1 Installation

1.1 Download/Install R

Go to https://www.r-project.org/ and follow the download and installation instructions provided.

1.2 (Recommended) Download/Install R Studio

Go to https://www.rstudio.com/products/rstudio/download/ and follow the download and installation instructions provided.

1.3 Download App Files from Github

Go to https://github.com/stefangraw/Allocation-Power-Optimizer and click "Clone or download" and choose Download ZIP. Once all files are downloaded, extract the main folder. Open the extracted folder and load the R script entitled "shiny-apo.R" with R Studio (or R if R Studio was not installed).

1.4 Install Required R Packages

Open R Studio (or the R console if R Studio was not installed) and execute the following line of code via the console (ignore >>, note that copy and pasting may produce incorrect quotation marks):

>> install.packages(c("shiny", "shinyjs", "parallel", "doParallel", "splines", "DescTools"))

Note steps 1.1 – 1.4 only need to be done once. Afterwards one can immediately proceed to 1.5.

1.5 Run Application

In the top-right corner of the scripting window of R Studio, click "Run App". If R Studio was not installed, enter the following command into the R console (ignore >>):

>> library(shiny); runApp('…/shiny-apo.R')

Where "…" denotes the path to the file "shiny-apo.R".

A new window should now appear entitled "Allocation & Power Optimizer". For visibility we suggest maximizing the window.

2 Explanation of Inputs

2.1 "Total sample size" tab

2.1.1 Basic Inputs

2.1.1.1 Critical value α (one-sided)

This is the type I error rate desired, or the probability of failing to reject the null hypothesis when it is true. A one-sided test is assumed. Typical values of α are 0.01, 0.05, and 0.10.

2.1.1.2 Inputs defining the sequence of total sample sizes

A sequence of the total number of subjects for which to determine the power for this study is given by the sequence $\{N_{start}, N_{start} + \Delta, N_{start} + 2\Delta, \dots, N_{end}\}$, where N_{start} is the user-defined starting N (see 2.1.1.2.1), Δ is the user-defined step size (see 2.1.1.2.3) and N_{end} is the user-defined ending N (see 2.1.1.2.2).

2.1.1.2.1 Start point of total sample size

This denotes the minimal sample size to be considered when determining power. Formally, defines N_{start} (see 2.1.1.2).

2.1.1.2.2 End point of total sample size

This denotes the upper bound of the maximal sample size to be considered when determining power. If a value is chosen that is not in the sequence defined by N_{start} and Δ (see 2.1.1.2), then N_{end} is the largest value of $N_{start} + \mathbb{N}\Delta$ which is less than the user-defined value ($\mathbb{N}$ is any non-negative integer). Formally, sets the upper bound for N_{end} (see 2.1.1.2).

2.1.1.2.3 Step size

This denotes the size of the increments between any two total sample sizes considered in the sequences of total sample sizes. Formally, defines Δ (see 2.1.1.2).

2.1.1.3 Effect size as

This determines the effect size of the treatment arms. Arm 1 is always the control arm and has effect size of 0. If either "Best arm" or "Linear trend" is selected, then the user will need to also provide the number of arms (see 2.1.1.4) and the standardized effect size (see 2.1.1.5). Additionally, if either "Best arm" or "Linear trend" are selected, the effect sizes for each arm $2, \ldots, k-1$ under the alternative hypothesis H_1, are determined as follows: "Best arm" sets each arm's effect to 0 besides arm k, which has effect size equal to the selected standardized effect size (see 2.1.1.4). "Linear trend" has the effect sizes grow linearly from the control (arm 1, effect of 0) to arm k (the best arm in this scenario, with effect of the selected standardized effect size). The linear function for calculating the effect of arm i is $d(i-1)/(k-1)$ where d is the selected standardized effect size. Typically, the linear trend scenario will have higher overall power but will be less likely to choose the "best" treatment arm.

If "Custom" is selected, then the user will need to enter each non-control arm's standardized effect size manually (see 2.1.1.6). Note that it is possible to use "Custom" to recreate "Linear trend" and/or "Best arm" schemes.

2.1.1.4 Number of arms

The number of treatment arms k (including control) for the trial. The first arm is always assumed to be the control arm. For example, if you have 5 treatment arms and 1 control arm, you have a total of 6 arms and would input 6 here. Only necessary if not using "Custom" Effect size as (see 2.1.1.3).

2.1.1.5 Standardized effect size

The standardized effect size of arm k, relative to the control arm's standardized effect size of 0. This is used only when "Best arm" or "Linear trend" is selected for Effect size as (see 2.1.1.3). Mathematically, this is the number of standard deviations by which the best treatment outperforms the control arm, on average. This is the effect size assuming the data has been standardized to have a sample variance of 1. Typically, this value comes from pilot data, or what the researcher feels is the "minimum improvement vs. the control that has meaning". For more information, the Wikipedia article on effect sizes provides a good overview of the topic https://en.wikipedia.org/wiki/Effect_size, as do many introductory statistical textbooks.

2.1.1.6 H_1 treatment means (comma separated)

This is only available when using "Custom" Effect size as (see 2.1.1.3). This allows the user to specify the standardized effect size of each treatment arm, relative to the control arm's standardized effect size of 0. The control arm is always assumed to be 0 and implicitly defined, so does not need to be entered in this field. One standardized effect size should be entered for each non-control treatment arm, separated by commas. For example, if one desires to recreate a linear trend for 4 total arms (3 treatment arms and 1 control arm), where the best arm has standardized effect size of 0.9, one would enter "0.3,0.6,0.9" (without the quotes) in this field.

Note it is possible to enter effect sizes of 0 or even negative.

2.1.1.7 Number of cores/threads

This determines the number of simultaneous threads for parallel processing. More cores result in faster computation time. Using more cores than your computer has available should default to the maximum number of cores but may cause it to crash. This defaults to half of the maximum number of cores available to the computer running the app.

2.1.1.8 Advanced settings

When checked, makes the Advanced settings available (see section 2.1.2).

2.1.1.9 Go!

After setting inputs as desired, this runs the app to determine the operating characteristics of the design and maximize the power (see section 3). While the app is running a screen indicating that the app is "Working…" is displayed.

2.1.2 Advanced Inputs

To view/change these, the "Advanced settings" box must be clicked (see 2.1.1.8).

2.1.2.1 Futility Stage I: δ and ϵ

Allows for more advanced early stopping rules. Both δ and ϵ default to zero, meaning that if unchanged then the only early stopping rule will be if the arm chosen at the end of Stage 1 is the control arm. Input δ defines the minimum number of standard deviations required for success between the arm chosen at the end of Stage 1 and the control arm. Input ϵ determines the probability required to exceeding the threshold set by δ at the end of Stage 1 (otherwise stop early for futility). If δ is zero, then ϵ defines the probability required for the winner arm to be superior to the control arm at the end of Stage 1 (otherwise stop early for futility). While δ can be defined as negative, we urge caution in interpretation of the results when doing so.

2.1.2.2 Number of n_1 and n_2 combinations per N

The number of combinations of Stage 1 and Stage 2 allocations to consider when attempting to optimize power for each total sample size N in the user-defined sequence of total sample sizes (see 2.1.1.2). This can be fairly coarse (defaults to 20) and still produce good results. For m checked sample sizes, and for each N total subjects, k treatment arms, and n_1 patients allocated to each arm in Stage 1, the choices of n_1 are chosen such that there are m equally spaced points between $n_1 = 2$ and $n_1 = \frac{N}{k} - 2$.

2.1.2.3 Post sample size

The number of draws to make from posterior distributions. Higher values of this will result in more accurate MCMC conclusions, but at the cost of computation time. This directly affects the accuracy of all operating characteristics.

2.1.2.4 Simulation runs H_0

The number of times to simulate under the null hypothesis H_0: each arm has 0 effect, meaning each treatment is equivalent to the control arm. This directly effects the accuracy of calculating the threshold τ for controlling the specified type I error rate α (see 2.1.1.1) as well as the width of the interval associated with α. Higher values of this will result in more accurate estimations of τ and α at the cost of computation time.

2.1.2.5 Simulation runs H_1

The number of times to simulate under the alternative hypothesis H_1, which is determined by inputs "Standardized effect size" and "Effect size as" (see 2.1.1.3 – 2.1.1.6). This directly effects the accuracy of calculating the statistical power of the trial (the probability of trial success given H_1 is true) and its associated interval. Higher values of this will result in more accurate power estimations at the cost of computation time.

2.1.2.6 Random seed

Allows the user to use a randomly generated seed or input a seed, for reproducibility of results. If a random seed is chosen, the value will be output in the Log of Inputs (see 3.2.5).

2.2 "Stage allocation optimization" tab

2.2.1 Basic Inputs

2.2.1.1 Critical value α (one-sided)

This is the type I error rate desired, or the probability of failing to reject the null hypothesis when it is true. A one-sided test is assumed. Typical values of α are 0.01, 0.05, and 0.10.

2.2.1.2 Total sample size

The total number of subjects available (or estimated to be available) for this study. Higher values result in more power, but may not be reasonably attainable.

2.2.1.3 Effect size as

This determines the effect size of the treatment arms. Arm 1 is always the control arm and has effect size of 0. If either "Best arm" or "Linear trend" is selected, then the user will need to also provide the number of arms (see 2.2.1.4) and the standardized effect size (see 2.2.1.5). Additionally, if either "Best arm" or "Linear trend" are selected, the effect sizes for each arm $2, \ldots, k-1$ under the alternative hypothesis H_1, are determined as follows: "Best arm" sets each arm's effect to 0 besides arm k, which has effect size equal to the selected standardized effect size (see 2.2.1.4). "Linear trend" has the effect sizes grow linearly from the control (effect of 0) to arm k (the best arm in this scenario, with effect of the selected standardized effect size). The linear function for calculating the effect of arm i is $d(i-1)/(k-1)$ where d is the selected standardized effect size. Typically, the linear trend scenario will have higher overall power but will be less likely to choose the "best" treatment arm.

If “Custom” is selected, then the user will need to enter each non-control arm’s standardized effect size manually (see 2.2.1.6). Note that it is possible to use “Custom” to recreate “Linear trend” and/or “Best arm” schemes.

2.2.1.4 Number of arms

The number of treatment arms k (including control) for the trial. The first arm is always assumed to be the control arm. For example, if you have 5 treatment arms and 1 control arm, you have a total of 6 arms and would input 6 here. Only necessary if not using “Custom” Effect size as (see 2.2.1.3).

2.2.1.5 Standardized effect size

The standardized effect size of arm k, relative to the control arm’s standardized effect size of 0. This is used only when “Best arm” or “Linear trend” is selected for Effect size as (see 2.2.1.3). Mathematically, this is the number of standard deviations by which the best treatment outperforms the control arm, on average. This is the effect size assuming the data has been standardized to have a sample variance of 1. Typically, this value comes from pilot data, or what the researcher feels is the “minimum improvement vs the control that has meaning”. For more information, the Wikipedia article on effect sizes provides a good overview of the topic https://en.wikipedia.org/wiki/Effect_size, as do many introductory statistical textbooks.

2.2.1.6 H_1 treatment means (comma separated)

This is only available when using “Custom” Effect size as (see 2.2.1.3). This allows the user to specify the standardized effect size of each treatment arm, relative to the control arm’s standardized effect size of 0. The control arm is always assumed to be 0 and implicitly defined, so does not need to be entered in this field. One standardized effect size should be entered for each non-control treatment arm, separated by commas. For example, if one desires to recreate a linear trend for 4 total arms (3 treatment arms and 1 control arm), where the best arm has standardized effect size of 0.9, one would enter “0.3,0.6,0.9” (without the quotes) in this field.

Note it is possible to enter effect sizes of 0 or even negative.

2.2.1.7 Number of cores/threads

This determines the number of simultaneous threads for parallel processing. More cores result in faster computation time. Using more cores than your computer has available should default to the maximum number of cores, but may cause it to crash. This defaults to half of the maximum number of cores available to the computer running the app.

2.2.1.8 Advanced settings

When checked, makes the Advanced settings available (see section 2.2.2).

2.2.1.9 Go!

After setting inputs as desired, this runs the app to determine the operating characteristics of the design and maximize the power (see section 3). While the app is running a screen indicating that the app is “Working…” is displayed.

2.2.1.10 Show table

When checked, shows the estimated power, threshold, probability of early stopping, probability of choosing best treatment arm, and total sample size for various Stage 1 sample sizes. Note that this

information is only available after the app has been run via the Go! Button (section 2.2.1.8). Also see 3.2.6.

2.2.2 Advanced Inputs

To view/change these, the "Advanced settings" box must be clicked (see 2.2.1.8).

2.2.2.1 Futility Stage I: δ and ϵ

Allows for more advanced early stopping rules. Both δ and ϵ default to zero, meaning that if unchanged then the only early stopping rule will be if the arm chosen at the end of Stage 1 is the control arm. Input δ defines the minimum number of standard deviations required for success between the arm chosen at the end of Stage 1 and the control arm. Input ϵ determines the probability required to exceeding the threshold set by δ at the end of Stage 1 (otherwise stop early for futility). If δ is zero, then ϵ defines the probability required for the winner arm to be superior to the control arm at the end of Stage 1 (otherwise stop early for futility). While δ can be defined as negative, we urge caution in interpretation of the results when doing so.

2.2.2.2 Number of n_1 and n_2 combinations

The number of combinations of Stage 1 and Stage 2 allocations to consider when attempting to optimize power. This can be fairly coarse (defaults to 20) and still produce good results. For m checked sample sizes, and for N total subjects, k treatment arms, and n_1 patients allocated to each arm in Stage 1, the choices of n_1 are chosen such that there are m equally spaced points between $n_1 = 2$ and $n_1 = \frac{N}{k} - 2$.

2.2.2.3 Post sample size

The number of draws to make from posterior distributions. Higher values of this will result in more accurate MCMC conclusions, but at the cost of computation time. This directly affects the accuracy of all operating characteristics.

2.2.2.4 Simulation runs H_0

The number of times to simulate under the null hypothesis H_0: each arm has 0 effect, meaning each treatment is equivalent to the control arm. This directly effects the accuracy of calculating the threshold τ for controlling the specified type I error rate α (see 2.2.1.1) as well as the width of the interval associated with α. Higher values of this will result in more accurate estimations of τ and α at the cost of computation time.

2.2.2.5 Simulation runs H_1

The number of times to simulate under the alternative hypothesis H_1, which is determined by inputs "Standardized effect size" and "Effect size as" (see 2.2.1.3 – 2.2.1.6). This directly effects the accuracy of calculating the statistical power of the trial (the probability of trial success given H_1 is true) and its associated interval. Higher values of this will result in more accurate power estimations at the cost of computation time.

2.2.2.6 Confidence level for simulated α and power

Confidence level used to calculate intervals for simulated type I error rate α and power of adaptive DTL design, Bonferroni-adjusted t and Dunnett-adjusted t (see 3.2.4.2 and 3.2.4.3).

2.2.2.7 Simulation runs Dunnett/Bonferroni

Number of simulation runs used to estimate power and associated confidence intervals for Bonferroni-adjusted t and Dunnett-adjusted t tests (see 3.2.4.2 and 3.2.4.3)..

2.2.2.8 Random seed

Allows the user to use a randomly generated seed or input a seed, for reproducibility of results. If a random seed is chosen, the value will be output in the Log of inputs (see 3.2.5).

3 Explanation of Outputs

3.1 "Total sample size" tab

3.1.1 Graph: Overall power for best allocation

This figure shows the estimated power as a function of selected total sample sizes (see 2.1.1.2) using the best allocation. Additionally, a smoothed fit (blue curve) is provided and used to estimated power for intermediate total sample sizes displayed below (see 3.1.3).

3.1.2 Log of inputs

This is a log of inputs (see 2.1.1 and 2.1.2) which can be copy/pasted so that the same analyses may be re-run. Additional outputs include the minimal total sample size required for 80% and 90% power (if within selected total sample size range), the randomly selected or chosen seed is also outputted and the run time.

3.1.3 Table: Total N and Power

This table provides estimates of power for all total sample sizes N within the total sample size range (see 2.1.1.2) based on the smooth fit (blue curve) provided in the graph of overall power for best allocation (see 3.1.1).

3.2 "Stage allocation optimization" tab

3.2.1 Simulated Type I Error Rate and confidence interval

This displays the estimated simulated type I error rate α and associated confidence interval (see 2.2.2.6) for the trial at the threshold τ associated with the optimal n_1. The estimated type I error should be approximately equal to the value provided by the user (see 2.2.1.1). The width of the confidence interval (see 2.2.2.6), which is calculated by Clopper-Pearson method for proportion of successful simulations under H_0, is a measure of the simulation error of the app. If the interval is too wide, increasing "Simulation runs H_0" (see 2.2.2.4) will shrink this interval.

3.2.2 Graph: Stage 2 Rejection Threshold

This shows the estimated threshold τ used to control the type I error rate α for various choices of n_1 (the number of subjects allocated per arm in Stage 1). The smooth blue line is the smoothed fit (smoothed quartic polynomial splines and a single knot) and is used to compute the power. The number of points on this graph is directly controlled by the "Number of checked sample sizes" input (see 2.2.2.2). If the red line connecting the points is especially chaotic, we recommend increasing the advanced inputs "Post sample size" and "Simulation runs H_0" (see 2.2.2.3 and 2.2.2.4).

3.2.3 Graph: Overall Power

This shows the estimated statistical power for various choices of n_1 (the number of subjects allocated per arm in Stage 1). The smooth blue line is the smoothed fit (smoothed via cubic splines, number of knots dynamically chosen) and is used to compute the power. The number of points on this graph is directly controlled by the "Number of checked sample sizes" input (see 2.2.2.2). If the red line connecting the points is especially chaotic, we recommend increasing the advanced inputs "Post sample size" and "Simulation runs H_1" (see 2.2.2.3 and 2.2.2.5). The vertical dashed line represents the choice of n_1 which maximizes the statistical power.

3.2.4 Table of Power for Adaptive, Bonferroni-adjusted, and Best arm known Designs

3.2.4.1 Adaptive DTL design

This design is the adaptive Bayesian two-stage drop-the-losers design. This row shows the power under the optimal choice of n_1 (see 3.2.3) and associated confidence interval (see 2.2.2.6), the choice of n_1 and n_2 that give this optimal power, the total subjects actually used (may be slightly less than what is inputted as Total sample size (see 2.2.1.3) due to rounding), and the associated τ which is the threshold for trial success that controls the type I error rate specified by the critical value α (see 2.2.1.1 and 3.2.1). Additionally, shown is the probability of early stop due to futility, which might depend on some advanced options (see 2.2.2.1), and the probability of successfully completing the trial with the best treatment arm being chosen as the winner in Stage 1 (defined as the treatment arm with largest standardized effect size, see 2.2.1.3-2.2.1.6. If multiple arms would be considered the best treatment arm, a successful trial that used any of them as the winner counts toward success with the best arm). The confidence interval associated with the power is calculated utilizing the Clopper-Pearson method for proportion of successful simulations under H_1, and if it is too wide can be reduced by increasing "Simulation runs H_1" (see 2.2.2.5).

3.2.4.2 Bonferroni adjusted t

In this single stage design, $k-1$ t-tests are performed between the control arm and the $k-1$ treatment arms using the pooled variance of all arms and allocating equal amounts of observations to each arm. Tests are adjusted for multiple testing using Bonferroni's method. The trial succeeds and $H_0$is rejected if the smallest p-value is significant. Power for this approach is estimated via simulation with precision determined by 2.2.2.7. Exact confidence intervals of the power are provided via the Clopper-Pearson method.

3.2.4.3 Dunnett adjusted t

In this single stage design, the $k-1$ treatment arms are compared to the control using Dunnett's method (Dunnett 1955). Power is estimated via simulation and by using the "DescTools" R package. Dunnett's method is more optimal for this type of testing than the Bonferroni method (see 3.2.4.2). Exact confidence intervals of the power are provided via the Clopper-Pearson method.

3.2.4.4 If best arm known

This design assumes the best arm (typically arm k unless using "Custom" Effect size as, see 2.2.1.3) is known. In this case, the best design is to assign $N/2$ patients to the best arm and control, and then perform a t-test. The power of such a test for the inputs given is outputted here, and represents the

upper bound of the power. Note that as this is a single stage design, there is no potential to stop early for futility.

3.2.5 Log of inputs

This is a log of inputs (see 2.2.1 and 2.2.2) which can be copy/pasted so that the same analyses may be re-run. The randomly selected or chosen seed is also outputted, along with the run time.

3.2.6 Show table

After the app has run and outputs are displayed, clicking this box shows a table displaying, for various choices of n_1, the estimated power and its confidence interval (see 2.2.2.6), n_2, N, τ, probability of stopping early for futility, and probability of successful trial where the best treatment arm is selected as the winner in Stage 1. The best treatment arm is defined as the arm with the largest standardized effect size (see 2.2.1.3-2.2.1.6). If multiple arms would be considered the best treatment arm, a successful trial that used any of them as the winner counts toward success with the best arm. The table may be sorted by any of these columns.

Note: More stringent stopping rules allow for a smaller expected number of patients that will have to be enrolled into the trial at the cost of potentially decreased power. The expected number of samples at the conclusion of a trial can be calculated via $E(\text{sample size}) = n_1 k + 2(1-P)n_2$ where P is the probability of stopping early for futility.

4 Mathematical Notation and Definitions

4.1 α

The type I error rate (see 2.1.1.1 or 2.2.1.1).

4.2 k

The number of treatment arms (see 2.1.1.4 or 2.2.1.4).

4.3 N

The total number of subjects available, equal to $N = kn_1 + 2n_2$ (see 2.1.1.2 or 2.2.1.2).

4.4 n_1 and n_2

The number of patients assigned to each arm in Stage 1 and Stage 2 of the design, respectively.

4.5 d

The standardized effect size (see 2.1.1.5 or 2.2.1.5). Equivalent to Cohen's d.

4.6 τ

The threshold for trial success, typically chosen to preserve the type I error rate α. In other words, the trial succeeds only if at the end of Stage 2 it is true that the probability that the best arm is better than the control arm exceeds τ.