Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

The R Package MAMS for Designing Multi-Arm Multi-Stage Clinical Trials

Thomas Jaki
Lancaster University

Philip Pallmann
Lancaster University

Dominic Magirr
AstraZeneca

Abstract

In the early stages of drug development there is often uncertainty about the most promising among a set of different treatments, different doses of the same treatment, or combinations of treatments. Multi-arm multi-stage (MAMS) clinical studies provide an efficient solution to determine which intervention is most promising. In this paper we discuss the R package **MAMS** that allows designing such studies within the group-sequential framework. The package implements MAMS studies with normal, binary, ordinal, or time-to-event endpoints in which either the single best treatment or all promising treatments are continued at the interim analyses. Additionally unexpected design modifications can be accounted for via the use of the conditional error approach. We provide illustrative examples of the use of the package based on real trial designs.

Keywords: adaptive designs; conditional error; multi-arm, R, step-down procedure.

1. Introduction

Developing new medicines and health technologies is time-consuming and expensive. The development of one novel treatment has been estimated to take 10–15 years and costs several hundred million pounds on average (**?**). Because early-phase studies frequently evaluate treatments on short-term endpoints, uncertainty often exists about which of a set of candidate treatments should be selected for testing in a confirmatory phase III clinical trial. These candidates can be truly different medications but can also be different doses of the same drug or different combinations of multiple drugs. The high failure rate of phase III trials of around 50% (**?**) combined with their substantial cost (**?**) make selecting an appropriate treatment for evaluation in phase III of paramount importance.

Seamless phase II/III multi-arm clinical trials that compare several active treatments with a common control group are one potentially efficient solution to overcome this problem. These

seamless studies use the initial part of the study (phase II) to learn about all treatments while an in-depth evaluation occurs only on the promising one(s) in the second part (phase III). The design typically applied for such an endeavour is called multi-arm multi-stage (MAMS). Using data accumulated across both parts of the trial implies that decisions about the superiority of (a) treatment(s) will be reached in a more efficient manner than with separate trials for phases II and III. Incorporating a series of interim analyses to allow early stopping either for efficacy or to drop ineffective treatments early has recently received attention (e.g., **?????**).

While group-sequential designs found their way into standard software like R (**?**), Stata (StataCorp, College Station, TX), and SAS (SAS Institute, Cary, NC) years ago and are well established (**?**), programs for adaptive methods are still relatively rare. Here we give a brief overview of available software implementations focusing on MAMS designs. Summaries of software for adaptive designs more generally can be found in **?** and the book chapters by **?** and **?**.

At present, there are three relevant R packages for adaptive trials on CRAN. **adaptTest** (**?**) implements four different adaptive tests for two-arm two-stage designs. **AGSDest** (**?**) computes estimates, p values, and lower one-sided confidence intervals for two-arm multi-stage designs, using the conditional error principle to adjust for adaptations. The only package that allows the design of multi-arm trials is **asd** (**?**). It provides a simulation function for multi-arm two-stage designs with different selection rules at interim and uses p value combination methods (inverse normal or Fisher's combination test) for the outcome analysis (**?**). We are not aware of any R package for adaptive trials with more than two arms and more than two stages.

Stata modules for group-sequential designs are available (e.g., **?**), but there are no routines for more general adaptive methods, with the exception of **nstage** (**?**) and **nstagebin** (**?**) that both implement MAMS designs. **nstage** has been available for several years (**?**) and was recently updated (**?**). This module differs from the **MAMS** package described in this paper in two principal ways. Firstly, the theory behind the **MAMS** package is based on score statistics allowing continuous, binary, ordinal, and time-to-event data to be used as the primary endpoints while **nstage** has been developed explicitly for time-to-event endpoints. Secondly, the **MAMS** package focuses on designs that strictly control the FWER in the strong sense. Similarly, **nstagebin** is only applicable to designs with binary outcomes, and its focus is not on strong FWER control either.

SAS does not offer any functionality for adaptive methods other than group-sequential: the only adaptation that can be made with the procedures **SEQDESIGN** and **SEQTEST** is early stopping. Extensions by means of user-written programs are of course possible. **?** provides a wealth of SAS macros for general adaptive designs. A SAS macro **ADCCT** for two-stage designs with two or three active groups and mid-course treatment selection (**?**) is available from `https://cemsiis.meduniwien.ac.at/user/koenig-franz/research/software/sas-macros-adcct`.

Besides implementations of adaptive designs in commonly used statistical software, there are also a few commercial standalone packages that cover certain types of confirmatory adaptive methods, most notably East (Cytel, Cambridge, MA) and ADDPLAN (ICON, Dublin, Ireland). The latter has an additional module **MC** with a wide range of functions for MAMS designs, including sample size and power calculations, treatment selection at interim, sample size reassessment, performance analyses (e.g., selection probabilities of specific treatment arms),

closed testing, multiplicity-adjusted p values, and simultaneous confidence intervals.

In this manuscript we introduce the R package **MAMS** (**?**), which is an implementation of the methods proposed in **?**, **?**, and **?**. It facilitates the design of MAMS trials in which all promising treatments are selected at a series of interim analyses. This has not been possible so far in R for designs involving more than two arms and stages. Other convenient features of **MAMS** are that unplanned design modifications are smoothly adjusted for, and it can use a parameterisation of effect sizes that does not require any knowledge about the variability in advance. A step-down variant where either the most promising treatment or all promising treatments can be selected at interim is available, too.

In Section 2 we give an overview of the methodology described in **?** and the ideas to allow for unexpected design modifications (**?**). In Section 3 we highlight a few computational aspects of the package. Section 4 provides examples of the use of the package based on real trials focusing on the interpretation of the results before we conclude with a discussion in Section 5.

2. The underlying methodology

We consider the situation where K treatments are compared to a common control group (indexed by zero). Of interest is to test if treatment k is superior to control, which formally corresponds to

$$H_{01} : \mu_1 \leq \mu_0, \quad \ldots, \quad H_{0K} : \mu_K \leq \mu_0 \ ,$$

where μ_k is the mean response of a patient on treatment $k = 0, 1, \ldots, K$. Observations of the primary endpoint are modelled as normally distributed random variables with equal known variance, σ^2, but potentially different mean levels, $\mu_0, \ldots, \mu_K$. At up to J time points the data accumulated so far will be analysed and each treatment arm either continued or stopped. If all experimental arms are stopped, the trial terminates. We denote the number of subjects on control during the first stage by n and the number of observations on arm k up to stage j by $r_k^{(j)} n$. Often equal numbers of observations on each experimental treatment are assumed so that $r_1^{(j)} = \ldots = r_K^{(j)} = r^{(j)}$. To allocate an equal number of subjects to each treatment and each stage, $r^{(j)} = r_0^{(j)} = j r^{(1)}$, for example.

2.1. Designing the study

To decide which treatments are continued and which are stopped at each analysis time point, comparative Z statistics are defined as

$$Z_k^{(j)} = \frac{\hat{\mu}_k^{(j)} - \hat{\mu}_0^{(j)}}{\sigma \sqrt{\frac{r_k^{(j)} + r_0^{(j)}}{r_k^{(j)} r_0^{(j)} n}}}, \quad k = 1, \ldots K \ ; \ j = 1, \ldots, J \ ,$$

where $\hat{\mu}_k$ is the mean measurement on treatment $k = 0, \ldots, K$ and σ denotes the known standard deviation. At stage j recruitment to treatment k is stopped for futility if $Z_k^{(j)} \leq l_j$, i.e., the corresponding test statistics is below a pre-defined lower boundary value l_j.

Similarly we can reject the null hypothesis of non-superiority of treatment k over control if the corresponding test statistic exceeds an upper boundary, $Z_k^{(j)} > u_j$. In the event that one or more treatments are shown to be superior to control, the trial stops. If $l_j < Z_k^{(j)} \leq u_j$ further patients are recruited to each remaining treatment k and control. We set the final boundary value $l_J = u_J$ to enforce a decision about every superiority hypothesis at the end of the trial.

To determine the boundaries $(l_j, u_j),\ j = 1, \ldots, J$ we require that the familywise error rate (FWER), defined as

$$P(\text{reject at least one true } H_{0k},\ k = 1, \ldots, K),$$

be controlled at a pre-specified level α. This probability can analytically be found as a J dimensional integral that depends on the boundaries, $l_1, \ldots l_{J-1}, u_1, \ldots, u_J$ and how subjects are allocated to the arms and stages. The latter is usually pre-specified, leaving $2 \times J - 1$ unknowns when finding the FWER of a design. To determine the boundary values to ensure FWER control different approaches can be taken. The first is to determine the boundaries to satisfy some optimality criterion subject to controlling the FWER as described in **?**. The second approach is to utilise an error-spending approach as described in **?**. The third approach pre-specifies functions that relate the boundaries to the final critical value, $l_j = g(u_J),\ u_j = f(u_J)\ j = 1, \ldots, J-1$, leaving one equation and one unknown. The third option is explicitly implemented in the **MAMS** package while it can be used as the basis for the other two ideas as well; this would require the user to write a suitable optimisation routine.

To obtain the required sample size, the least favourable configuration (LFC, **?**) can be utilised. The power under the LFC is, without loss of generality, defined as the probability to reject H_{01} given $\mu_1 - \mu_0 = \delta$ and $\mu_k - \mu_0 = \delta_0$ for $k = 2, \ldots, K$ where δ is an effect that, if present, we would like to detect with high probability and δ_0 is an effect that, if present, would not be of interest.

In addition to the standard parameterisation of the effect sizes in terms of δ and δ_0, the **MAMS** package implements a slightly non-standard parameterisation as probabilities. The interesting treatment effect, p, that if present we would like to find with high probability and an uninteresting effect, p_0 are both parameterised as $P(X_k > X_0)$ where X_i denotes the random response on treatment i, i.e., the probability of a randomly selected person on treatment k observing a better outcome than a random person on control. As a consequence $p = 0.5$ implies that both the experimental treatment and control perform equally well. We utilise this parameterisation as no knowledge about the variance, σ^2, is required. To obtain p from the traditional effect size, δ, one can simply use $p = \Phi\left(\frac{\delta}{\sqrt{2\sigma^2}}\right)$.

2.2. Dealing with unexpected design modifications

In the previous section we have outlined how a MAMS design can be determined. Unfortunately, certain aspects of the design are often quite difficult to specify a-priori, and it is well known that deviations from the planned design, such as changing the sample size or dropping treatment arms, can compromise the operating characteristics of the design, and in particular FWER and power, dramatically. It is therefore useful to be able to adjust the design to account for such unexpected deviations. Here we will outline the underlying methodology to do so, which is described comprehensively for MAMS designs in **?**.

The first ingredient to allow design modifications is the conditional error principle (**?**). It utilises the conditional probability of rejecting H_{0k} under the null hypothesis given a design and the data observed so far, called the conditional error. The conditional error principle then states that a new design controls the FWER if the conditional error of this design does not exceed the conditional error of the original design. As a consequence the conditional error approach can be used to adjust for design modifications within each of the pairwise comparisons contrasting one experimental arm against control.

Secondly, to ensure FWER control over all comparisons, the closure principle (**?**) can be utilised. In order to apply the closure principle to the family of null hypotheses, $H_1, \ldots, H_K$ (omitting the index 0 for brevity), local hypothesis tests for all intersection hypotheses, $H_I = \cap_{k=1}^{K} H_K$, are found. The elementary hypothesis H_k can then be rejected at level α if and only if all H_I containing H_k can be rejected at level α.

3. Computational aspects of the MAMS package

After a brief description of the underlying methodology we will now highlight some of the computational aspects that have been used in the implementation of the **MAMS** package. The first notable aspect is the computation of the FWER and power, which do not have a closed-form solution. Instead an integral of dimension J (the number of stages) over multivariate normal distributions needs to be evaluated. The multivariate normal densities are evaluated using the package **mvtnorm** (**?**), which utilises the algorithms of **?**. The outer integrals are solved using quadrature and the midpoint rule. The number of points used for the quadrature can be controlled with the argument `N`, but the default will hardly ever have to be changed by the user. The computational complexity in solving the FWER and power constraint lies in the number of stages rather than the number of treatment arms. Consequently, the implementation used in the package is efficient for any number of treatment arms provided that the number of stages is small. For designs with more than two stages, however, computing a design might take several minutes.

When computing the power of the design we use a "divide and conquer" strategy. In order to determine the power it is necessary to find the probability of stopping which, by construction, could happen at any stage. To simplify the computation, it is helpful to observe that it is still only possible to stop at exactly one of the stages. As a consequence it is possible to find the probability of stopping at each stage separately and simply sum over all these probabilities in order to find the overall power. To determine the sample size required to achieve a pre-specified power a simple loop is used. To potentially speed computation up further it is possible to set the starting value of the loop via `nstart`. This is particularly helpful as it is well known that the maximum sample size of a design with interim analyses will not typically be smaller than the equivalent fixed-sample design for continuous endpoints. In a similar manner, the loop is terminated at a maximum value for the sample size, which is by default 3 times the sample size of the equivalent fixed-sample design with all other parameters unchanged, but this can be overridden by specifying a value for `nstop`.

4. Applications

In this section we showcase some uses of the **MAMS** package and how to interpret the cor-

responding R output. The TAILoR study (**?**) serves as the motivating example, and so we consider a design that evaluates three different experimental treatment arms against control, using a one-sided type I error rate of 5% and 90% power. The interesting effect size is set to $p = 0.65$, which corresponds to an effect of $\delta = 0.545\sigma$ on the traditional scale. The uninteresting treatment effect is chosen as $p_0 = 0.55$ ($\delta_0 = 0.178\sigma$). **MAMS** allows the user to choose whichever parameterisation they prefer for specifying the effect sizes.

4.1. A single-stage design

Designing studies including finding the boundaries of the design and the required sample size can be achieved with the function `mams`. The parameters of the function correspond to the definition in Section 2 so that `K`, e.g., specifies the number of experimental treatments that are to be compared to control, and `J` the number of stages. We begin by considering a single-stage design (`J=1`), which corresponds to a design based on a standard Dunnett test (**?**) involving `K=3` experimental treatments. We use equal allocation between treatment arms, which is specified via `r=1` for the experimental arms and `r0=1` for control.

```
R> library("MAMS")
R> m1 <- mams(K=3, J=1, p=0.65, p0=0.55, r=1, r0=1, alpha=0.05, power=0.9)
```

An overview of the design is displayed with `print(m1)` or `summary(m1)` or simply `m1`.

```
R> m1
Design parameters for a 1 stage trial with 3 treatments

                                           Stage 1
Cumulative sample size per stage (control):      79
Cumulative sample size per stage (active):       79

Maximum total sample size:  316

             Stage 1
Upper bound:   2.062
Lower bound:   2.062
```

The output produced specifies the number of patients required on control and each treatment arm as well as the boundaries of the design. A total of 316 patients, 79 on control and 79 on each of the 3 experimental treatments, are required for this study. The null hypothesis for treatment k can be rejected if the corresponding test statistic is larger than 2.062.

The same design can also be specified on the scale of traditional effect sizes rather than probabilities, by setting `p` and `p0` to `NULL` and specifying values for `delta`, `delta0`, and `sd`. The output will be exactly the same as for `m1`.

```
R> m1d <- mams(K=3, J=1, p=NULL, p0=NULL, delta=0.545, delta0=0.178, sd=1,
+    r=1, r0=1, alpha=0.05, power=0.9)
```

In the remainder of this section we will specify all effect sizes on the probability scale, but converting them is straightforward in R:

```
R> pnorm(0.545/sqrt(2))
[1] 0.6500195
R> qnorm(0.65) * sqrt(2)
[1] 0.5449254
```

4.2. Multi-stage designs with different boundary shapes

Since only a single stage was used in this initial example, no form of the boundaries had to be specified. For multi-stage designs the shape of the lower and upper boundary can be defined via the arguments `lshape` and `ushape`. These arguments can either invoke the pre-defined shapes following **?**, **?** or the triangular test (**?**) using options `"pocock"`, `"obf"`, or `"triangular"`, respectively. Alternatively a constant value (option `"fixed"`) can be specified. Finally, custom boundaries can be defined as a function that requires exactly one argument for the number of stages and returns a vector of the same length. The lower boundary shape must be non-decreasing while the upper boundary shape must be non-increasing to ensure reasonable trial designs are found.

In the following example we calculate a two-stage design investigating three experimental treatments. Triangular boundaries are used with a cumulative sample size ratio of `r=1:2` between first and second stage, i.e., the interim analysis is scheduled after half of the maximum number of patients have been recruited and their outcome observed, and twice as many subjects on control as on the experimental arms, as specified by `r0=c(2, 4)`.

```
R> m2 <- mams(K=3, J=2, p=0.65, p0=0.55, r=1:2, r0=c(2, 4), alpha=0.05,
+    power=0.9, ushape="triangular", lshape="triangular")
R> m2
Design parameters for a 2 stage trial with 3 treatments

                                           Stage 1 Stage 2
Cumulative sample size per stage (control):      76     152
Cumulative sample size per stage (active):       38      76

Maximum total sample size:  380

             Stage 1 Stage 2
Upper bound:   2.359   2.225
Lower bound:   0.786   2.225
```

The cumulative sample sizes at stages 1 and 2 are given in tabular form in the R output. The trial may be stopped after the first analysis, either for futility (if all the Z statistics are less than 0.786) or superiority (if at least one Z statistic exceeds 2.359). In all other cases the trial is to be taken to the second stage where additional patients are randomised to any experimental treatment whose Z statistic falls between the boundary values of the first stage and control. A critical value of 2.225 is used at the second analysis to decide whether a treatment shall be deemed superior to control or not.

Our next example involves three treatment arms in a three-stage design with equal numbers of subjects added at every stage as well as balance of sample size between control and treatment

groups; this requires us to specify the cumulative sample sizes as `r=1:3` and `r0=1:3`. To illustrate the versatility of the function `mams`, we do not use any of the pre-defined boundary shapes. Instead we implement a fixed lower bound of zero (with `lshape="fixed"` and `lfix=0`) and an upper boundary where the first-stage critical value is three times as large as the final critical value. To achieve this, `ushape` is specified as a function that returns the vector (3, 2, 1) (`return(x:1)`).

```
R> m3 <- mams(K=3, J=3, p=0.65, p0=0.55, alpha=0.05, power=0.9, r=1:3, r0=1:3,
+    ushape=function(x) return(x:1), lshape="fixed", lfix=0)
R> m3
Design parameters for a 3 stage trial with 3 treatments

                                          Stage 1 Stage 2 Stage 3
Cumulative sample size per stage (control):      27      54      81
Cumulative sample size per stage (active):       27      54      81

Maximum total sample size:  324

             Stage 1 Stage 2 Stage 3
Upper bound:   6.125   4.083   2.042
Lower bound:   0.000   0.000   2.042
```

The maximum total sample size is considerably lower than with design `m2` (324 versus 380), and so is the critical value at the final stage (2.042 versus 2.225). These feigned advantages come, however, at the cost of very large upper boundary values at stages 1 and 2 (6.125 and 4.083) that make it extremely hard to stop the trial early, so this is unlikely to be a useful design in practice. On a related note, if a design should not allow stopping for one of efficacy or futility at all, we can achieve this by setting `lfix=-Inf` or `ufix=Inf`, respectively.

We compare the boundaries of our "own" design `m3` with those of the corresponding standard designs (Pocock, O'Brien-Fleming, triangular) graphically using the `plot` function that comes with the **MAMS** package. First we have to compute the boundaries of the standard designs for `J=3` stages and sample size allocations as in `m3`. Notice that the computation of designs with more than 2 stages can take several minutes.

```
R> poc <- mams(K=3, J=3, p=0.65, p0=0.55, r=1:3, r0=1:3, alpha=0.05, power=0.9,
+    ushape="pocock", lshape="pocock")
R> obf <- mams(K=3, J=3, p=0.65, p0=0.55, r=1:3, r0=1:3, alpha=0.05, power=0.9,
+    ushape="obf", lshape="obf")
R> tri <- mams(K=3, J=3, p=0.65, p0=0.55, r=1:3, r0=1:3, alpha=0.05, power=0.9,
+    ushape="triangular", lshape="triangular")
```

Then we plot the boundaries with identical scaling of the vertical axes (using the argument `ylim`) to make the graphs visually comparable:

```
R> par(mfrow=c(2, 2))
R> plot(poc, ylim=c(-5, 7), main="Pocock")
```

```
R> plot(obf, ylim=c(-5, 7), main="O'Brien-Fleming")
R> plot(tri, ylim=c(-5, 7), main="Triangular")
R> plot(m3, ylim=c(-5, 7), main="Self-designed")
```

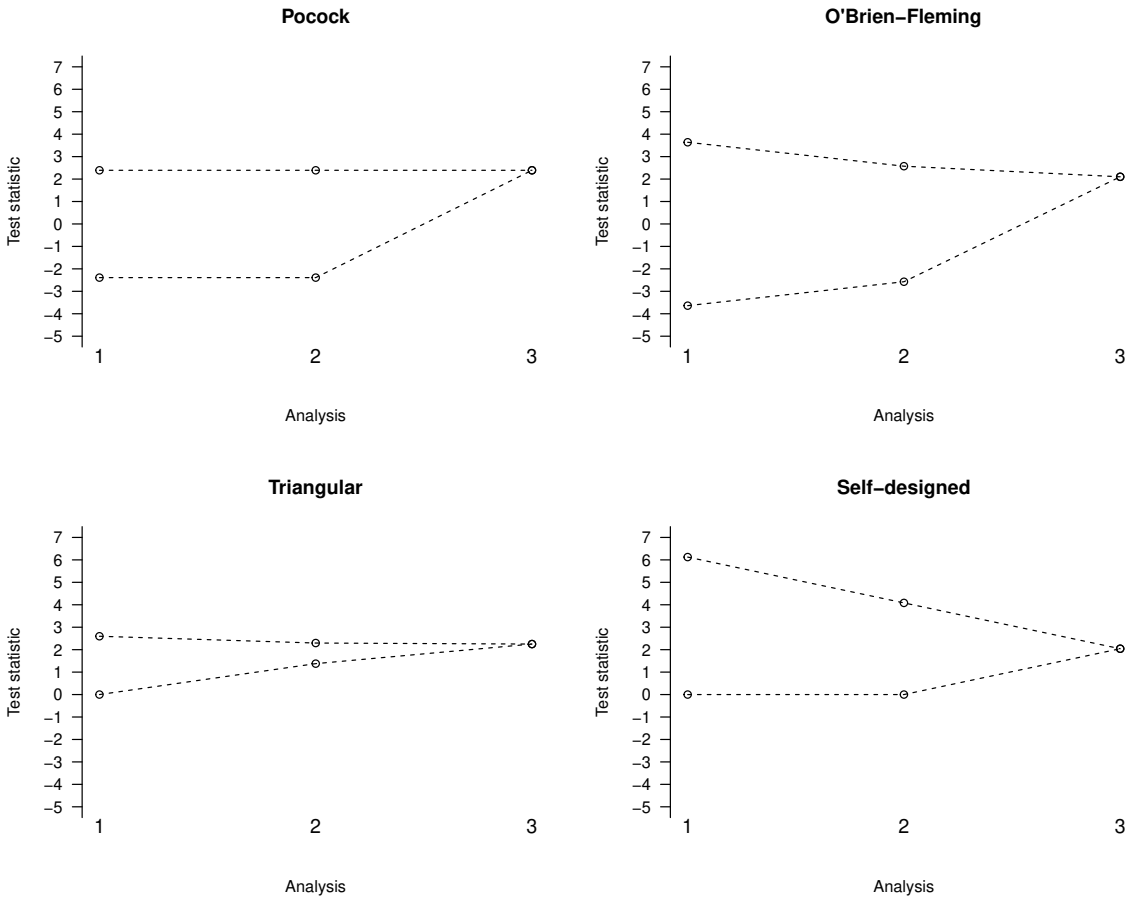

Figure 1: Stopping boundaries for a three-arm three-stage design using the methods of Pocock (top left), O'Brien-Fleming (top right), the triangular test (bottom left), and our own design with a fixed lower bound of zero and an upper bound whose first-stage critical value is three times as large as the final one (bottom right).

Figure 1 displays the shapes of all four designs. We see that the triangular design has clearly the narrowest boundaries (and therefore the highest chances of stopping the trial early) whereas the self-designed variant leads to extraordinarily high upper boundary values at the first two interim analyses.

4.3. Evaluating the properties of a design

To evaluate the properties of a particular design via simulation, the function `mams.sim` can be employed. It allows for flexible numbers of subjects per arm and stage in the form of a $J \times (K + 1)$ matrix `nMat`. In addition to the upper and lower boundaries (`u` and `l`), a vector

of true success probabilities (`pv`) is required (or alternatively a vector of true effect sizes (`deltav`) and a standard deviation (`sd`)). The parameter `ptest` allows to specify rejection of which hypotheses should be counted in the power calculation. We evaluate the properties of the two-stage design `m2` under the global null hypothesis (i.e., a true effect size of $p = 0.5$ or $\delta = 0$ for all treatments) with 100,000 simulation runs.

```
R> m2sim <- mams.sim(nsim=1e5, nMat=t(m2$n * m2$rMat), u=m2$u, l=m2$l,
+    pv=rep(0.5, 3), ptest=1:2)
R> m2sim
Simulated error rates based on 1e+05 simulations


Prop. rejecting at least 1 hypothesis:               0.049
Prop. rejecting first hypothesis (Z_1>Z_2,...,Z_K)   0.017
Prop. rejecting hypotheses 1 or 2:                   0.034
Expected sample size:                              244.578
```

The probability of rejecting at least one hypothesis is 0.049, and since we simulated under the global H_0, this corresponds to a FWER of 5% as desired. The power to reject the first hypothesis when it has the largest estimated effect is 0.017, and the power to reject either H_1 or H_2 or both of them (as specified by `ptest=1:2`) is 0.034. The expected number of patients required for the trial under the global H_0 is 244.6 in contrast to the maximum required of 380.

The function `mams.sim` is also useful to simulate and compare expected sample sizes of different designs. We illustrate this for the designs `poc`, `obf` and `tri` (whose boundaries are shown in Figure 1) under the LFC of the alternative, i.e., one treatment's effect size equals $p = 0.65$ whereas the effect sizes for all other treatments are equal to $p_0 = 0.55$, using 100,000 simulation runs.

```
R> pocsim <- mams.sim(nsim=1e5, nMat=t(poc$n * poc$rMat), u=poc$u, l=poc$l,
+    pv=c(0.65, rep(0.55, 2)), ptest=1)
R> obfsim <- mams.sim(nsim=1e5, nMat=t(obf$n * obf$rMat), u=obf$u, l=obf$l,
+    pv=c(0.65, rep(0.55, 2)), ptest=1)
R> trisim <- mams.sim(nsim=1e5, nMat=t(tri$n * tri$rMat), u=tri$u, l=tri$l,
+    pv=c(0.65, rep(0.55, 2)), ptest=1)
```

Similarly, we can obtain the design properties under the global null hypothesis by setting `pv=rep(0.5, 3)`. Table 1 summarises minimum, maximum, and expected sample sizes of the three designs. We see that under both the LFC and the global H_0 the triangular design is expected to require the lowest number of patients. On the other hand, O'Brien-Fleming has the lowest minimum and maximum but the highest expected sample size under the LFC of all three designs. Under the global H_0 both the Pocock and O'Brien-Fleming designs have expected sample sizes that are very close to their respective maxima.

4.4. A step-down design

Design	Minimum	Maximum	Expected (LFC)	Expected (H_0)
Pocock	165	396	232.4	385.6
O'Brien-Fleming	140	336	259.2	334.0
Triangular	170	408	217.3	222.3

Table 1: Minimum, maximum, and (simulated) expected sample sizes of three-stage designs involving three experimental treatment arms with Pocock, O'Brien-Fleming, and triangular boundaries under the least favourable configuration of the alternative (LFC) and the global null hypothesis (H_0).

A direct improvement over the basic design can be achieved by using a step-down version of the test. The function `stepdown.mams` implements such a design that selects at interim either all promising treatments (`selection="all.promising"`) or only the best performing treatment (`selection="select.best"`) from those whose test statistics are between the upper and lower boundaries (**?**). If the trial is stopped early, making a selection becomes obsolete, but note that we consider stopping boundaries as non-binding. The step-down procedure makes use of closed testing, as described in Section 2.2 and **?**.

We reuse the three-arm two-stage design `m2` with 76 and 38 observations in stage 1 and cumulative sample sizes at stage two of 152 and 76 on control and active treatments, respectively. The sample size of the study can be specified through the matrix `nMat` that has J rows and K columns where the first column contains the values for the control group. A lower boundary can be set via `lb`, and we set it to 0.786 as in the original design. We can then choose how much of the total familywise error we want to spend at each stage using the argument `alpha.star`, and we choose to spend, in line with the triangular test, $\alpha_1^* = 0.026$ at the first interim analysis, with the remaining α level being used at the second analysis. We compare the selection rules `all.promising` and `select.best`:

```
R> m2.all <- stepdown.mams(nMat=matrix(c(76, 152, rep(c(38, 76), 3)),
+    nrow=2, ncol=4), lb=m2$l[1], alpha.star=c(0.026, 0.05),
+    selection="all.promising")
R> m2.best <- stepdown.mams(nMat=matrix(c(76, 152, rep(c(38, 76), 3)),
+    nrow=2, ncol=4), lb=m2$l[1], alpha.star=c(0.026, 0.05),
+    selection="select.best")
```

The output for all intersections and stages is verbose, therefore we summarise only the upper boundary values in Table 2 with the full output provided in Appendix A. One can see that the option to proceed with more than one promising treatment comes at the cost of higher stopping boundaries for the intersection hypotheses at stage 2. So in order to reject the global null hypothesis at the second stage, a test statistic needs to exceed 2.22 while it only needs to be larger than 2.17 if only the best arm is chosen at the interim analysis. The boundaries are the same for the elementary hypotheses H_1, H_2, and H_3 as well as for the intersection hypotheses H_{12}, H_{13}, and H_{23} because the sample sizes were chosen to be equal in all active treatment arms. A graphical display of the stopping boundaries using the `plot` function is shown in Figure 2.

```
R> par(mfrow=c(1, 2))
R> plot(m2.all, main="Select all promising", col=c(1, 1, 2, 1, 2, 2, 4))
R> plot(m2.best, main="Select the best",  col=c(1, 1, 2, 1, 2, 2, 4))
```

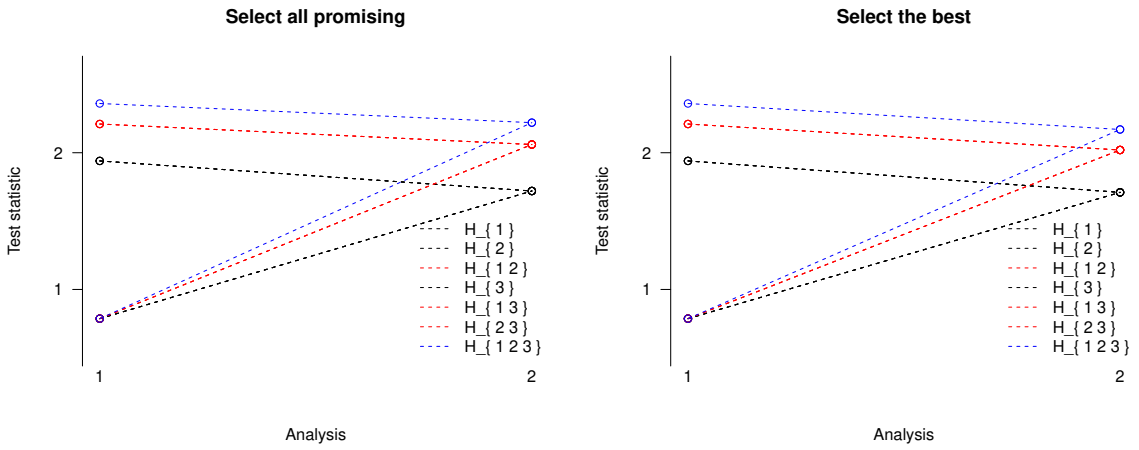

Figure 2: Stopping boundaries for a three-arm two-stage step-down design with selection of all promising treatments at interim (left) or just the single best treatment (right).

Hypotheses	Stage 1	Stage 2 (best)	Stage 2 (all)
H_1, H_2, H_3	1.94	1.71	1.72
H_{12}, H_{13}, H_{23}	2.21	2.02	2.06
H_{123}	2.36	2.17	2.22

Table 2: Upper boundaries for the elementary (H_1, H_2, H_3), intersection (H_{12}, H_{13}, H_{23}), and global (H_{123}) hypotheses of a three-arm two-stage step-down design involving selection of either the single best or all promising treatments at interim.

4.5. Dealing with unforeseen design modifications

Two other functions of the package, `new.bounds` and `stepdown.update`, allow for unexpected design modifications to be taken into account. The function `new.bounds` recalculates the boundary values when the sample sizes achieved are not as planned in advance. We consider again the two-stage design `m2` where 76 patients were required per stage in the control arm and 38 patients per stage for each of the three experimental treatment arms. Now assume these requirements could not be met and the observed sample sizes at the interim analysis were 75 for control and 40, 35, and 41 for the experimental treatments. We can recalculate the final boundary value with `new.bounds` in which we specify the interim bounds `u=2.359` and `l=0.786` (as obtained for `m2`). The sample sizes as observed at stage 1 and planned for stage 2 are given in the $J \times (K + 1)$ matrix `nMat`.

```
R> m2.nb <- new.bounds(K=3, J=2, nMat=matrix(c(75, 152, 40, 76, 35, 76, 41, 76),
+    nrow=2, ncol=4), alpha=0.05, u=m2$u[1], l=m2$l[1], ushape="triangular",
+    lshape="triangular")
R> m2.nb
Design parameters for a 2 stage trial with 3 treatments

             Stage 1 Stage 2
Upper bound:   2.359   2.224
Lower bound:   0.786   2.224
```

We find that as a result of the deviation from the planned sample size at the interim analysis, the final boundary value has been lowered from 2.225 in the original `m2` design to 2.224.

The function `stepdown.update` uses the conditional error approach to incorporate unplanned sample size reassessment and/or treatment selection (e.g., elimination of treatment arms due to safety issues) while maintaining control of the desired FWER. We once again base this example on the original three-arm two-stage design but consider the step-down version (`m2.all`) and assume there were some unforeseen design changes during the course of the trial. Initially the sample sizes at interim were planned to be 76 for the control group and 38 per active treatment arm. At the interim analysis, we now wish to take into account three deviations from the planned study. Firstly, we want to account for the deviation from the desired sample size which, as in the previous example, turned out to be 75 for control and 40, 35, and 41 for the experimental treatments, which translates to `nobs=c(75, 40, 35, 41)` in the function `stepdown.update`. Secondly, treatment 2 has been dropped from the study due to safety, so that only treatment arms 1 and 3 (`selected.trts=c(1, 3)`) are to be continued. Finally, following a reassessment of the sample size, we wish to increase the cumulative sample size at the second stage by 50% from 152 to 228 in the control arm and 76 to 114 in the active arms. We can specify this using `nfuture=matrix(c(228, 114, 35, 114), 1, 4)`. Notice that since treatment arm 2 has already been abandoned, no additional patients are recruited beyond the 35 already in the study. Further supposing the interim evaluation yielded Z statistics of `zscores=c(1.1, 0.9, 0.9)`, we can calculate the modified design.

```
R> m2.update <- stepdown.update(current.mams=m2.all, nobs=c(75, 40, 35, 41),
+    zscores=c(1.1, 0.9, 0.9), selected.trts=c(1, 3),
+    nfuture=matrix(c(228, 114, 35, 114), nrow=1, ncol=4))
```

The complete output of `m2.update` is provided in Appendix B and we summarise it in Table 3 (column "Updated"). The boundaries for the elementary hypothesis H_1 and H_3 have been slightly increased to account for the change in sample size while the boundary for H_2 has been slightly decreased. Similarly, the boundary for the intersection hypothesis involving only the remaining treatments (H_{13}) has been increased while the others have been decreased. No change in the threshold for the global null hypothesis (H_{123}) is observed in this example. As before we can also illustrate the updated design using the `plot` function.

4.6. Non-normal endpoints

Up to this point we have focused on normally distributed endpoints. Based on asymptotic

Hypothesis	Initial	Updated (cond. error)
H_1	1.72	1.73 (0.088)
H_2	1.72	1.71 (0.069)
H_3	1.72	1.79 (0.058)
H_{12}	2.06	1.92 (0.056)
H_{13}	2.06	2.14 (0.051)
H_{23}	2.06	1.90 (0.043)
H_{123}	2.22	2.22 (0.041)

Table 3: Initial and updated upper boundaries (with conditional errors) for the elementary (H_1, H_2, H_3), intersection (H_{12}, H_{13}, H_{23}), and global (H_{123}) hypotheses of a three-arm two-stage step-down design involving selection of all promising treatments at interim. Treatment 2 has been dropped at the interim analysis and the sample size for the remaining comparisons increased.

theory, **MAMS** can also handle non-normal endpoints by exploiting the asymptotic properties of efficient score statistics (**?**), as we will demonstrate for ordinal, binary, and time-to-event outcome data.

Ordinal and binary endpoints

Ordinal data consist of multiple different categories that have a natural order, which is common for quality-of-life scores, pain scores, and similar questionnaire-based outcomes. Our illustration here is motivated by the ASCLEPIOS study (**?**) and its example analyses in **?** and **?**.

We design a MAMS trial with three experimental treatments and a control arm, one interim analysis after half the patients have provided an outcome measure, and triangular boundaries in a setting with an ordinal primary endpoint, under the assumption of proportional odds. We expect that under control conditions the probabilities of falling into each of six categories, ordered from best to worst, are 0.075, 0.182, 0.319, 0.243, 0.015, and 0.166. Suppose the interesting effect is a doubling in the probability of falling into one of the two best categories combined, from 25.7 to 51.4%, for any experimental arm. This corresponds to an odds ratio (OR) of 3.06 and a log-OR of 1.12. The uninteresting effect shall be one quarter of the interesting effect on the log-OR scale i.e., a log-OR of 0.28 or an OR of 1.32.

To find the boundary values and sample sizes, we can use the function `ordinal.mams`, which is a wrapper for `mams` with additional inputs `prob` for the probabilities of falling into each category (which must sum up to one), as well as `or` and `or0` for the interesting and uninteresting treatment effects, respectively, on the OR scale:

```
R> prob <- c(0.075, 0.182, 0.319, 0.243, 0.015, 0.166)
R> mord <- ordinal.mams(prob=prob, or=3.06, or0=1.32, K=3, J=2, alpha=0.05,
+    power=0.9, r=1:2, r0=1:2, ushape="triangular", lshape="triangular")
R> mord
Design parameters for a 2 stage trial with 3 treatments

                                                Stage 1 Stage 2
```

```
Cumulative sample size per stage (control):      34      68
Cumulative sample size per stage (active):       34      68

Maximum total sample size:  272

             Stage 1 Stage 2
Upper bound:   2.330   2.197
Lower bound:   0.777   2.197
```

The function `ordinal.mams` can also be used for binary endpoints as they are a simple special case of ordinal data where `prob` has only two categories (success/failure, yes/no, etc.) and the proportional odds assumption becomes obsolete.

Time-to-event endpoints

Another useful extension of **MAMS** is to event-time outcomes e.g., when the primary endpoint is survival. In that case the effect sizes δ and δ_0 must be specified in terms of log-hazard ratios (log-HRs), which are assumed to be asymptotically normal, and the standard deviation is $\sigma = 1$. Sample sizes are expressed in terms of events (e.g., deaths), e, rather than numbers of patients, n. As a consequence, we set $r_k^{(1)} = 1$ and $r_k^{(j)} = e_0^{(j)}/e_0^{(1)}$. The underlying approximation should work well if the effect size is small, the number of allocated patients patients per arm is equal at each stage, and there are few ties in relation to the number of different event times.

Assume we want to design a MAMS trial with three experimental treatment arms and a control, using triangular boundaries. One interim analysis is to be conducted upon observing $e_0^{(1)}$ events in the control arm, set to half of the total number of events in that arm. Our interesting effect size is a HR of 1.5, corresponding to a log-HR of 0.405, and the uninteresting effect size is a HR of 1.1 i.e., a log-HR of 0.095.

We can calculate the boundary values and sample sizes with the function `tite.mams`, which is another wrapper for `mams` with additional inputs `hr` and `hr0` for the interesting and uninteresting treatment effects, respectively, on the HR scale:

```
R> mtite <- tite.mams(hr=1.5, hr0=1.1, K=3, J=2, alpha=0.05, power=0.9,
+    r=1:2, r0=1:2, ushape="triangular", lshape="triangular")
R> mtite
Design parameters for a 2 stage trial with 3 treatments

                                              Stage 1 Stage 2
Cumulative number of events per stage (control):      81     162
Cumulative number of events per stage (active):       81     162

Maximum total number of events:  648

             Stage 1 Stage 2
Upper bound:   2.330   2.197
Lower bound:   0.777   2.197
```

The sample size output here is given as the required number of events, which is obviously smaller than the required number of patients. We refer to **?** for guidance how to estimate the maximum total number of patients to be recruited.

5. Discussion

In this paper we have described the R package **MAMS** for designing multi-arm multi-stage clinical trials, as well as some of the underlying statistical methodology. We have shown how to design a study with the package's basic function `mams` or its step-down variant `stepdown.mams`, how to evaluate properties like the expected sample size and power of a design with `mams.sim`, how to incorporate unforeseen changes using `new.bounds` and `stepdown.update`, and how to use the functionality of **MAMS** for ordinal, binary, and time-to-event endpoints with `ordinal.mams` and `tite.mams`. In addition to studying the numerical R output, it is often instructive to assess and compare designs using graphics, which can be accomplished with the package's automated `plot` function for the boundaries.

All this provides a convenient toolkit for planning and adapting efficient and highly flexible trials: inclusion of multiple experimental treatment arms increases the chances of a success; interim analyses allow to eliminate futile treatments early on and to stop the trial as soon as efficacy of any treatment is established; spontaneous design modifications (e.g., due to a safety issue) or sample size reassessment are smoothly taken into account using the conditional error principle. The theoretical foundations have been around for a few years (**??**), but they will only make a real impact on how clinical trials are conducted if user-friendly software is available and accessible. Given the limited range of adaptive design software that has been published to date, we consider **MAMS** a big step forward. It is the first package in R for adaptive trials with more than one active treatment group and at the same time more than one interim analysis for various types of endpoints.

The functionality of **MAMS** is extendable beyond pre-defined methods; for example, the design functions allow the user to implement arbitrarily shaped boundaries, with the only restrictions that the lower boundary be non-decreasing and the upper one non-increasing. In practice, however, it is often advisable to apply established methods whose properties are well understood. Especially the triangular test of **?** is an appealing candidate: although it may raise the *maximum* sample size in comparison to, e.g., the O'Brien-Fleming design, it usually has a substantially lower *expected* sample size compared to other designs (**?**).

Acknowledgments

This work is independent research arising in part from Dr Jaki's Senior Research Fellowship (NIHR-SRF-2015-08-001) supported by the National Institute for Health Research. Funding for this work was also provided by the Medical Research Council (MR/J004979/1). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

We are grateful to the Associate Editor and two anonymous referees whose comments helped us improve both the manuscript and the software.

A. A step-down test

```
R> m2.all
Design parameters for a 2 stage trial with 3 treatments

                                                 Stage 1 Stage 2
Cumulative sample size  (control):                    76     152
Cumulative sample size per stage (treatment  1 ):     38      76
Cumulative sample size per stage (treatment  2 ):     38      76
Cumulative sample size per stage (treatment  3 ):     38      76

Maximum total sample size:  380


Intersection hypothesis H_{ 1 }:

                   Stage 1 Stage 2
Conditional error 0.0260000    0.05
Upper boundary    1.9400000    1.72
Lower boundary    0.7864987    1.72

Intersection hypothesis H_{ 2 }:

                   Stage 1 Stage 2
Conditional error 0.0260000    0.05
Upper boundary    1.9400000    1.72
Lower boundary    0.7864987    1.72

Intersection hypothesis H_{ 1 2 }:

                   Stage 1 Stage 2
Conditional error 0.0260000    0.05
Upper boundary    2.2100000    2.06
Lower boundary    0.7864987    2.06

Intersection hypothesis H_{ 3 }:

                   Stage 1 Stage 2
Conditional error 0.0260000    0.05
Upper boundary    1.9400000    1.72
Lower boundary    0.7864987    1.72

Intersection hypothesis H_{ 1 3 }:

                   Stage 1 Stage 2
Conditional error 0.0260000    0.05
```

```
Upper boundary    2.2100000    2.06
Lower boundary    0.7864987    2.06

Intersection hypothesis H_{ 2 3 }:

                    Stage 1 Stage 2
Conditional error 0.0260000    0.05
Upper boundary    2.2100000    2.06
Lower boundary    0.7864987    2.06

Intersection hypothesis H_{ 1 2 3 }:

                    Stage 1 Stage 2
Conditional error 0.0260000    0.05
Upper boundary    2.3600000    2.22
Lower boundary    0.7864987    2.22
R> m2.best
Design parameters for a 2 stage trial with 3 treatments

                                                Stage 1 Stage 2
Cumulative sample size  (control):                   76     152
Cumulative sample size per stage (treatment  1 ):    38      76
Cumulative sample size per stage (treatment  2 ):    38      76
Cumulative sample size per stage (treatment  3 ):    38      76

Maximum total sample size:  380


Intersection hypothesis H_{ 1 }:

                    Stage 1 Stage 2
Conditional error 0.0260000    0.05
Upper boundary    1.9400000    1.71
Lower boundary    0.7864987    1.71

Intersection hypothesis H_{ 2 }:

                    Stage 1 Stage 2
Conditional error 0.0260000    0.05
Upper boundary    1.9400000    1.71
Lower boundary    0.7864987    1.71

Intersection hypothesis H_{ 1 2 }:

                    Stage 1 Stage 2
Conditional error 0.0260000    0.05
```

```
Upper boundary    2.2100000    2.02
Lower boundary    0.7864987    2.02

Intersection hypothesis H_{ 3 }:

                    Stage 1 Stage 2
Conditional error 0.0260000    0.05
Upper boundary    1.9400000    1.71
Lower boundary    0.7864987    1.71

Intersection hypothesis H_{ 1 3 }:

                    Stage 1 Stage 2
Conditional error 0.0260000    0.05
Upper boundary    2.2100000    2.02
Lower boundary    0.7864987    2.02

Intersection hypothesis H_{ 2 3 }:

                    Stage 1 Stage 2
Conditional error 0.0260000    0.05
Upper boundary    2.2100000    2.02
Lower boundary    0.7864987    2.02

Intersection hypothesis H_{ 1 2 3 }:

                    Stage 1 Stage 2
Conditional error 0.0260000    0.05
Upper boundary    2.3600000    2.17
Lower boundary    0.7864987    2.17
```

B. A design with unexpected modifications

```
R> m2.update
Design parameters for a 2 stage trial with 3 treatments

                                                Stage 1 Stage 2
Cumulative sample size  (control):                   75     228
Cumulative sample size per stage (treatment  1 ):    40     114
Cumulative sample size per stage (treatment  2 ):    35      35
Cumulative sample size per stage (treatment  3 ):    41     114

Maximum total sample size:  491
```

```
Intersection hypothesis H_{ 1 }:

                    Stage 1    Stage 2
Conditional error 0.0000000 0.08843773
Upper boundary    1.9400000 1.73000000
Lower boundary    0.7864987 1.73000000

Intersection hypothesis H_{ 2 }:

                    Stage 1    Stage 2
Conditional error 0.0000000 0.06896401
Upper boundary    1.9400000 1.71000000
Lower boundary    0.7864987 1.71000000

Intersection hypothesis H_{ 1 2 }:

                    Stage 1    Stage 2
Conditional error 0.0000000 0.05613374
Upper boundary    2.2100000 1.92000000
Lower boundary    0.7864987 1.92000000

Intersection hypothesis H_{ 3 }:

                    Stage 1   Stage 2
Conditional error 0.0000000 0.0576666
Upper boundary    1.9400000 1.7900000
Lower boundary    0.7864987 1.7900000

Intersection hypothesis H_{ 1 3 }:

                    Stage 1   Stage 2
Conditional error 0.0000000 0.0510653
Upper boundary    2.2100000 2.1400000
Lower boundary    0.7864987 2.1400000

Intersection hypothesis H_{ 2 3 }:

                    Stage 1    Stage 2
Conditional error 0.0000000 0.04326494
Upper boundary    2.2100000 1.90000000
Lower boundary    0.7864987 1.90000000

Intersection hypothesis H_{ 1 2 3 }:

                    Stage 1    Stage 2
Conditional error 0.0000000 0.04112744
```

```
Upper boundary    2.3600000 2.22000000
Lower boundary    0.7864987 2.22000000
```

Affiliation:

Thomas Jaki
Medical and Pharmaceutical Statistics Research Unit
Department of Mathematics and Statistics
Lancaster University
Lancaster, LA1 4YF, United Kingdom
E-mail: jaki.thomas@gmail.com

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd