of reducing the time and resource costs of clinical trials. This is achieved in part by stopping enrollment into poorly performing arms which forces

some degree of binding on to the decision. Therefore we have assumed throughout that stopping rules are binding. If a stopping rule is overruled, so long as it is included in k during hypothesis testing at the end of the study, it seems unlikely that continuing enrollment into said arm would be responsible for renewed FWER inflation. Being certain of this, as well as understanding how such a test would perform with respect to power and other characteristics would require additional investigations.

There are also several generalizations of and extensions to our method that could be pursued. First, one may wish to consider the possibility of stopping treatment arms early for benefit as opposed to just for futility. The appeal of this is readily apparent, although implementing it would raise questions about how stopping a single treatment arm early for benefit would affect the rest of the trial. One may wish to stop enrollment in the control arm, but whether enrollment in the other treatment arms should be discontinued is less clear.

We have explored the operating characteristics of our proposed multi-arm, multi-stage design only for a single, continuous endpoint. A second generalization would be to explore the effects of other endpoints: either using different endpoints for the intermediate and final analyses or using different types of endpoint, such as binary or time-to-event. As an example of the former, a study of cardiovascular health might use reduction in blood pressure as the intermediate outcome, while the final endpoint might be survival free of major cardiac event. Early simulations have suggested that using an intermediate endpoint that differs from the final endpoint results in less precision at intermediate analyses and in turn reduced power. Unexpectedly, there also appears to be increased FWER inflation, though the reason for this is unclear and bears further investigation. As for different types of endpoints, while summary measures quantifying these endpoints will tend to be normally distributed in a large sample setting, their unique properties may need to be taken into account in small sample settings.

These generalizations could only improve a method that we consider already suitable for implementation in the clinical trials setting. In sum, we believe that correctly designed multi-arm, multi-stage studies have great potential for speeding up the clinical trials process without sacrificing desirable statistical properties.

Appendix

Let $\bar{x}_0^{(j)}$ represent the control arm mean (i.e., the mean of the j^{th} batch of control group observations). Let $\bar{x}_i^{(j)}$ represent the mean for treatment arm $i \in 1, \dots, m$ (i.e., the mean of the j^{th} batch of treatment group i 's observations). Both are evaluated at stage $j \in 1, \dots, J$. Further suppose that both are normally distributed and standardized so that their standard errors are 1.

Assuming a 1:1 randomization ratio with equal sample sizes at each stage, we can write the probability that a particular arm (we set $i = 1$ without loss of generality) makes it to the end of the study conditional on the control arm means at each stage as

$$p_A = P \left(\frac{\bar{x}_1^{(1)} - \bar{x}_0^{(1)}}{\sqrt{2 \times 1}} > b_1, \frac{\bar{x}_1^{(1)} + \bar{x}_1^{(2)} - \bar{x}_0^{(1)} - \bar{x}_0^{(2)}}{\sqrt{2 \times 2}} > b_2, \dots, \frac{\sum_{j=1}^J (\bar{x}_1^{(j)} - \bar{x}_0^{(j)})}{\sqrt{2 \times J}} > b_J \mid \bar{x}_1^{(j)} = x^{(j)} \forall j \in 1, \dots, J \right),$$

where $b_j, j \in 1, \dots, J$ are the z -score thresholds for keeping an arm in the study. This quantity can be rewritten as

$$p_A = P \left(\bar{x}_1^{(1)} > \sqrt{2 \times 1} b_1 + x^{(1)}, \right. \\ \frac{\bar{x}_1^{(1)} + \bar{x}_1^{(2)}}{\sqrt{2}} > \frac{\sqrt{2 \times 2} b_2 + x^{(1)} + x^{(2)}}{\sqrt{2}}, \dots, \\ \frac{\sum_{j=1}^J \bar{x}_1^{(j)}}{\sqrt{J}} > \frac{\sqrt{2 \times J} b_J + \sum_{j=1}^J x^{(j)}}{\sqrt{J}} \\ \left. \mid \bar{x}_0^{(j)} = x^{(j)} \forall j \in 1, \dots, J \right).$$

Note that the left hand sides of the inequalities may be viewed as cumulative z -scores $Z_1, \dots, Z_J$ from a one-sample test of $H_0: \mu = 0$ with known variance 1, computed after 1, 2, ..., J equally-sized blocks of information accrue. Thus,

$$p_A = P \left(Z_1 > \gamma_1, \dots, Z_J > \gamma_J \mid \bar{x}_0^{(j)} = x^{(j)} \forall j \in 1, \dots, J \right), \quad (2)$$

where

$$\gamma_j = \gamma_j(b_j) = \frac{\sqrt{2j} b_j + \sum_{i=1}^j x^{(i)}}{\sqrt{j}}$$

and $Z_1, \dots, Z_J$ are multivariate normal with zero means, unit variances, and correlation $\text{cor}(Z_i, Z_j) = (i/j)^{1/2}$ for $i < j$.

Using existing integration techniques developed for the monitoring setting, we can easily calculate such probabilities¹¹. Using nearly identical methods, we can calculate the probability that a particular arm both makes it to the end of the study and is determined to be significant at the level α/k (still conditional on the control arm means at each stage). This probability can be written as

$$p_B = P \left(Z_1 > \gamma_1, \dots, Z_J > \gamma'_J \mid \bar{x}_0^{(j)} = x^{(j)} \forall j \in 1, \dots, J \right). \quad (3)$$

Here, γ_J has been replaced with

$$\gamma'_J = \gamma_J(b'_J),$$

where

$$b'_J = \max\{b_J, \Phi^{-1}(1 - \alpha/k)\},$$

α is our nominal significance level and k is the number of arms that make it to the final stage. We evaluate p_B with the same methods used for evaluating p_A .

Using these quantities, we can write the probability that a *particular* set of k treatment arms makes it to the end of the study, at least one is found to be significant, and the remaining $m - k$ treatment arms do not make it to the end as

$$\{p_A^k (p_A - p_B)^k\} (1 - p_A)^{m-k}.$$

As previously stated, conditioning on the control arm means at each stage implies conditional independence across the treatment arms of a particular treatment arm making it to the end of the study. Therefore, we are able to express this probability relatively succinctly.

To find the conditional probability (given the control arm means at each stage) that *some* set of exactly k treatment arms makes it to the end, at least one is found to be significant, and the remaining $m - k$ treatment arms do not make it to the end, we multiply by the number of ways to select k arms from m :

$$\binom{m}{k} \{p_A^k (p_A - p_B)^k\} (1 - p_A)^{m-k}.$$

Recall that p_A and p_B are both conditional on the sample means in the control arm at each stage and p_B is additionally dependent on k . Therefore, to get the FWER, we must first sum across all possible values of k and then integrate across the assumed distribution of the sample mean in the control arm under the null hypothesis. Here we use the standard normal. This gives,

$$FWER = \int \cdots \int_{\mathbb{R}^J} \sum_{k=0}^m \binom{m}{k} \{p_Z^k - (p_A - p_B)^k\} \times (1 - p_A)^{m-k} \phi(x^{(1)}) \cdots \phi(x^{(J)}) dx^{(1)} \cdots dx^{(J)}.$$

(4)

References

1. Sydes MR, Parmar MK, James ND, et al. Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the mrc stampede trial. *Trials*. 2009; 10(1):39. [PubMed: 19519885]
2. Phillips PP, Gillespie SH, Boeree M, et al. Innovative trial designs are practical solutions for improving the treatment of tuberculosis. *J Infect Dis*. 2012; 205(suppl 2):S250–S257. [PubMed: 22448027]
3. Packer, M. A Clinical and Regulatory Perspective. Novartis Presentations for the March 27, 2014 Meeting of the Cardiovascular and Renal Drugs Advisory Committee. URL <http://www.fda.gov/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/CardiovascularandRenalDrugsAdvisoryCommittee/ucm392157.htm>
4. Royston P, Barthel F, Parmar MK, et al. Designs for clinical trials with time-to-event outcomes based on stopping guidelines for lack of benefit. *Trials*. 2011; 12(1):81. [PubMed: 21418571]
5. Bratton DJ, Phillips PP, Parmar MK. A multi-arm multi-stage clinical trial design for binary outcomes with application to tuberculosis. *BMC Med Res Methodol*. 2013; 13(1):139. [PubMed: 24229079]
6. Wason J, Jaki T. Optimal design of multi-arm multi-stage trials. *Stat Med*. 2012; 31(30):4269–4279. [PubMed: 22826199]
7. Proschan MA, Dodd LE. A modest proposal for dropping poor arms in clinical trials. *Stat Med*. 2014; 33(19):3241–3252. [PubMed: 24757049]
8. Magirr D, Jaki T, Whitehead J. A generalized dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika*. 2012; 99(2):494–501.
9. Halpern SD, Karlawish JH, Casarett D, et al. Hypertensive patients' willingness to participate in placebo-controlled trials: implications for recruitment efficiency. *Am Heart J*. 2003; 146(6):985–992. [PubMed: 14660989]
10. Wason J, Magirr D, Law M, et al. Some recommendations for multi-arm multi-stage trials. *Stat Methods in Med Res*. 2016; 25(2):716–727. [PubMed: 23242385]
11. Proschan, MA., Lan, KG., Wittes, JT. *Statistical Monitoring of Clinical Trials*. New York: Springer; 2006.
12. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat*. 1979; 6(2):65–70.
13. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. 1988; 75(4):800–802.
14. Choodari-Oskoei B, Parmar MK, Royston P, et al. Impact of lack-of-benefit stopping rules on treatment effect estimates of two-arm multi-stage (TAMS) trials with time to event outcome. *Trials*. 2013; 14(1):1. [PubMed: 23286245]
15. Phillips, P. Adaptive designs for clinical trials 2-day workshop. Cambridge: MRC BSU; Adaptive designs for developing better regimens for the treatment of TB: The PanACEA MAMS-TB trial. URL http://www.mrc-bsu.cam.ac.uk/wp-content/uploads/2_Phillips.pdf

Table 1Simulated FWERs for nominal $\alpha = 0.025$ under the global null

Arms	Stages	Threshold		
		$Z > -0.5$	$Z > 0$	$Z > 0.5$
$m = 2$	$J = 1$	0.0243	0.0260	0.0299
	$J = 2$	0.0262	0.0289	0.0328
	$J = 3$	0.0271	0.0297	0.0311
$m = 5$	$J = 1$	0.0218	0.0237	0.0297
	$J = 2$	0.0238	0.0295	0.0405
	$J = 3$	0.0256	0.0329	0.0428
$m = 10$	$J = 1$	0.0200	0.0211	0.0266
	$J = 2$	0.0217	0.0274	0.0417
	$J = 3$	0.0235	0.0321	0.0477

Table 2

α, γ values required to obtain target FWERs

m	FWER = 0.025			FWER = 0.050		
	$J = 1$	$J = 2$	$J = 3$	$J = 1$	$J = 2$	$J = 3$
2	0.0241	0.0215	0.0209	0.0473	0.0425	0.0420
3	0.0246	0.0210	0.0195	0.0488	0.0409	0.0389
4	0.0257	0.0209	0.0192	0.0509	0.0410	0.0380
5	0.0264	0.0211	0.0191	0.0530	0.0415	0.0377
6	0.0272	0.0216	0.0189	0.0545	0.0423	0.0376
7	0.0279	0.0218	0.0191	0.0563	0.0430	0.0378
8	0.0288	0.0223	0.0193	0.0578	0.0436	0.0384
9	0.0289	0.0226	0.0195	0.0592	0.0446	0.0384
10	0.0297	0.0228	0.0197	0.0604	0.0452	0.0392

Table 3

Simulated relative expected sample sizes

Arms	Stages	Setting		
		Global null	Single effective	All effective
$m = 2$	$J = 2$	0.78	0.91	1.00
	$J = 3$	0.66	0.86	0.99
$m = 5$	$J = 2$	0.78	0.83	1.00
	$J = 3$	0.66	0.74	0.98
$m = 10$	$J = 2$	0.77	0.79	1.00
	$J = 3$	0.65	0.69	0.98

Table 4

Simulated powers

		Single effective						All effective					
Arms	Stages	Basic			Sequential			Basic			Sequential		
		Ours	Dunnett	Ours	Ours	Dunnett	Ours	Ours	Dunnett	Ours	Ours	Dunnett	Ours
$m = 2$	$J = 2$	0.94	0.92	0.94	0.92	0.92	0.90	0.92	0.92	0.93	0.93	0.94	0.94
	$J = 3$	0.93	0.92	0.93	0.92	0.92	0.90	0.90	0.92	0.92	0.92	0.94	0.94
$m = 5$	$J = 2$	0.92	0.86	0.92	0.86	0.86	0.83	0.86	0.86	0.91	0.89	0.92	0.92
	$J = 3$	0.91	0.86	0.91	0.86	0.86	0.82	0.86	0.86	0.89	0.89	0.92	0.92
$m = 10$	$J = 2$	0.90	0.81	0.90	0.81	0.81	0.78	0.81	0.81	0.88	0.88	0.90	0.90
	$J = 3$	0.89	0.81	0.89	0.81	0.81	0.76	0.81	0.81	0.86	0.86	0.90	0.90

Table 5

Simulated percent relative biases

Arms	Stages	Setting	
		Single effective	All effective
$m = 2$	$J = 2$	0.32	0.24
	$J = 3$	0.82	0.82
$m = 5$	$J = 2$	0.32	0.30
	$J = 3$	0.76	0.80
$m = 10$	$J = 2$	0.26	0.34
	$J = 3$	0.78	0.80