# STOR601 Research Topic 1: Multi-Arm Multi-Stage Trials

By: Peter Greenstreet, Supervisor: Thomas Jaki

February 21, 2020

## 1 Introduction

Multi-Arm Multi-Stage trials (MAMS) are a class of trial designs which have the aim of improving the efficiency of bringing new treatments to the market. This is important as getting a treatment to market is a long and expensive process, with novel treatments taking between 10-15 years to bring to the market (Kola and Landis (2004) and Dimasi et al. (2003)). Therefore being able to improve the efficiency of introducing new treatments to the market can result in large savings in money and time.

In a MAMS trial several experimental treatments are tested simultaneously against a common control (either an active control or a placebo). We then conduct interim analyses on the treatments to decide which we should continue with, by comparing our different experimental treatments test statistics against our boundaries. MAMS designs therefore have several advantages over classic trial designs, where we run separate controlled treatments against each of our experimental treatments (Wason et al. (2016)):

1. A shared control can be used instead of needing a new one for each treatment group.
2. A direct comparison between each treatment can be conducted which reduces bias, compared to comparisons of treatments which have been tested in separate trials.
3. Interim analysis means we are able to drop ineffective treatments. If we find a clearly superior treatment we are also able to stop early.

MAMS designs can be very useful in phase II trials because these trials are explorative, so we are likely to be testing many treatments, which may not work. The ability to drop ineffective treatments part way through is very useful. Phase III trials are confirmatory trials so we are normally quite sure that the treatment works, we just need the evidence to back it up. However MAMS designs can still be used in phase III trials as discussed in Gaunt et al. (2015).

There are many types of MAMS designs. Such as those discussed in Kelly et al. (2005), Posch et al. (2005) and Whitehead and Jaki (2009). During this report, we are going to focus on a method called group-sequential MAMS design. This design has two or more stages, in which all the treatments are allowed to continue to each stage, given they are sufficiently promising. This method is discussed in Magirr et al. (2012) and Lin and Bunn (2017).

In this report, we are going to study issues faced by group-sequential MAMS designs when working out the sample size required to conduct the study. We begin by defining our notation; then look at the errors we need to control; after which we study the sample size calculations; followed by a section covering the error in current minimum sample size calculations; finally we will summarise our work and look into further research areas.

## 2 Family of null hypotheses

When conducting a MAMS trial our family of null hypotheses is:

$$H_1 : \mu_1 \leq \mu_0, ..., H_K : \mu_k \leq \mu_0, \tag{2.1}$$

where $\mu_1 \dots \mu_k$ are the mean responses on $K$ experimental treatments and $\mu_0$ is the mean response of the control. We will assume that each patient's response to a treatment is independent and that it is normally distributed with known variance $\sigma^2$. At each interim analysis indexed by $j = 1, \dots, J$ we will test our family of null hypotheses 2.1. In our study, we also assume that the sample sizes are equal for each experimental treatment across each stage.

# 3 Errors

In clinical trials there are 2 types of error which we need to control, type I ($\alpha$) and type II ($\beta$) errors. However, in MAMS this is more complicated than in traditional randomised control trials, due to the fact there are multiple hypotheses we are interested in. We will therefore look at how we control the type I error as well as the power of our MAMS trial. Power is one minus the type II error and is the value more commonly used in clinical trials.

## 3.1 Type 1 error

***Definition:*** *Type I error is the probability of any experimental treatment being declared effective when the global null hypothesis is true. Wason and Jaki (2012)*

For a family of hypotheses 2.1 as we have in MAMS trials, to control the family-wise error (FWER) means that the probability of us falsely rejecting the null hypotheses is less than a pre-specified level ($\alpha$). There are two different opinions when it comes to FWER. (Freidlin et al. (2008)) argue not to adjust for FWER in multi-arm trials where each arm corresponds to a different treatment. Their argument is if we did each treatment in separate trials then we would not be subject to multiple testing adjustments. Therefore, they suggest controlling the pairwise type I error rate. This is where you control the type I error between each arm and the control separately, so you therefore ignore any correlation between each of the experimental treatment arms. This method was used in the MRC STAMPEDE trial (Sydes et al. (2009)).

In (Wason et al. (2016)) they say this opinion has it merits however a MAMS trial in its construction is quite different to doing lots of separate trials. Furthermore, FWER provides the maximum probability of recommending an ineffective treatment, which is important in phase III trials. In the European Medicines Agency (EMEA (2002)) state that controlling FWER is required for confirmatory trials. This is becoming more important in phase II trials, as they are sometimes being used as a second pivotal study, when making a confirmatory claim. When conducting a phase II trial, if you want to be able to use this data to back up findings in a phase III trial, it is important that FWER is used.

## 3.2 Power

In a MAMS trial our objective is to find the best treatment, this means that the power of our test depends on both the mean effect of the best treatment and the mean effect of all the other treatments in our trial. In order to find the power in a MAMS trial we use the least favourable configuration. This method was used in the TAILoR trial as discussed in Pushpakom et al. (2015). This method requires us to specify both a clinically relevant difference $\delta_1$ and an uninteresting treatment difference threshold $\delta_0$. The uninteresting treatment difference threshold is the minimum difference between an experimental treatment and the control treatment, that would make that experimental treatment clinically interesting (Wason et al. (2016)). Therefore if $\mu_k - \mu_0 < \delta_0$ we would prefer not to continue investigating treatment $k$.

***Definition:*** *Power is the probability that without loss of generality, $H_1$ is rejected and treatment 1 is recommended given that $\mu_1 - \mu_0 = \delta_1$ and $\mu_k - \mu_0 = \delta_0$ for $k = 2, \dots K$. Van Montfort et al. (2014)*

The values of $\delta_1$ and $\delta_0$ should be made by the clinicians as they have an understanding of what the clinically relevant thresholds are. Both these values will have a large effect when it comes to calculating the sample size.

# 4 Sample size calculation

In MAMS trials the sample size calculations are not as simple as in a classical clinical trial. The key issue is defining what our trial sample size actually is. For example, is it the maximum number of patients you might need or the minimum number of patients, or is it the expected number of patients? During this report we are mainly going to focus on calculating the minimum sample size required with a summary of calculating maximum sample size in section 4.3 and expected sample size in section 4.4. We need to find a sample size which results in our power being satisfied under the least favourable configuration therefore without loss of generality we need to find:

$$P(\text{reject } H_1 | \mu_1 = \delta, \mu_2 = \delta_0, \dots, \mu_K = \delta_0) \geq 1 - \beta, \tag{4.1}$$

where $\delta$ is the difference between treatment 1 and the control.

## 4.1 Boundary calculation

In order to find 4.1 we first have to begin by finding the boundaries at which we are either going to conclude a treatment is superior, or that a treatment is inferior, at each stage. We use the type I error in order to calculate these bounds. We want our boundaries to be such that:

$$P(\text{reject at least one true } H_k, k = 1, \dots, K) \leq \alpha \tag{4.2}$$

This means that our family-wise error rate is less than our pre-specified level. There are different types of designs, which will give you different bounds for each. There are advantages and disadvantages to each as shown in Magirr et al. (2012). Popular examples of boundary designs are from Pocock (1977), O'Brien and Fleming (1979) and the triangular test which is described in Whitehead (1997). The formula for calculating these bounds for these different designs can be found in Magirr et al. (2012).

## 4.2 Number of patient per arm

We begin by calculating the number of patients that we need for each arm at each stage $n$. Under the least favourable configuration that no individual null hypotheses are rejected at analysis $1, \dots, J-1$ and then at analysis $J$ without loss of generality, the null hypothesis $H_1$ is rejected and a treatment 1 is recommended at $\Pi_J$. The power of the study is:

$$\Pi_1 + \Pi_2 + \dots + \Pi_J \tag{4.3}$$

This is the power as $\Pi_j$ is the probability that we reject our null hypothesis at $j$, therefore the power is the probability of correctly rejecting a null hypothesis at either $1, \dots, J$.

In order for us to find the number of patient per arm we solve:

$$\Pi_1 + \Pi_2 + \dots + \Pi_J = 1 - \beta \tag{4.4}$$

To solve this we use the fact that the correlation between test statistics for the experimental treatments comes from the common control treatment, which means that $\Pi_1 + \Pi_2 + \ldots + \Pi_J$ are independent from one another. We then use this fact to calculate each $\Pi_j$. We then find 4.3 for different $n$ until we get our required power 4.4. We calculate $\Pi_j$ using the formula in Magirr et al. (2012).

### 4.3 Maximum number of patients

In order to find the maximum sample size required we look at the worst-case scenario. This is when we are unable to drop any of our treatments or accept them as being superior until the final analysis $J$. This is where we will either accept an experimental treatment, or will reject all of them as there is inefficient evidence to reject any of our null hypotheses 2.1. In order to find the number of patients we need, we solve 4.4 to find $n$ then find $n \times J \times K$.

### 4.4 Expected number of patients

When looking at the expected number we do not know how the treatments are going to perform before doing the trial. For example are we expecting only one treatment to be clinically relevant or multiple treatments, or maybe none and when will we find this treatment? For a method on calculating the expected sample size see Magirr et al. (2012). Their method works by us calculating the number of patients we will need in all the different scenarios then multiplying each of these by the probability that the scenarios happen.

## 5 Calculation of minimum number of patients

In Magirr et al. (2012) they have also calculated the minimum number of patients needed for a MAMS trial. This can be calculated by taking the maximum number of patients needed, then dividing this by the number of stages $J$. This is a very intuitive method for calculating this, as in the best scenario you find a treatment, which is clinically relevant by the end of your first stage. However, this calculation underestimates the true number, for a couple of key reasons:

1. It takes a period of time from collecting the data to do the statistical analysis on it.
2. Most treatments take time before we can measure the treatments effect.

Both these issues could be resolved if, once you have recruited the right number of patients to your trial for each stage, you stop recruiting new patients and wait for the results. However, this has two key issues. The first of which is this would drastically increase the length of time your trial could take. For example if we had a 3 stage MAMS trial where it takes 3 months before we can take the treatments effect measurements and the data takes 1 month to analyse, we could potentially have to wait an additional 8 months compared to if we did not stop recruitment. The second, and possibly the biggest issue, is that by human nature if we stop a clinical trial part way through it is then incredibly hard to recruit new patients to the trial after the pause. This is because people think that the trial may have stopped for other reasons such as issues with treatment safety. Therefore pausing the trial at different intervals is not a viable option.

We are going to see how the impacts of different treatment effect periods and statistical analysis periods affect the minimum patient numbers in our trial. We are also going to study the affect of different recruitment rates on the minimum number of patients. We are going to assume that we recruit patients at a constant rate throughout the trial. This results in the minimum number of patients ($N_{\min}$) equalling:

$$N_{\min} = \min\Big(\frac{N_{\max}}{J} + d \times (a + t), N_{\max}\Big), \tag{5.1}$$

where $a$ statistical analysis period, $t$ treatment effect period, $d$ recruit rate. In figures 1 and 2 we have set the default $N_{\max} = 540$, $J = 3$, $a = 14$ days, $d = 30$ days, $r = 2$ per day we have chosen $N_{\max} = 540$ and $J = 3$ as this is one of the examples given in Magirr et al. (2012).

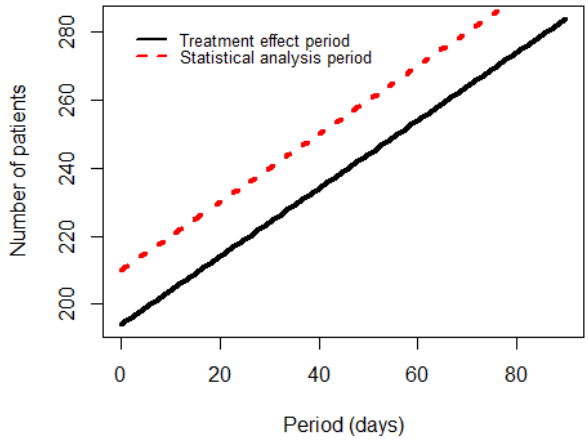


Figure 1: Affect of treatment effect period and Statistical analysis period on the minimum number of patients needed for a trial

As you can see from equation 5.1 in figure 1 as we increase both the treatment effect period and statistical analysis period we increase $N_{\min}$. In most cases we are unable to change the treatment effect period as there is normally no way of making a treatment faster, especially if you are using a control which has a set treatment period time. We may be able to decrease the statistical analysis period, however this will involve an increase in resources, resulting in an increased cost of doing the analysis.

Furthermore as we can see in equation 5.1 and figure 2 as we decrease the recruitment rate we decrease the minimum number of patients. We might think of designing our trial so we have a slow recruitment rate however this also has issues. As you can see in figure 2 the recruitment rate has a huge influence on the time our trial will take, even if the trial only needs to recruit the minimum patient numbers.

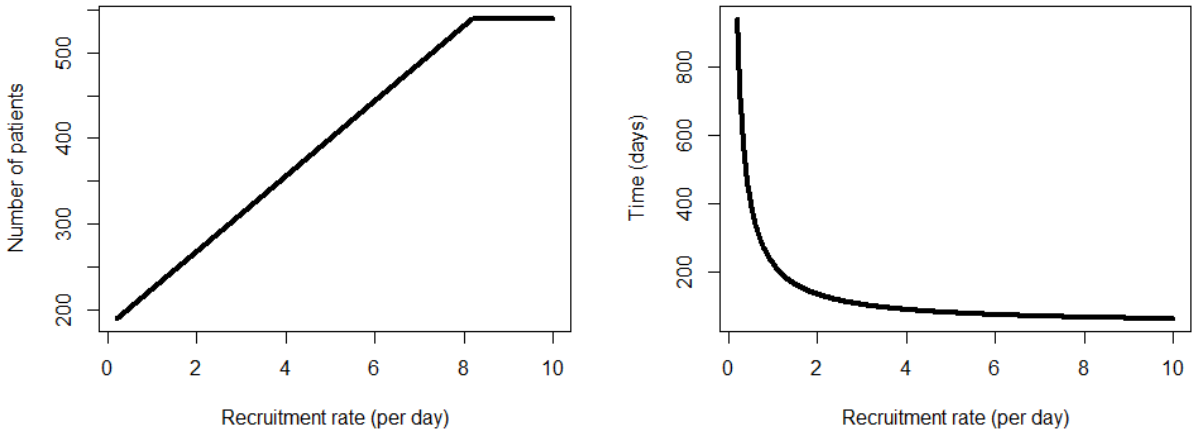


Figure 2: Recruitment rate effect on the minimum number of patients needed for a trial and on the time the trial takes.

# 6 Conclusion

In this report we have studied the group-sequential MAMS design with a focus on looking at the sample size for the maximum and minimum number of patients needed. We began by looking at the type I and type II errors as these are both important factors in calculating sample size. After which we looked at calculating sample size for the maximum number and expected number of patients needed for our MAMS design. Finally, we studied how we calculate the minimum number of patients.

When calculating the minimum number of patients we showed that Magirr et al. (2012) calculation for this is a best case scenario. We have shown that when you count for recruitment rate, treatment period and statistical analysis period their estimate can hugely underestimate the minimum patients required. This leads us to some important questions about group-sequential MAMS designs, before we consider using this in a clinical trial. The more stages we have the more interim analyses we will have to conduct, each of which costs money. Therefore, if we are unlikely to save very much on recruitment levels to a trial, is it worth the increased cost?

In figure 2 we saw in our example if we have a recruitment rate of above 8 patients per day, then our minimum number of patients will be the same as our maximum number of patients. However the maximum number of patients we need to recruit for a $J$ stage MAMS trial is more than a $J-1$ stage MAMS trial. It may be that a multi-stage multi-arm trial has a larger minimum number of patients than the maximum number needed in a single-stage multi-arm trial. Therefore, it is very important that we take into account recruitment rate, treatment period and statistical analysis period before choosing to use a MAMS trial design.

# 7 Further research

Further work to be done in this area would be to study the effects of recruitment rate, treatment period and statistical analysis period on the expected number of patients needed. This would involve first calculating the expected number of treatments to drop at each stage or the expected time until a treatment is classed as superior. Then from this we could calculate the expected number of patients needed. We could also look into ways of calculating how many stages would be optimal for patient numbers and overall cost. Finally, we would also like to look into the effect of having a non-uniform recruitment rate such as those discussed in Minois et al. (2017). This is more realistic as when we recruit more centres to our trial the recruitment rate will increase. This will affect the minimum number of patients we will need for a trial.

# References


Dimasi, J. A., Hansen, R. W., and Grabowski, H. G. (2003). The price of innovation: new estimates of drug development costs. *Journal of Health Economics*, 22(2):151–185. 1

EMEA (2002). Committee for proprietary medicinal products (cpmp). 2

Freidlin, B., Korn, E. L., Gray, R., and Martin, A. (2008). Multi-arm clinical trials of new agents: some design considerations. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 14(14):4368. 2

Gaunt, P., Mehanna, H., and Yap, C. (2015). The design of a multi-arm multi-stage (mams) phase iii randomised controlled trial comparing alternative regimens for escalating (compare) treatment of intermediate and high-risk oropharyngeal cancer with reflections on the complications of introducing a new experimental arm. *Trials*, 16(s2). 1

Kelly, P. J., Stallard, N., and Todd, S. (2005). An adaptive group sequential design for phase ii/iii clinical trials that select a single treatment from several. *Journal of Biopharmaceutical Statistics*, 15(4):641–658. 1

Kola, I. and Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery*, 3(8):711. 1

Lin, J. and Bunn, V. (2017). Comparison of multi-arm multi-stage design and adaptive randomization in platform clinical trials. *Contemporary Clinical Trials*, 54:48–59. 1

Magirr, D., Jaki, T., and Whitehead, J. (2012). A generalized dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika*, 99(2):494–501. 1, 3, 4, 5, 6

Minois, N., Savy, S., Lauwers-Cances, V., Andrieu, S., and Savy, N. (2017). How to deal with the poisson-gamma model to forecast patients' recruitment in clinical trials when there are pauses in recruitment dynamic? *Contemporary Clinical Trials Communications*, 5(23):144–152. 6

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, 35(3):549–556. 3

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199. 3

Posch, M., Koenig, F., Branson, M., Brannath, W., Dunger-Baldauf, C., and Bauer, P. (2005). Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine*, 24(24):3697–3714. 1

Pushpakom, S. P., Taylor, C., Kolamunnage-Dona, R., Spowart, C., Vora, J., Garcia-Finana, M., Kemp, G. J., Whitehead, J., Jaki, T., Khoo, S., Williamson, P., and Pirmohamed, M. (2015). Telmisartan and insulin resistance in hiv (tailor): protocol for a dose-ranging phase ii randomised open-labelled trial of telmisartan as a strategy for the reduction of insulin resistance in hiv-positive individuals on combination antiretroviral therapy. *BMJ Open*, 5(10). 2

Sydes, M. R., Parmar, M. K. B., James, N. D., Clarke, N. W., Dearnaley, D. P., Mason, M. D., Morgan, R. C., Sanders, K., and Royston, P. (2009). Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the mrc stampede trial. *Trials*, 10:39. 2

Van Montfort, K., Oud, J., and Ghidey, W. (2014). Developments in statistical evaluation of clinical trials. 3

Wason, J., Magirr, D., Law, M., and Jaki, T. (2016). Some recommendations for multi-arm multi-stage trials. *Statistical Methods in Medical Research*, 25(2):716–727. 1, 2

Wason, J. M. S. and Jaki, T. (2012). Optimal design of multi-arm multi-stage trials. *Statistics in Medicine*, 31(30):4269–4279. 2

Whitehead, J. (1997). The design and analysis of sequential clinical trials. *Biometrics*, 53(4):1564. 3

Whitehead, J. and Jaki, T. (2009). One and two stage design proposals for a phase ii trial comparing three active treatments with control using an ordered categorical endpoint. *Statistics in Medicine*, 28(5):828–847. 1