interAdapt – An Interactive Tool for Designing and Evaluating Randomized Trials with Adaptive Enrollment Criteria

Aaron Fisher, Michael Rosenblum and Harris Jaffee

Abstract

We consider the problem of designing a randomized trial when there is prior evidence that the experimental treatment may be more effective for certain groups of patients, such as those with a certain biomarker or risk score at baseline. Randomized trial designs have been proposed that dynamically adapt enrollment criteria to target precisely those benefiting from treatment. Such trial designs aim to learn if the treatment benefits the overall population, only a certain subpopulation, or neither. We introduce the interAdapt software tool, which provides a user friendly interface for constructing and evaluating certain adaptive trial designs. These designs are automatically compared to standard (non-adaptive) designs in terms of the following performance criteria: power, sample size, and trial duration. Unlike existing software, interAdapt is open-source and cross-platform, and is the first to implement the group sequential, adaptive enrichment designs of [6].

1. Introduction

Group sequential, randomized trial designs involve rules for early stopping based on analyses of accrued data. Such early stopping could occur if there is strong evidence early in the trial of benefits or harms of the new treatment being studied. Adaptive enrichment designs involve rules for changing enrollment criteria, based on data accrued during a trial. We consider designs that incorporate features of both group sequential and adaptive enrichment designs. Such designs have rules for early stopping and for modifying enrollment criteria, based on accrued data in an ongoing trial. A class of such designs is presented by [6]. We focus on these designs, which we refer to simply as "adaptive designs." We compare the performance of these designs to "standard designs," defined to be group sequential designs where the enrollment criteria cannot be changed during the trial (but the trial may be stopped early).

To demonstrate our designs and software, we consider the problem of planning a Phase III trial for a new surgical treatment of stroke, which is considered by (Rosenblum et al. 2013)[6]. The new treatment is known as Minimally-Invasive Surgery Plus rt-PA for Intracerebral Hemorrhage (MISTIE), and is described in more detail in (Morgan 2008)[5]. There was more prior data on the efficacy of the treatment for patients with little or no intraventricular hemorrhage (IVH) at baseline (referred to as small IVH patients) than for patients with large IVH volume. The goal of the trial was to determine whether MISTIE is effective for the combined population of those with small or large IVH, and also for the small IVH population. A standard trial design, e.g., one enrolling the combined population throughout the trial, or enrolling only small IVH participants throughout the trial, may be inefficient at simultaneously answering these questions. An alternative is to use an adaptive trial design, which would first recruit from the combined population, and then decide whether to restrict enrollment based on results from interim analyses. Though we focus on this stroke trial application throughout, our software tool can be applied in many disease areas.

In Section 2, we formally define the hypothesis testing problem to be addressed different trial designs. In Section 3, we compare our software to the most similar, currently available commercial software, AptivSolutions ADDPLAN PE (Patient Enrichment). In Section 4, we describe how to install interAdapt on a personal computer as an R package, and how to access it online through a web browser. Section 5 describes the inputs available when using interAdapt, and discusses the interpretation of the application's output. In Section 6, we present an example demonstrating how an adaptive design is created and analyzed with interAdapt.

2. Problem Description

Consider the case where we have two subpopulations, referred to as subpopulation 1 and subpopulation 2. These must be specified before the trial starts, and be defined in terms of participant attributes measured at baseline (e.g., having a high initial severity of disease or a certain biomarker value). We focus on situations where there is suggestive, prior evidence that the treatment may be more likely to benefit subpopulation 1. Let π_1 and π_2 denote the proportion of patients in subpopulations 1 and 2, respectively.

Both the adaptive and standard designs discussed here consist of ongoing enrollment, and include predetermined rules for stopping the trial early based on interim analyses of currently enrolled patients. We discretize each trial into K stages, and say that the k^{th} stage will be completed once a pre-specified number of additional patients (n_k) have been enrolled. We allow the user to separately set the number of patients enrolled in the standard and adaptive designs, so these designs need not nessecarily have the same number of patients enrolled at each stage. In stages when both subpopulations are being recruited, we assume that $\pi_1 n_k$ of the patients recruited in stage k are from subpopulation 1, and $\pi_2 n_k$ are from subpopulation 2. An interim analysis is done at the end of each stage, which may lead us to stop the trial early if there is either strong evidence of treatment efficacy, or strong evidence of treatment futility.

Let $Y_{i,k}$ be the a binary outcome variable for the i^{th} patient recruited in stage k, where $Y_{i,k} = 1$ indicates a successful outcome. Let $T_{i,k}$ be an indicator of the event that the i^{th} patient recruited in stage k is assigned to the treatment. interAdapt assumes that the probability of being assigned to treatment is .5.

For subpopulation 1, denote the probability of a successful outcome under treatment as p_{1t}, and the probability of a successful outcome under control as p_{1c}. Similarly for population 2, let p_{2t} denote the probability of a success under treatment, and p_{2c} denote the probability of a success under control. We define the average treatment effect for a given population as difference in the probability of a successful outcome between the treatment and control groups.

In the remained of this section we give an overview of the relevant concepts and notation needed to understand and use interAdapt. A more detailed discussion of the theoretical context, and of the parameter calculation procedure, can be found in (Rosenblum et al. 2013)[6].

Hypotheses

We focus on testing the null hypothesis of no treatment effect in subpopulation 1, and the null hypothesis of no treatment effect in the combined population. The two hypotheses are defined respectively as

- H_{01}: $p_{1t} - p_{1c} \leq 0$
- H_{0C}: $\pi_1(p_{1t} - p_{1c}) + \pi_2(p_{2t} - p_{2c}) \leq 0$

interAdapt generates decision rules for an adaptive design that are able to test both of these hypotheses. The application compares the properties of the adaptive design to the properties of a standard design testing only H_{0C}, and to the properties of a standard design testing only H_{01}. In this paper, we refer to the adaptive design as AD, and refer to the two standard designs as SC and SS respectively. All three trials contain K stages, and the decision to entirely stop the trial early can be made at the end of any stage. Again, the trials differ in that SC and SS never change their enrollment criteria, while AD may switch to enroll only patients from subpopulation 1.

Note that the designs discussed here are not the same as the standard designs discussed in section 6.1 of (Rosenblum et al. 2013)[6], which test both hypothesis simultaneously. Implementing standard designs such as those discussed in (Rosenblum et al. 2013)[6] into the software is an area of future research.

Whenever any of the trials AD, SC or SS is stopped early, there will be some patients who have been enrolled but who's outcomes have not yet been measured. These patients are sometimes referred to as "overrunning" or "pipeline" patients. interAdapt currently discards measurements from these overruning patients in the final analysis. Incorporating these measurements is a goal for future work.

Test Statistics

We calculate three z-scores at the end of each stage, one for the treatment effect in the combined population, and one for the treatment effect in each of the subpopulations.

Denote $Z_{C,k}$ as the standardized Z-score for the estimated treatment effect in the combined population, which is based on the data from all patients with outcomes recorded by the end of stage k. When we assume an equal probability of being randomized to either treatment or control, the test statistic $Z_{C,k}$ takes the following form:

$$Z_{C,k} = \left[\frac{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} Y_{i,k'} T_{i,k'}}{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} T_{i,k}} - \frac{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} Y_{i,k'} (1 - T_{i,k'})}{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} (1 - T_{i,k})} \right] \times \left\{ \left(\frac{2}{\sum_{k'=1}^{k} n_{k'}} \right) \left(\sum_{s \in \{1,2\}} \pi_s [p_{sc}(1 - p_{sc}) + p_{st}(1 - p_{st})] \right) \right\}^{-1/2}$$

The term in square brackets is the difference in sample means between the treatment and control groups. The term in curly brackets is the variance of this difference in sample means.

Let $Z_{1,k}$ and $Z_{2,k}$ denote the analogous test statistics for the z-scores of the estimated treatment effect in subpopulations 1 and 2. The explicit form of $Z_{1,k}$ can be written as follows, where $A_{i,k}$ is an indicator that the i^{th} subject recruited in stage k is a member of subpopulation 1:

$$Z_{1,k} = \left[\frac{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} Y_{i,k'} T_{i,k'} A_{i,k'}}{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} T_{i,k} A_{i,k'}} - \frac{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} Y_{i,k'} (1 - T_{i,k'}) A_{i,k'}}{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} (1 - T_{i,k}) A_{i,k'}} \right] \times \left\{ \left(\frac{2}{\sum_{k'=1}^{k} \sum_{i=1}^{n_{k'}} A_{i,k'}} \right) (\pi_1 [p_{1c}(1 - p_{1c}) + p_{1t}(1 - p_{1t})]) \right\}^{-1/2}$$

We can write $Z_{2,k}$ in an analagous form using an indicator of memebership in population 2.

The decision rules defined later on in this section for testing H_{0C} and H_{01} will consist of critical boundaries for $(Z_{C,1}, Z_{C,2}, ...Z_{C,K})$, $(Z_{1,1}, Z_{1,2}, ...Z_{1,K})$, and $(Z_{2,1}, Z_{2,2}, ...Z_{2,K})$. To calculate the family-wise Type I error of any given set of decision rules, we make use of the multivariate distribution of $(Z_{C,1}, Z_{C,2}, ...Z_{C,K}, Z_{1,1}, Z_{1,2}, ...Z_{1,K})$, which can be shown to be normal with a known covariance matrix (Jennison and Turnbull, 1999, Chapter 3)[4].

Family-wise Type I Error

In the context of our hypotheses, the Family-wise Type I error rate refers to the combined rate of false positives from testing either H_{0C} and H_{01}. We say that the Family-wise Type I error rate is controlled at level α when the probability of rejecting at least one true hypothesis is less than α, under all possible true underlying states of the world.

For all three designs, AD, SC, and SS, we require the same family-wise type I error rate, denoted by α. Since the two standard designs SS and SC each only test a single hypothesis, their family-wise Type I error rates are simply equal to the type I error rates of their respective hypothesis tests. A multiple hypothesis correction would have to be made in order to analyze a combination of the results of the two standard designs. We discuss the control of the family-wise Type I error rate for the AD design in the next section.

Decision Rules for Stopping the Trial Early

In the SC trail, our decision rules consist of efficacy and futility boundaries for H_{0C}. At each stage k, we calculate the test statistic $Z_{C,k}$. If $Z_{C,k}$ is above the efficacy boundary for stage k, we reject H_{0C} and end the trial. If the $Z_{C,k}$ is between the efficacy and futility boundaries for stage k, we make no conclusion and continue the trial. If $Z_{C,k}$ is below the futility boundary for stage k, we end the trial with the conclusion that we have failed to reject H_{0C}. interAdapt makes the simplification that the number of patients enrolled in each stage of SC is constant (n_{SC}), and allows the user to input this per-stage enrollment rate.

The efficacy boundaries for SC are set to be proportional to those described by Wang and Tsiatis (1987). This means that the efficacy boundary for the k^{th} stage is set to $e_{SC}\{(\sum_{k'=1}^{K} n_{k'})/n_k\}^{-\delta}$, where K is the total number of stages, δ is a constant in the range $[-.5, .5]$, and e_{SC} is the constant calibrated to ensure the desired family-wise Type I error rate. In order to calculate e_{SC}, we make use of the fact

that the random vector of test statistics $(Z_{C,1}, Z_{C,2}, \ldots Z_{C,K})$ follows a multivariate normal distribution with a known covariance structure (Jennison and Turnbull, 1999, Chapter 3)[4]. Under H_{0C} we assume this vector has mean zero. Using the "mvtnorm" package [2] in R to evaluate the multivariate normal distribution function, we find the proportionality constant e_{SC} such that the null probability of $Z_{C,k}$ exceeding $e_{SC}\{(\sum_{k'=1}^{K} n_{k'})/n_k\}^{-\delta}$ at any stage k is less than or equal to α.

In SC, as well as in SS and AD, we make use of non-binding futility constants that the study administrators can choose to ignore. All three designs are calibrated such that family-wise type I error rate is controlled at level α regardless of whether the futility boundaries are ignored. In calculating power however, we do assume that the futility boundaries are adhered to.

Futility boundaries for the first $K-1$ stages of SC are also proportional to $\{(\sum_{k'=1}^{K} n_{k'})/n_k\}^{-\delta}$, but with a different proportionality constant, denoted by f_{SC}. The constant f_{SC} is traditionally set to be negative, though this is not required. Since these futility boundaries are nonbinding, f_{SC} can be changed by the user without affecting the calculated Type I error rate. In the K^{th} stage of the trial, interAdapt sets the futility bound to be equal to the efficacy bound. This ensures that $Z_{C,K}$ eventually crosses either the efficacy bound or less futility bound, and that we are always able to make some kind of decision regarding H_{0C} by the time the trial has concluded.

The decision boundaries for $Z_{1,k}$ in the SS design are defined by exactly the same form. The efficacy boundary for the k^{th} stage is set equal to $e_{SS}\{(\sum_{k'=1}^{K} n_{k'})/n_k\}^{-\delta}$, where e_{SS} is the constant that ensures the appropriate type I error rate. The first $K-1$ futility boundaries for H_{01} are again set equal to $f_{SS}\{(\sum_{k'=1}^{K} n_{k'})/n_k\}^{-\delta}$, where f_{SS} is a constant that can be set by the user. The futility boundary in stage K is set equal to the final efficacy boundary in stage K. Again, the user can also specify the number of patients to enroll in each stage (n_{SS}). This per-stage enrollment rate is set to be constant across stages.

Decision boundaries for AD vary from those of the standard designs two ways. First, because AD simultaneously tests H_{0C} and H_{01} it must have two sets of decision boundaries rather than one. For the k^{th} stage of AD, let $u_{C,k}$ and $u_{1,k}$ denote the efficacy boundaries for H_{0C} and H_{01} respectively. The boundaries $u_{C,k}$ and $u_{1,k}$ are set equal to $e_{AD,C}\{(\sum_{k'=1}^{K} n_{k'})/n_k\}^{-\delta}$ and $e_{AD,1}\{(\sum_{k'=1}^{K} n_{k'})/n_k\}^{-\delta}$ respectively, where $e_{AD,C}$ and $e_{AD,1}$ are constants set such that the probability of rejecting either hypothesis under the global null hypothesis is zero. To correctly calibrate $e_{AD,C}$ and $e_{AD,1}$, interAdapt first chooses $e_{AD,C}$ such the probability of falsely rejecting H_{0C} is $a_c\alpha$, where a_c is a fraction between 0 and 1 that can be specified by the user. Then, conditional on $e_{AD,C}$, interAdapt finds the smallest constant $e_{AD,1}$ such that

$$P\left(Z_{C,k} > e_{AD,C}\left\{\frac{\sum_{k'=1}^{K} n_{k'}}{n_k}\right\}^{-\delta} \text{ or } Z_{1,k} > e_{AD,1}\left\{\frac{\sum_{k'=1}^{K} n_{k'}}{n_k}\right\}^{-\delta} \text{ for any } k\right) \leq \alpha$$

The futility boundaries for the AD design, $l_{1,k}$ and $l_{2,k}$, are defined relative to the test statistics $Z_{1,k}$ and $Z_{2,k}$. These boundaries are set equal to $f_{AD,C}\{(\sum_{k'=1}^{K} n_{k'})/n_k\}^{-\delta}$ and $f_{AD,S}\{(\sum_{k'=1}^{K} n_{k'})/n_k\}^{-\delta}$ respectively, where $f_{AD,C}$ and $f_{AD,S}$ can be set by the user.

The second way that the decision boundaries of AD differ from those of the standard designs is that we allow for more exibility in the futility boundaries. In each stage k, our adaptive design has the option of stopping enrollment in subpopulation 2, based on the treatment effect estimate $Z_{2,k}$. interAdapt also allows the user to specify a final stage for testing an effect in the total population, denoted by stage k^*. Regardless of the results at stage k^*, we always stop enrolling from subpopulation 2 at the end stage k^*, if we have not done so already. The futility boundaries $l_{2,k}$ are not defined for $k > k^*$.

For the AD design, the user can specify two stage specific sample sizes, one for stages when both populations are enrolled ($k \leq k^*$), and one for stages where only patients in subpopulation 1 are enrolled ($k > k^*$). We refer to these two sample sizes as n_1^* and n_k^* respectively.

As described in (Rosenblum et al. 2013)[6], our decision rules in AD consist of the following steps for each stage k:

1. (Assess Efficacy) If $Z_{C,k} > u_{C,k}$, reject H_{0C}. If $Z_{1,k} > u_{1,k}$, reject H_{01}. If either, or both null hypothesis are rejected, stop all enrollment and end the trial.

2. (Assess Futility of the entire trial) Else, if $Z_{1,k} \leq l_{1,k}$ or if this is the final stage of the trial, stop all enrollment and end the trial for futility, failing to reject either H_{0C} or H_{01}.

3. (Assess Futility for H_{0C}) Else, if $Z_{2,k} \leq l_{2,k}$, or if $k \leq k^*$, stop enrollment from subpopulation 2 in all future stages. In this case, the following steps must then be done:

   (a) If $Z_{1,k} > u_{1,k}$, reject H_{01} and stop all enrollment.

   (b) If $Z_{1,k} \leq l_{1,k}$ or if this is the final stage of the trial, conclude that we've fail to reject either H_{0C} or H_{01}, and stop all enrollment.

   (c) Else, we continue to patients from subpopulation 1, and re-evaluate steps (a)-(b) at the end of the next stage. If $k < k^*$ then $\pi_1 n_1^*$ patients should be enrolled in the next stage. If $k \geq k^*$, then n_k^* patients should be enrolled in the next stage.

4. (Continue Enrollment from Combined Population) Else, continue enrolling from both subpopulations and repeat step 1 at the end of the next stage.

The decision rules outputted by interAdapt represent the feature that enrollment of subpopulation 2 cannot continue after stage k^* by setting the futility boundary l_{2,k^*} equal infinity. This ensures that $Z_{2,k^*} < l_{2,k^*}$.

3. Related Software

The most comparable available software is AptivSolutions ADDPLAN PE (Patient Enrichment), an impressive, commercial software that implements certain types of adaptive enrichment designs. It has many features that our software does not have. Conversely, there are features of our software that ADDPLAN PE does not have. First, ADDPLAN PE does not implement the class of designs from (Rosenblum etal. 2013)[6]. Second, in ADDPLAN PE, the user must a priori designate a particular stage (e.g., stage 2) at which a change to enrollment may be made, even though there may be large a priori uncertainty as to when sufficient information will have accrued to make such a decision. In contrast, our software is more flexible, in that one can select designs (by setting k^* to the maximum number of stages) in which the decision to change enrollment criteria can be made at any stage.

interAdapt also has the benefits of being cross-platform and open-source, while ADDPLAN PE is commercial software that is only compatible with the Windows OS.

4. Running interAdapt

interAdapt is an interactive application built on the "Shiny" package for the R programming language (`http://www.r-project.org/`). The user interface is shown in the user's default web browser, while the back-end calculations are all done in R. Users can run interAdapt either by installing R and the interAdapt R package locally on their computer, or by simply using a web browser view interAdapt online. Both options are free and quick to set up. However, because online application will slow down noticeably when accessed by multiple users, we encourage heavy users to install interAdapt locally.

Running interAdapt Over the Web

interAdapt is currently hosted on the RStudio webserver, and can be accessed simply visiting the link below. `http://spark.rstudio.com/mrosenblum/interAdapt`

Running interAdapt Locally

To run interAdapt locally, one must first install the R programming language. R runs on both Windows & MacOS, with the most current versions available for download at (`http://www.r-project.org/`). After downloading and installing R, activating the R application will open an "R Console" window where typed commands are executed by R. interAdapt is available as a package for R, and can be installed by typing the lines below into the R Console, while connected to the Internet. The return key must be pressed after each line of code. The first and third lines will cause R to give feedback on the installation progress, which we do not show here.

```
install.packages('devtools')
library('devtools')
install_github(username='aaronjfisher',repo='interAdapt',subdir='r_package')
```

Once interAdapt has been installed, the application can be run without an internet connection by the opening the R Console and typing the code below.

```
library('interAdapt')
runInterAdapt()
```

5. User Interface

Inputs to interAdapt can be entered in the side panel on the left, with outputs are shown in the main panel on the right. The parameters in the input panel let the user describe known or assumed characteristics of their populations of interest, as well as their trial design parameters. Input parameters include the proportion of patients in each subpopulation, the patient recruitment rate in each subpopulation, and the desired Family-wise Type I error rate. The output section displays the decision boundaries and trial designs that will satisfy the requirements specified by the user. It also compares the performance of the three designs, AD, SC and SS. Performance is compared in terms of power, expected sample size, expected trial duration, and expected number of overrunning patients.

All tables generated by interAdapt can be downloaded as csv files by clicking on the download button beneath the table. Users can also download an automated report of the results by clicking the "Generate Report" button at the bottom of the output panel. This report is generated with the "knitr" package for R [7]. Citations in the report are created using the "knitcitations" package [1].

Inputs

Parameters in the input panel are organized into two sections, basic parameters and advanced parameters. To view the different sets of parameters, click the drop down menu titled "Show basic parameters."

Basic parameters can be entered using either "Batch mode" or "Interactive mode". In Batch mode, interAdapt will not analyze the entered parameters until the "Apply" button is pressed. This allows for several parameters to be changed at once without waiting for interAdapt to recalculate the results after each individual change. In Interactive mode, interAdapt will automatically recalculate the results after each change, allowing the user to quickly see the effect of changing one specific input parameter. Switching between Batch mode and Interactive mode can be done using the dropdown menu at the top of the Basic Parameters section. Interactive mode is not available when entering advanced parameters.

To save the current set of inputs, select the dropdown menu titled "Show basic parameters" and select "Show All Parameters and Save/Load Option". From here, you can save the current parameters as a csv file, or load a previously saved csv file of inputs. Regardless of whether interAdapt is being run online or locally, these saved csv files are always stored on the user's computer. You may also load a 3-column dataset into interAdapt in the form of a csv, where each row contains information about a patient in the trial. The first column must contain binary indicators of subpopulation, where 1 denotes subpopulation 1, and 2 denotes subpopulation 2. The second column must contain an indicator of the treatment arm (T_i), and the third column must contain the binary outcome measurement (Y_i). The first row of this dataset file is expected to be a header row of labels, rather than values for the first individual. From this dataset, interAdapt will calculate π_1, p_{1c}, p_{1t}, p_{2c}, and p_{2t}, and adjust the input sliders accordingly.

A detailed explanation of each input is given below.

Basic Parameters

- Subpopulation 1 proportion (π_1): The proportion of the population in subpopulation 1. This is the subpopulation in which we have prior evidence of a stronger treatment effect.
- Probability outcome = 1 under control, subpopulation 1 (p_{1c}): The probability of experiencing a successful outcome for control patients in subpopulation 1. This is used in estimating power and expected sample size of each design.
- Probability outcome = 1 under control, subpopulation 2 (p_{2c}): The probability of experiencing a successful outcome for control patients in subpopulation 2. This is used in estimating power and expected sample size of each design.

- Probability outcome = 1 under treatment for sub-population 1 (p_{1t}): The probability of experiencing a successful outcome for treated patients in subpopulation 1. Note that a specific effect size is not specified for subpopulation 2. Instead, interAdapt generates the relevant performance metrics for a range of several possible effect sizes in subpopulation 2. This range can be specified in the Advanced Parameters section.

- Per stage sample size, combined population (n_1^*): The number of patients enrolled in stages 1 through k^* of AD. Per stage enrollment for SC and SS can be entered in the advanced parameters section.

- Per stage sample size for stages where only sub-population 1 is enrolled (n_k^*): The number of patients required for each stage after stage k^*, in the AD design.

- Alpha (FWER) Requirement (α): The family-wise Type I error rate for all hypotheses in the trial. In AD, this is the probability of falsely rejecting either H_{0C} or H_{01}. In SC it is the probability of falsely rejecting H_{0C}. In SS it is the probability of falsely rejecting H_{01}.

- Proportion of Alpha allocated to H0C (a_C): To control the family-wise Type I error rate in the AD design, the test of H_{0C} is first calibrated to have a Type I error rate equal to $a_C\alpha$. The decision rules for H_{01} are then calibrated so that the overall family-wise Type I error rate is equal to α.

Advanced Parameters

- Delta (δ): This parameter defines the curvature of the efficacy and futility boundaries, which are all proportional to $\{(\sum_{k'=1}^{K} n_{k'})/n_k\}^{-\delta}$.

- Number of Iterations for simulation: Z-statistics are simulated generate the power, expected sample size, expected trial duration, and expected number of overrunning patients. Generally, about 10,000 simulations are needed for reliable results. It is our experience that a simulation with 10,000 iterations takes about 15 seconds on a modern personal computer.

- Time limit for simulation, in seconds: If the simulation time exceeds this threshold, calculations will stop and the user will get an error message saying that the application has "reached CPU time limit". To remove the error, either the number of iterations can be reduced, or the time limit for simulation can be extended. interAdapt does not allow for this time limit to exceed 90 seconds.

- Total number of stages (K): The total number of stages for all three designs.

- Participants enrolled per year from combined population: The number of patients that can be recruited per year in the combined population. This affects the estimated duration of the trials. The recruitment rates for subpopulations 1 and 2 are equal to the combined population recruitment rate multiplied by π_1 and π_2 respectively.

- Delay time from enrollment to primary outcome in years: The between when an individual is enrolled in the study, and when their final outcome is measured. This value affects the expected number of overruning patients, and the expected duration of the trial.

- Lower bound for treatment effect in sub-population 2: interAdapt simulates performance metrics under a range of treatment effect sizes for subpopulation 2. This sets the lower bound for this range.

- Upper bound for treatment effect in sub-population 2: interAdapt simulates performance metrics under a range of treatment effect sizes for subpopulation 2. This sets the upper bound for this range.

- Last stage sub-population 1 is enrolled under an adaptive design (k^*): In the adaptive design, we don't enroll any patients from subpopulation 2 after stage k^*.

- Per stage sample size for standard group sequential design enrolling combined pop (n_{SC}): The number of patients enrolled in each stage for SC.

- Per stage sample size for standard group sequential design enrolling only subpop. 1 (n_{SS}): The number of patients enrolled in each stage for SS.

- H_{0C} futility boundary proportionality constant for the adaptive design ($f_{AD,C}$): This is used to calculate the futility boundary for H_{0C} in the adaptive design, which is set to $f_{AD,C}\{(\sum_{k'=1}^{K} n_{k'})/n_k\}^{-\delta}$ in stage k.
- H_{01} futility boundary proportionality constant for the adaptive design ($f_{AD,S}$): This is used to calculate the futility boundary for H_{01} in the adaptive design, which is set to $f_{AD,S}\{(\sum_{k'=1}^{K} n_{k'})/n_k\}^{-\delta}$ in stage k.
- H_{0C} futility boundary proportionality constant for the standard design (f_{SC}): This is used to calculate the futility boundary for H_{0C} in SC, which is set to $f_{SC}\{(\sum_{k'=1}^{K} n_{k'})/n_k\}^{-\delta}$ in stage k.
- H_{01} futility boundary proportionality constant for the standard design (f_{SS}): This is used to calculate the futility boundary for H_{01} in SS, which is set to $f_{SS}\{(\sum_{k'=1}^{K} n_{k'})/n_k\}^{-\delta}$ in stage k.

Outputs

The output panel of the user interface is split into three sections, "About interAdapt", "Designs" output and "Performance" output. The "About interAdapt" section gives a brief introduction to the software, and a link to the full software documentation. The Designs output gives a road plan for how to conduct each of the three trials: FA, AD and SC. This includes the efficacy boundaries; user specified non-binding futility boundaries, and number of patients to recruit by the end of each stage. Performance output compares the three designs in terms of their power, expected sample size, expected duration, and expected number of overrunning patients. The radio buttons at the top of the output section can be used to switch between these three sections.

Designs Output

The designs output gives information on how to conduct each of the three trials. Tabs at the top of the page can be used to navigate between the results for each design. Each of the first three tabs each correspond with one of the designs, and the fourth tab shows all three designs side by side.

In the "Adaptive" tab, the table at the bottom of the page shows the required number of patients that must be recruited by the end of each stage.For each stage k, the table also gives efficacy boundaries for $Z_{1,k}$ and $Z_{C,k}$, and futility boundaries for $Z_{1,k}$ and $Z_{2,k}$. Because we always stop enrolling subpopulation 2 after stage k^*, futility boundaries for $Z_{2,k}$ in stage k^* and later stages are not given. For the same reason, efficacy boundaries for $Z_{C,k}$ are not given for stages $k > k^*$. A plot at the top of the page shows these efficacy and futility boundaries for $Z_{C,k}$, $Z_{1,k}$ and $Z_{2,k}$ over all stages of the trial.

The two tabs for the standard designs have a comprable layout. Note that the efficacy boundaries for SS and SC are identical. This is because the efficacy boundary depends only on the null distribution of z-statistics, which unaffected by the choice of study population. The two standard trials each pull from only one subpopulation, so their efficacy boundaries are the same.

The final tab combines the tables from the first three tabs, and omits plots of the decision boundaries.

Performance Output – Layout & Interpretation

interAdapt shows performance of each of the three designs in terms of four metrics: power, expected sample size, expected duration, and expected number of overrunning patients. These metrics all depend, among other things, on the true treatment effect in each subpopulation. A treatment effect for subpopulation 1 can be specified in the Basic Parameters section, and a range of values for the treatment effect in subpopulation 2 can be specified in the Advanced Parameters section. interAdapt will calculate performance metrics for the specified range of treatment effects, and generate charts of each metric plotted against the underlying treatment effect in subpopulation 2. These four plots can be accessed via the tabs at the top of the page. The table at bottom of the output section shows all four metrics side by side, with each column of the table denoting a different treatment effect in subpopulation 2.

When the true treatment is very strong, trials will tend to be able to detect the treatment effect more easily, and will be more likely to stop early for efficacy. This translates to an overall increase in power, a decrease in expected sample size, a decrease in expected trial duration, and an increase in the expected number of overrunning patients. The increase in overruning patients is due to the fact that it is precisely the process of stopping early that generates overrunning patients. Conversely, if the true underlying treatment effect is significantly harmful, the trials will be more likely to stop early for futility. This leads

to trials with smaller expected sample sizes, shorter expected durations, and larger expected numbers of overrunning patients. Trials will tend to last the longest when the treatment effect is positive, but not overwhelmingly strong. These patterns are reflected in the plots shown by interAdapt.

The power plot shows the power of AD to reject H_{0C}, to reject H_{01}, and to reject at least one of H_{0C} or H_{01}. As the standard design SC only tests H_{0C}, interAdapt only shows it's power to reject H_{0C}. Likewise, interAdapt only shows the power of SS to reject H_{01}. Note that the power of SC and AD to reject H_{0C} both increase as the treatment effect for subpopulation 2 increases. The power of AD to reject H_{01} decreases as the treatment effect in subpopulation 2 increases, but this is only because AD does not bother to test H_{01} after a treatment effect in the combined population is discovered.

In general, power of a trial can be increased by increasing the per-stage sample size (n_1^*, n_k^*, n_{SS} and n_{SC}), increasing the number of stages (K), lowering the futility boundaries (f_{SC}, f_{SS}, $f_{AD,C}$, or $f_{AD,S}$), or relaxing the required type I error rate (α).

The power of SS is constant with respect to the true treatment effect in subpopulation 2, as we'd expect since SS does not take any data from subpopulation 2. The expected sample size, expected duration, and the expected number of overrunning patients for SS are also constant with respect to the true treatment effect in subpopulation 2.

In the plot of expected sample size for each design, we see that trials tend to need to recruit more patients when the treatment effect is weak. For designs testing for an effect in the combined population, this means that the expected sample size will be highest when the weighted average treatment effect across subpopulations is weak. If the treatment effect is significantly positive in subpopulation 1, the highest possible expected sample size may come at a negative value for the true treatment effect in subpopulation 2. In general, lowering K or k^*, increasing the futility bounds (f_{SC}, f_{SS}, $f_{AD,C}$, or $f_{AD,S}$), or relaxing the required type I error rate (α), can all decrease the expected sample size.

The plot of expected trial duration for each design shows patterns very similar to those in the plot of expected sample size. A trial's duration is defined as the time until the last patient's outcome is measured. Like expected sample size, the expected duration can be decreased by lowering K or k^*, increasing the futility bounds (f_{SC}, f_{SS}, $f_{AD,C}$, or $f_{AD,S}$), or relaxing α. Increasing the recruitment rate, or decreasing the delay time to outcome measurement, can also shorten the expected duration of a trial.

The plot of the expected number of overrunning patients will generally show a minimum when the treatment effect is just on the cusp of significance, as this will require trials to gather more data. Whenever a trial gathers as much data as possible and reaches stage K, there will be no overrunning patients. When applicable, decreasing K or k^*, decreasing the futility bounds (f_{SC}, f_{SS}, $f_{AD,C}$, or $f_{AD,S}$), decreasing the delay time to outcome measurement, or lowering the recruitment rate can all decrease the expected number of overrunning patients.

6. Example of Entering Input and Interpreting Output

The default inputs to interAdapt come from the motivating example of the MISTIE Phase III trial. This section presents a summary of this trial, and of the design goals of the investigators, as described in (Rosenblum et al. 2013)[6]. The MISTIE trial studied a new surgical treatment for stroke, and measured patient's outcomes by their disability score on the modified Rankin Scale (mRS) 180 days after enrollment. A successful outcome was defined as a mRS score less than or equal to 3.

The Phase II trial for the MISTIE treatment had only enrolled patients with with little or no intraventricular hemorrhage (IVH). More specifically, patients had been categorized as "small IVH" if their IVH volume was less than 10ml, and did not require a catheter for intracranial pressure monitoring. Otherwise, pateints were classified as "large IVH." The Phase II trial only recruited small IVH patients, and yeiled a treatment effect estimate of 12.1% [95% CI: (-2.7%, 26.9%)]. The investigators thought that the treatment could also be effective in large IVH pateints, but no data had yet been collected to test this. Thus, we refer to the subpopulation of small IVH patients as subpopulation 1, as there was more prior evidence of treatment efficacy in this subpopulation.

The study designers were concerned with the calibrating power and alpha level of the Phase III trial under the following three scenarios:

(a) The average treatment effect is 12.5% for both small and large IVH pateints;

(b) The average treatment effect is 12.5% for small IVH patients, and zero large IVH patients;

(c) The treatment effect is zero both subpopulations.

In the context of these scenarios, the study coordinators had three goals:

(i) At least 80% power for testing H_{0C} in scenario (a);

(ii) At least 80% power for testing H_{01} in scenario (b);

(iii) A family-wise Type I error rate (α) of .025.

Prior research by (Hanley 2012)[3] indicated that the proportion of patients with small IVH (π_1) was .33, that the probability of a positive outcome under control was .25 for small IVH patients (p_{1c}), and that the probability of a positive outcome under control was .2 for large IVH patients (p_{2c}). If the true treatment effect in subpopulation 1 was 12.5% then the probability of a positive outcome under treatment for patients in subpopulation 1 (p_{1t}) would be approximately 12.5%+25%=37.5%.

Since the adaptive design AD tests H_{0C} as well as H_{01}, it must achieve all three goals (i)-(iii). The standard design SC need only achieve (i) and (iii), and the standard design SS need only achieve (ii) and (iii). Recall that interAdapt allows the user to specify a range of treatment values for subpopulation 2, and will display the power of the trial designs across this range. By default, interAdapt sets the range of values for the treatment affect in subpopulation 2 to [-.2, .2], letting the user see the power of all three designs under scenarios (a) and (b).

The remaining default input parameters come from the analysis section of (Rosenblum et al. 2013)[6]. Here, the authors first fixed $K = 5$ and $\delta = -.5$, and then searched for values of the remaining parameters that minimize the average expected sample size over scenarios (a)-(c) for the adaptive design, while still achieving goals (i)-(iii). They found a minimum average expected sample size at $k^* = 4$, $n_1^* = 150$, $n_k^* = 311$, and $f_{AD,C} = f_{AD,S} = 0$.

Now we turn to the output of interAdapt that results from the default parameters, and show that each of the three designs achieves its relevant goals. In the power plot, we see that AD has 80% power to reject H_{0C} in scenario (a), and 80% power to reject H_{01} in scenario (b). SC has 80% power to reject H_{0C} in scenario (a), and SS has 80% power to reject H_{01} in scenario (b). Although it is not shown, we know that the family-wise type I error rate is less than .025, as this was specified as an input to interAdapt.

Summary

We described the interAdapt application for designing and simulating trials with adaptive enrollment criteria. We provided an overview of the theoretical problem the application addresses, and gave an explanation of the application's inputs and outputs.

Acknowledgements

References

[1] Carl Boettiger. *knitcitations: Citations for knitr markdown files*, 2013. R package version 0.4-7.

[2] Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, and Torsten Hothorn. *mvtnorm: Multivariate Normal and t Distributions*, 2013. R package version 0.9-9996.

[3] Daniel Hanley. `http://braininjuryoutcomes.com/studies/mistie/entry/mistie/international-stroke-conference-2012-mistie-phase-2-results`, 2012.

[4] Christopher Jennison and Bruce W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC Press, 1999.

[5] T Morgan, M Zuccarello, R Narayan, P Keyl, K Lane, and D Hanley D. Preliminary findings of the minimally-invasive surgery plus rtpa for intracerebral hemorrhage evacuation (mistie) clinical trial. *Acta Neurochir Suppl.*, 105:147–51, 2008.

[6] Michael Rosenblum, Richard E. Thompson, Brandon Luber, and Daniel Hanley. Adaptive group sequential designs that balance the benefits and risks of expanding inclusion criteria. *Under Revision*, 2013.

[7] Yihui Xie. *knitr: A general-purpose package for dynamic report generation in R*, 2013. R package version 1.4.1.