The Journal of Open Source Software

gsMAMS: an R package for Designing Multi-Arm Multi-Stage Clinical Trials

Tushar Patni [1], **Yimei Li** [1], **and Jianrong Wu** [2]

1 St. Jude Children's Research Hospital, Memphis, TN, United States of America **2** University of New Mexico Comprehensive Cancer Center, Albuquerque, NM, United States of America

DOI: 10.21105/joss.06322

Software

- Review
- Repository
- Archive

Editor: Antonia Mey

Reviewers:

- @njtierney
- @RhysPeploe

Submitted: 12 January 2024
Published: 28 May 2024

Summary

In the field of clinical trial design, a multi-arm trial allows simultaneous comparison of multiple experimental treatments with a common control and provides a substantial efficiency advantage compared to conventional randomized controlled trials. A multi-stage trial allows multiple interim looks at the trial outcome so that ineffective arms can be stopped early. In the current R ecosystem, `MAMS` Jaki et al. (2019) is one of the few packages that can handle multiple stages, multiple treatment arms, and different types of endpoints, but the computational effort of obtaining sample size and sequential stopping boundaries is very high when the number of stages exceeds 3. More importantly, their method may lead to an under powered study when the endpoint is time-to-event data. Therefore, we introduce the R package `gsMAMS` (group sequential multi-arm multi-stage (MAMS)) for designing MAMS trials with continuous, ordinal, and survival outcomes, which is computationally very efficient even for a number of stages greater than 3. We also discuss the applications of the package. The `gsMAMS` package facilitates the design and planning of multi-arm clinical trials, which simultaneously compare multiple experimental treatments with a common control group. It is intended to be used by clinical researchers and statisticians so that they can determine appropriate sample sizes and stopping criteria for these trials.

Statement of Need

Traditional two-arm randomized controlled trials are not an optimal choice when multiple experimental arms are available for testing efficacy (Parmar et al. (2014)). In such situations, a multiple-arm trial should be preferred, which allows simultaneous comparison of multiple experimental arms with a common control and provides a substantial efficiency advantage. In multi-arm trials, several arms are monitored in a group sequential fashion, with ineffective arms being dropped out of the study. Therefore, multi-arm trials offer a more efficient, cost-effective, and patient-centered approach to clinical research, with the potential to identify superior treatments more reliably than traditional two-arm trials (Parmar et al. (2014)).

Some packages that are available in R (e.g., `adaptTest` Vandemeulebroecke (2022), `asd` Hack et al. (2022), and `AGSDEST` Parsons et al. (2011)) have limitations either in the number of treatment arms that can be incorporated in the package, the number of interim analyses that can be implemented in the package, or the different kinds of outcomes that the package can handle, but the `MAMS` package works well for both multiple treatment arms and multiple stages. It also works for continuous, ordinal, and survival outcomes. However, the computational effort of obtaining stopping boundaries is very high when the number of stages exceeds 3. The long computational time is the major drawback of using the `MAMS` package, and as a result, the process of generating design parameters and evaluating potential modifications can be time-intensive, sometimes taking several hours. This lengthy computation time becomes

problematic when researchers or clinicians need to iteratively tweak design parameters and promptly assess the resulting changes in the trial design.

This paper introduces the R package gsMAMS available at https://cran.r-project.org/web/packages/gsMAMS/, which provides functions to obtain sample size, and efficacy and futility boundaries for multiple stages and multiple experimental arms. It also provides functions that generate operating characteristics for designing trials with continuous, ordinal, and survival outcomes. It is computationally very fast compared to the MAMS package, even for a number of stages greater than 3.

This package will serve well for clinicians, researchers, and statisticians who are involved in designing phase-II clinical trials with multiple treatment arms because it is easy to use, provides comprehensive information regarding the trial characteristics, and provides a substantial efficiency advantage compared to traditional randomized controlled trials.

Computational Aspects

The computational complexity of this package is very low because our proposed method is based on the sequential conditional probability ratio test (SCPRT) procedure, which provides analytical solutions for both futility and efficacy boundaries for an arbitrary number of stages and arms. Thus, it avoids complicated computational efforts in obtaining stopping boundaries (Wu et al. (2023)). The family-wise error rate (FWER) is controlled by Dunnett correction, which entails finding the root of an integral of a multivariate normal distribution. The multivariate normal densities are evaluated using the package mvtnorm (Genz & Bretz (2009)). The package is efficient for any number of treatment arms and stages, but it has a limitation that it is only configured for 10 stages. In practice, a study rarely needs to have more than 10 interim looks planned. To give an example, the computational time of the MAMS package to obtain stopping boundaries and sample size for a multi-arm trial with continuous outcome, four experimental arms, and three stages is approximately 7 minutes, and for four stages, it is approximately 4.5 hours. But for the gsMAMS package with the same trial configuration, the computational time to obtain stopping boundaries and sample size for three stages and four stages design is approximately 0.06 seconds for both cases. This is approximately 7000-27000 times faster compared to the MAMS package.

The operating characteristics for continuous and ordinal outcomes require less computational effort than the survival outcomes. The computational burden of sample size and sequential conditional probability ratio test (SCPRT) boundary calculation for continuous, ordinal, and survival outcomes is minimal because there are only two critical components of the algorithm, which are the roots of FWER and power, i.e., critical value and sample size.

The below code shows an example demonstration about the computational efficiency of our package, where we are designing a four-stage trial for four treatment arms, and we have used the same equivalent parameters for both designs.

```
##MAMS design
system.time(mams(K = 4,
                 J = 4,
                 p = NULL,
                 p0 = NULL,
                 delta = 0.545,
                 delta0 = 0.178,
                 sd = 1,
                 r = 1:4,
                 r0 = 1:4,
                 alpha = 0.05,
                 power = 0.9,
                 ushape = "obf",
```

```
                lfix = 0))
#   user system  elapsed
#5220.67   5.84  8860.42

##gsMAMS design
system.time(design_cont(delta0 = 0.178,
                        delta1 = 0.545,
                        alpha = 0.05,
                        beta = 0.1,
                        k = 4,
                        frac = c(1:4 / 4)))
#  user  system elapsed
#  0.07    0.00    0.06
```

Based on the results, the elapsed time is very high for the MAMS package.

Application

In this section, we will demonstrate the use of gsMAMS package and provide a separate example for each type of outcome.

Continuous outcome

For the continuous outcome, we will consider the TAILoR trial (Pushpakom et al. (2015)), which is a phase II trial, and it compares three doses of telmisartan (20, 40, 80mg) with no intervention (control) for the reduction of insulin resistance in human immunodeficiency virus-positive patients receiving combination antiretroviral therapy. The primary outcome measure is a reduction in the mean homeostasis model assessment of insulin resistance (HOMA-IR) score at 24 weeks. The standardized desirable and minimal effect sizes for efficacy are set as $\delta^{(1)} = 0.545$ for the 80mg group and $\delta^{(0)} = 0.178$ for the 20 and 40 mg groups, respectively, for the trial design.

The sample size calculation is based on a one-sided type I error of 5% and a power of 90%. Based on the trial characteristics, we will design the trial for a two-stage design.

The design parameters of the trial can be calculated using the `design_cont()` function, and the arguments in the function correspond to the standardized effect size in the ineffective arm (`delta0`) and effective arm (`delta1`), type I error (`alpha`), type II error (`beta`), total number of treatment arms (`k`), and the information time (0.5, 1) is denoted by the (`frac`) argument in the function.

```
#Installing the package from CRAN
install.packages("gsMAMS")
#Loading the library
library(gsMAMS)

set.seed(1234)
design_cont(delta0 = 0.178,
            delta1 = 0.545,
            alpha = 0.05,
            beta = 0.1,
            k = 3,
            frac = c(0.5, 1))

## $`Sample size`
##                                          Stage 1 Stage 2
## Cumulative sample size for treatment group      40      79
```

```
## Cumulative sample size for control group        40      79
##
## $`Maximum total sample size for the trial`
## [1] 316
##
## $`Boundary values`
##             Stage 1 Stage 2
## Lower bound   0.006   2.062
## Upper bound   2.910   2.062
```

The design output shows the cumulative sample size for the treatment and control groups at each stage. The SCPRT lower and upper boundaries are (0.006, 2.062) and (2.91, 2.062), respectively. Based on the design parameters, the first interim analysis can be conducted after the enrollment of 40 patients in the control arm. If the test statistic $Z_{k,l} < 0.006$, the k_{th} arm is rejected for futility at the 1^{st} stage, and the trial continues with the remaining treatment arms and the control. If the test statistic $0.006 \leq Z_{k,l} \leq 2.91$ for $k = 1, 2, 3$, then the trial continues to the next stage, and 39 patients are further enrolled per arm. If $Z_{k,1} > 2.91$ for some k, the trial is terminated, and the arm with the maximum value of $Z_{k,1}$, $k = 1, 2, 3$, would be recommended for further study.

For FWER and Stagewise FWER:

The operating characteristics of the trial can be generated using the `op_power_cont()` and `op_fwer_cont()` functions for power under the alternative hypothesis and FWER under the global null hypothesis, respectively. Most of the arguments in the functions are similar to the `design_cont()` function, with the exception of the number of simulations (`nsim`) and the seed number (`seed`).

```
op_fwer_cont(alpha = 0.05,
             beta = 0.1,
             p = 3,
             frac = c(0.5, 1),
             delta0 = 0.178,
             delta1 = 0.545,
             nsim = 10000,
             seed = 10)
```

```
## $FWER
## [1] 0.05
##
## $`Stagewise FWER`
##  look1  look2
## 0.0050 0.0466
##
## $`Stopping probability under null`
##  look1  look2
## 0.2599 0.7401
##
## $`Probability of futility under null`
##  look1  look2
## 0.2549 0.6949
##
## $`Average sample size used per arm under null`
## [1] 61.645
```

Based on the simulation results, the type I error at the first interim analysis is 0.5%, and at the

second interim analysis, it is approximately 4.6%. Therefore, the overall type I error of the trial is close to 5%. The sample size required for the trial was 79 patients per arm, but the trial used approximately an average of 62 subjects per arm. The stopping probability (probability of stopping the trial either due to futility or efficacy) should add up to 1, which is the case here. Since this is under the null configuration, the probability of futility (probability of stopping the trial when all the treatment arms become futile in the trial) should be approximately 95%, and it holds true in this case.

For Power and Stagewise Power:

```
op_power_cont(alpha = 0.05,
              beta = 0.1,
              p = 3,
              frac = c(0.5, 1),
              delta0 = 0.178,
              delta1 = 0.545,
              nsim = 10000,
              seed = 10)
```

```
## $Power
## [1] 0.893
##
## $`Stagewise Power`
##  look1  look2
## 0.3126 0.5804
##
## $`Stopping probability under alternative`
##  look1  look2
## 0.3258 0.6742
##
## $`Probability of futility under alternative`
##  look1  look2
## 0.0035 0.0821
##
## $`Average sample size used per arm under alternative`
## [1] 62.652
```

Based on the simulation results, the probability of success at the first stage is 31.26% and at the second stage is approximately 58.04%. Therefore, the overall power is approximately 90%. The sample size required for the trial was 79 patients per arm, but the trial used approximately an average of 62 subjects per arm. The stopping probability should add up to 1, which is the case here, and under the alternate configuration, the probability of futility is approximately 8.5%, which is less than 10% type II error. The reason is that the type II error comes from both failing to find any efficacious arm (futility) and finding the less efficacious arm as the most efficacious arm. The latter part was not included when the probability of futility was calculated.

Ordinal Outcome

For the ordinal outcome, we will consider the ASCLEPIOS trial (Whitehead (1993)), a phase II trial for patients with stroke. The primary outcome response is the patient's Barthel index assessed 90 days after randomization. This is an ordered categorical score ranging from 0 (vegetative state) to 100 (complete recovery) in steps of 5, and relates to the activities of daily living that the patient is able to undertake. Following the ASCLEPIOS study, we group the outcome categories of the score into six larger categories. We will consider the treatment

worthwhile if the odds ratio between the effective and control arms is 3.06, and we set the null odds ratio to be 1.32, which is the odds ratio between the ineffective and control arms.

The sample size calculation is based on a one-sided family-wise error rate (FWER) of 5% and a power of 90%. Based on the trial characteristics, we will design the trial for a three-stage design. The design parameters for a five-arm (k = 4) trial can be calculated using the `design_ord()` function, and the arguments in the function correspond to the probability of outcomes in the control group (`prob`), the odds ratio of the ineffective treatment group vs. control (`or0`), the odds ratio of the effective treatment group vs. control (`or`), and the remaining arguments are similar to the `design_cont()` function for the continuous outcome.

```
design_ord(prob = c(0.075, 0.182, 0.319, 0.243, 0.015, 0.166),
           or = 3.06,
           or0 = 1.32,
           alpha = 0.05,
           beta = 0.1,
           k = 4,
           frac = c(1/3, 2/3, 1))
```

```
## $`Sample size`
##                                           Stage 1 Stage 2 Stage 3
## Cumulative sample size for treatment group      21      42      62
## Cumulative sample size for control group        21      42      62
##
## $`Maximum total sample size for the trial`
## [1] 310
##
## $`Boundary values`
##             Stage 1 Stage 2 Stage 3
## Lower bound  -0.630   0.437   2.161
## Upper bound   3.126   3.092   2.161
```

Based on the design parameters, the first interim analysis can be conducted after the enrollment of 21 patients (information time is 21/62, which is approximately 1/3 at this stage) in the control and treatment arms. If the test statistic $Z_{k,l} < -0.63$, the k_{th} arm is rejected for futility at the 1^{st} stage, and the trial continues with the remaining treatment arms and the control. If the test statistic $-0.63 \leq Z_{k,l} \leq 3.126$ for $k = 1, 2, 3, 4$, then the trial continues to the next stage, and 21 patients are further enrolled per arm. If $Z_{k,1} > 3.126$ for some k, the trial is terminated, and the arm with the maximum value of $Z_{k,1}$, $k = 1, 2, 3, 4$, would be recommended for further study. A similar procedure is followed if the trial goes to the second stage of interim analysis.

The operating characteristics can be generated using the functions `op_fwer_ord()` and `op_power_ord()`, which are similar to those of the continuous outcome.

Survival Outcome

For the survival outcome, we will consider a MAMS trial with five arms (four treatment arms and a control arm, k = 4) and two interim looks with balanced information time (`frac` = c(0.5, 1)). The null hazard ratio is (`hr0`) 1, and the alternative hazard ratio is (`hr1`) 0.67. The median survival time of the control group is 20 months, and the survival distribution is exponential without loss to follow-up. The sample size calculation is based on a one-sided type I error of 5% and a power of 90%.

The design parameters for a two-stage design can be calculated using the `design_surv()` function, and the arguments in the function correspond to the median survival time of the control group (`m0`), the hazard ratio of the ineffective treatment vs. control (`hr0`), the hazard

ratio of the effective treatment vs. control (`hr1`), the accrual time (`ta`), the follow-up time (`tf`), the shape parameter of the Weibull distribution (kappa), and the rate of loss to follow-up (`eta`) (assumed loss to follow-up follows an exponential distribution with rate parameter eta).

```
design_surv(m0 = 20,
            hr0 = 1,
            hr1 = 0.67032,
            ta = 40,
            tf = 20,
            alpha = 0.05,
            beta = 0.1,
            k = 4,
            kappa = 1,
            eta = 0,
            frac = c(0.5, 1))
```

```
## $`Sample size`
##                                                                 Stage 1 Stage 2
## Cumulative number of events for combined treatment & control      164     328
##
## $`Maximum total number of events for the trial`
## [1] 820
##
## $`Total number of subjects required for the trial`
## [1] 1170
##
## $`Boundary values`
##             Stage 1 Stage 2
## Lower bound   0.075    2.16
## Upper bound   2.980    2.16
```

The operating characteristics of the trial can be generated using the `op_power_surv()` and `op_fwer_surv()` function.

The design output shows the cumulative number of events for the treatment and control groups combined at each stage, along with the futility and efficacy boundaries. The total number of subjects required for the trial per arm is 234.

Based on the design parameters, the first interim analysis can be conducted after the incidence of 164 events (information time is 164/328, which is 0.5 at this stage), which results from the aggregation of events in the control and treatment arms. If the test statistic $Z_{k,l} < 0.075$, the k_{th} arm is rejected for futility at the 1^{st} stage, and the trial continues with the remaining treatment arms and the control. If the test statistic $0.075 \leq Z_{k,l} \leq 2.98$ for $k = 1, 2, 3, 4$, then the trial continues to the next stage, and we will wait for the incidence of an additional 164 events to conduct the next interim analysis. If $Z_{k,1} > 2.98$ for some k, the trial is terminated, and the arm with the maximum value of $Z_{k,1}$, $k = 1, 2, 3, 4$, would be recommended for further study.

The operating characteristics of the trial can be generated using the `op_power_surv()` and `op_fwer_surv()` functions.

For FWER and Stagewise FWER:

```
op_fwer_surv(m0 = 20,
             alpha = 0.05,
             beta = 0.1,
             p = 4,
```

```
                frac = c(1/2, 1),
                hr0 = 1,
                hr1 = 0.6703,
                nsim = 10000,
                ta = 40,
                tf = 20,
                kappa = 1,
                eta = 0,
                seed = 12)
```

```
## $FWER
## [1] 0.05
##
## $`Stagewise FWER`
##  look1  look2
## 0.0049 0.0460
##
## $`Stopping probability under null`
##  look1  look2
## 0.2318 0.7682
##
## $`Probability of futility under null`
##  look1  look2
## 0.2269 0.7234
##
## $`Average number of events happened per arm under null`
## [1] 129.5902
##
## $`Average duration of trial(months)`
## [1] 54.27148
```

Based on the simulation results, the type I error is approximately 0.4% at the first interim analysis and approximately 4.6% at the second interim analysis. Therefore, the overall type I error is maintained at 5%. The stopping probability should be approximately 1, which is the case here, and since this is under the null configuration, the probability of futility should be approximately 95%, which holds true in this case. The average duration of the trial is 54 months, which is reasonable as the total duration of the trial is 60 months. The average number of events that happened per arm is 130.

For Power and Stagewise Power :

```
op_power_surv(m0 = 20,
              alpha = 0.05,
              beta = 0.1,
              p = 4,
              frac = c(1/2, 1),
              hr0 = 1,
              hr1 = 0.6703,
              nsim = 10000,
              ta = 40,
              tf = 20,
              kappa = 1,
              eta = 0,
              seed = 12)
```

```
## $Power
```

```
## [1] 0.913
##
## $`Stagewise Power`
##  look1  look2
## 0.3270 0.5863
##
## $`Stopping probability under alternative`
##  look1  look2
## 0.3334 0.6666
##
## $`Probability of futility under alternative`
##  look1  look2
## 0.0059 0.0800
##
## $`Average number of events happened per arm under alternative`
## [1] 115.0483
##
## $`Average duration of trial(months)`
## [1] 52.27229
```

Based on the simulation results, the probability of success at the first stage is 32.7% and at the second stage is approximately 58.63%. Therefore, the desired power of 90% has been met. The overall stopping probability should add up to 1, which is the case here, and the probability of futility is approximately 8.5%, which is less than 10% type II error because of the same reason as mentioned in the continuous outcome. The average duration of the trial is 52 months, which is reasonable as the total duration of the trial is 60 months. The average number of events that happened per arm is 115.

Acknowledgements

Dr. Wu's research was supported by the University of new Mexico Comprehensive Cancer Center Support Grant National Cancer Institute (NCI) P30CA118100 and Dr. Li's research was supported by the Comprehensive Cancer Center at St. Jude Children's Research Hospital and American Lebanese Syrian Associated Charities (ALSAC).

References

Genz, A., & Bretz, F. (2009). *Computation of multivariate normal and t probabilities*. https://doi.org/10.1007/978-3-642-01689-9

Hack, N., Brannath, W., & Brueckner, M. (2022). *AGSDest: Estimation in adaptive group sequential trials*. https://CRAN.R-project.org/package=AGSDest

Jaki, T., Pallmann, P., & Magirr, D. (2019). The R package MAMS for designing multi-arm multi-stage clinical trials. *Journal of Statistical Software*, *88*(4), 1–25. https://doi.org/10.18637/jss.v088.i04

Parmar, M. K. B., Carpenter, J., & Sydes, M. R. (2014). More multiarm randomised trials of superiority are needed. *The Lancet*, *384*(9940), 283–284. https://doi.org/10.1016/S0140-6736(14)61122-3

Parsons, N., Friede, T., Todd, S., & Stallard, N. (2011). Software tools for implementing simulation studies in adaptive seamless designs: Introducing r package ASD. *Trials*, *12*(1), A8. https://doi.org/10.1186/1745-6215-12-S1-A8

Pushpakom, S., Taylor, C., Kolamunnage-Dona, R., Spowart, C., Vora, J., García-Fiñana, M., Kemp, G., Whitehead, J., Jaki, T., Khoo, S., Williamson, P., & 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), e009566. https://doi.org/10.1136/bmjopen-2015-009566

Vandemeulebroecke, M. (2022). *adaptTest: Adaptive two-stage tests*. https://CRAN.R-project.org/package=adaptTest

Whitehead, J. (1993). Application of sequential methods to a phase III clinical trial in stroke. *Drug Information Journal*, *27*(3), 733–740. https://doi.org/10.1177/009286159302700315

Wu, J., Li, Y., & Zhu, L. (2023). Group sequential multi-arm multi-stage trial design with treatment selection. *Statistics in Medicine*, *42*(10), 1480–1491. https://doi.org/10.1002/sim.9682