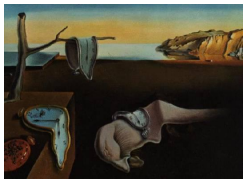


Journal of Statistical Software

MMMMMM YYYY, Volume VV, Code Snippet II. http://www.jstatsoft.org/

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

Aaron Fisher
Johns Hopkins University

Harris Jaffee
Johns Hopkins University

Michael Rosenblum
Johns Hopkins University

Abstract

The **interAdapt** R package is designed to be used by statisticians and clinical investigators to plan randomized trials. It can be used to determine if certain adaptive designs offer tangible benefits compared to standard designs, in the context of investigators' specific trial goals and constraints. Specifically, **interAdapt** compares the performance of trial designs with adaptive enrollment criteria versus standard (non-adaptive) group sequential trial designs. Performance is compared in terms of power, expected trial duration, and expected sample size. Users can either work directly in the R console, or with a user-friendly **shiny** application that requires no programming experience. Several added features are available when using the **shiny** application. For example, the application allows users to immediately download the results of the performance comparison as a csv-table, or as a printable, html-based report.

Keywords: adaptive design, adaptive enrollment, group sequential design, shiny application.

Introduction

Group sequential, randomized trial designs involve rules for early stopping of an entire trial 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 restricting enrollment criteria based on data accrued in an ongoing trial. For example, enrollment could be stopped for a certain subpopulation if there is strong early evidence that the treatment does not benefit that group. We focus on the class of designs introduced by Rosenblum, Thompson, Luber, and Hanley (2013), which combines features of both group sequential and adaptive enrichment designs. For conciseness, we refer to designs in this class as "adaptive designs." These are contrasted with "standard designs," defined to be group sequential designs where the enrollment criteria cannot be changed during the trial (except that the entire trial may be stopped early for efficacy or futility).

We introduce the **interAdapt** R package, a user friendly set of tools for exploring certain types of adaptive enrichment designs. The package contains a densely featured **shiny** application, ideal for users with little to no R programming experience, as well as an R function

(`compute_design_performance`) which provides the same core computational functionality that underlies the **shiny** application. Several input parameters are available to allow the user to describe the context of his/her trial. Computations for generating output typically require less than 15 seconds on a standard commercial laptop. The **shiny** application is also hosted on the RStudio webserver (`http://spark.rstudio.com/mrosenblum/interAdapt`), and can be accessed online without installing the R programming language.

To demonstrate our 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). The new treatment is called Minimally-Invasive Surgery Plus rt-PA for Intracerebral Hemorrhage (MISTIE), and is described in detail by Morgan, Zuccarello, Narayan, Keyl, Lane, and Hanley (2008). Previous trials had almost exclusively enrolled participants with little or no intraventricular hemorrhage (IVH) at baseline (referred to as small IVH participants). However, it was conjectured that the treatment may also benefit participants with large IVH volume at baseline. The goal of the Phase III trial being planned was to determine whether MISTIE is effective for the combined population of those with small or large IVH, and, if not, to determine whether MISTIE is effective for the small IVH population (for whom there was greater prior evidence). A standard trial design may be inefficient at simultaneously answering these questions. An alternative is to use an adaptive trial design that first recruits from the combined population, and then decides 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 1, we formally define the hypothesis testing problem to be addressed by the different trial designs. In Section 2, we compare our software to the most similar, currently available software: AptivSolutions ADDPLAN PE (Participant Enrichment), and the **asd** R package. In Section 3, we describe how to install and run **interAdapt** locally, as well as how to access the application online through a web browser. In Section 4, we describe the inputs available when using **interAdapt**, and discuss the interpretation of the package's output. In Section 5, we present an example demonstrating how an adaptive design can be created and analyzed with **interAdapt**.

1. Problem description

We consider the problem of designing a randomized trial to test whether a new treatment is superior to control, for a given population (e.g., those with intracerebral hemorrhage in the MISTIE example). Consider the case where we have two subpopulations, referred to as subpopulation 1 and subpopulation 2, which partition the overall population of interest. 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. In the MISTIE trial example, subpopulation 1 refers to small IVH participants, and subpopulation 2 refers to large IVH participants. Let π_1 and π_2 denote the proportion of the population in subpopulations 1 and 2, respectively.

Both the adaptive and standard designs discussed here involve enrollment over time, and include predetermined rules for stopping the trial early based on interim analyses. Each trial consists of K stages, indexed by k. In stages where both subpopulations are enrolled, we assume that the proportion of newly recruited participants in each subpopulation $s \in \{1, 2\}$ is equal to the corresponding population proportion π_s.

For a given design, let n_k denote the maximum number of participants to be enrolled during stage k. The number enrolled during stage k will be less than n_k if the trial is entirely stopped before stage k (so that no participants are enrolled in stage k) or if in the adaptive design enrollment is restricted to only subpopulation 1 before stage k (as described in Section 1.4).

For each subpopulation $s \in \{1, 2\}$ and stage k, let $N_{s,k}$ denote the maximum cumulative number of subpopulation s participants who have enrolled by the end of stage k. Let $N_{C,k}$ denote the maximum cumulative number of enrolled participants from the combined population by the end of stage k, i.e., $N_{C,k} = N_{1,k} + N_{2,k}$. The sample sizes will generally differ for different designs.

Let $Y_{i,k}$ be a binary outcome variable for the i^{th} participant recruited in stage k, where $Y_{i,k} = 1$ indicates a successful outcome. Let $T_{i,k}$ be an indicator of the i^{th} participant recruited in stage k being assigned to the treatment. We assume for each participant that there is an equal probability of being assigned to treatment ($T_{i,k} = 1$) or control ($T_{i,k} = 0$), independent of the participant's subpopulation. We also assume outcomes are observed very soon after enrollment, so that all outcome data is available from currently enrolled participants at each interim analysis.

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 subpopulation 2, let p_{2t} denote the probability of a success under treatment, and p_{2c} denote the probability of a success under control. We assume each of $p_{1c}, p_{1t}, p_{2c}, p_{2t}$ is in the interval $(0, 1)$. We define the true average treatment effect for a given population to be the difference in the probability of a successful outcome comparing treatment versus control.

In the remainder of this section we give an overview of the relevant concepts needed to understand and use **interAdapt**. A more detailed discussion of the theoretical context, and of the efficacy boundary calculation procedure, is provided by Rosenblum *et al.* (2013).

1.1. Hypotheses

We focus on testing the null hypothesis that, on average, the treatment is no better than control for subpopulation 1, and the analogous null hypothesis for the combined population. Simultaneous testing of null hypotheses for these two populations was also the goal for the two-stage, adaptive enrichment designs of Wang, O'Neill, and Hung (2007). We define our two null hypotheses, 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 compares different designs for testing these null hypotheses. An adaptive design testing both null hypotheses (denoted AD) is compared to two standard designs. The first standard design, denoted SC, enrolls the combined population and only tests H_{0C}. The second standard design, denoted SS, only enrolls subpopulation 1 and tests H_{01}. All three trial designs consist of K stages; the decision to entirely stop the trial early can be made at the end of any stage, based on a preplanned rule. The trials differ in that SC and SS never change their enrollment criteria, while AD may switch from enrolling the combined population to enrolling only participants from subpopulation 1.

The standard designs discussed here are not identical to those discussed in Section 6.1 of (Rosenblum *et al.* 2013), which test both hypotheses simultaneously. Implementing standard designs such as those discussed in (Rosenblum *et al.* 2013) into the **interAdapt** software is an area of future research.

Though it is not of primary interest, we occasionally refer below to the global null hypothesis, defined to be that $p_{1t} - p_{1c} = p_{2t} - p_{2c} = 0$, i.e., zero mean treatment effect in both subpopulations.

1.2. Test statistics

Three (cumulative) z-statistics are computed at the end of each stage k. The first is based on all enrolled participants in the combined population, the second is based on all enrolled participants

in subpopulation 1, and the third is based on all enrolled participants in subpopulation 2. Each z-statistic is a standardized difference in sample means, comparing outcomes in the treatment arm versus the control arm. Let $Z_{C,k}$ denote the z-statistic for the combined population at the end of stage k, which takes the following form:

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

The term in square brackets is the difference in sample means between the treatment and control groups. The term in curly braces is the variance of this difference in sample means. $Z_{C,k}$ is only computed at stage k if the combined population has been enrolled up through the end of stage k (otherwise it is undefined). Our designs never use $Z_{C,k}$ after stages where the combined population has stopped being enrolled. Let $Z_{1,k}$ and $Z_{2,k}$ denote analogous z-statistics restricted to participants in subpopulation 1 and subpopulation 2, respectively. These are formally defined in (Rosenblum *et al.* 2013).

1.3. Type I error control

The familywise (also called study-wide) Type I error rate is the probability of rejecting one or more true null hypotheses. For a given design, we say that the familywise Type I error rate is strongly controlled at level α if for any values of $p_{1c}, p_{1t}, p_{2c}, p_{2t}$ (assuming each is in the interval $(0, 1)$), the probability of rejecting at least one true null hypothesis (among H_{0C}, H_{01}) is at most α. To be precise, we mean such strong control holds asymptotically, as sample sizes in all stages go to infinity, as formally defined by Rosenblum *et al.* (2013). For all three designs, AD, SC, and SS, we require the familywise Type I error rate to be strongly controlled at level α. Since the two standard designs SS and SC each only test a single null hypothesis, the familywise Type I error rate for each design is equal to the Type I error rate for the corresponding, single hypothesis test.

1.4. Decision rules for early stopping and for modifying enrollment criteria

The decision rules for the standard design SC consist of efficacy and futility boundaries for H_{0C}, based on the statistics $Z_{C,k}$. At the end of each stage k, the test statistic $Z_{C,k}$ is calculated. If $Z_{C,k}$ is above the efficacy boundary for stage k, the design SC rejects H_{0C} and stops the trial. If $Z_{C,k}$ is between the efficacy and futility boundaries for stage k, the trial is continued through the next stage (unless the last stage $k = K$ has been completed). If $Z_{C,k}$ is below the futility boundary for stage k, the design SC stops the trial and fails to reject H_{0C}. **interAdapt** makes the simplification that the number of participants n_k enrolled in each stage of SC is a constant, denoted n_{SC}, that the user can set.

The efficacy boundaries for SC are set to be proportional to those described by Wang and Tsiatis (1987). Specifically, the efficacy boundary for the k^{th} stage is set to $e_{SC}(N_{C,k}/N_{C,K})^{\delta}$, where K is the total number of stages, δ is a constant in the range $[-.5, .5]$, and e_{SC} is the constant computed by **interAdapt** to ensure the familywise Type I error rate is at most α. Since n_k is set equal to n_{SC} for all values of k, the maximum cumulative sample size $N_{C,k}$ reduces to $\sum_{k'=1}^{k} n_{SC} = k n_{SC}$, and the boundary at stage k reduces to the simpler form $e_{SC}(k/K)^{\delta}$. By default, **interAdapt** sets δ to be -0.5, which corresponds to the efficacy boundaries of O'Brien and Fleming (1979).

In order to calculate e_{SC}, **interAdapt** makes use of the fact that the random vector of test statistics $(Z_{C,1}, Z_{C,2}, \ldots Z_{C,K})$ converges asymptotically to a multivariate normal distribution with a known covariance structure (Jennison and Turnbull 1999). Using the **mvtnorm** package (Genz, Bretz, Miwa, Mi, Leisch, Scheipl, and Hothorn 2013) in R to evaluate the multivariate normal distribution function, **interAdapt** computes the proportionality constant e_{SC} to ensure the probability of $Z_{C,k}$ exceeding $e_{SC}(N_{C,k}/N_{C,K})^{\delta}$ at one or more stages k is less than or equal to α at the global null hypothesis defined in Section 1.1.

In SC, as well as in SS and AD, **interAdapt** uses non-binding futility boundaries. That is, the familywise Type I error rate is controlled at level α regardless of whether the futility boundaries are adhered to or ignored. The motivation is that regulatory agencies may prefer non-binding futility boundaries to ensure Type I error control even if a decision is made to continue the trial despite a futility boundary being crossed.

In calculations of power, expected sample size, and expected trial duration, **interAdapt** assumes futility boundaries are adhered to.

Futility boundaries for the first $K-1$ stages of SC are set equal to $f_{SC}(N_{C,k}/N_{C,K})^{\delta}$, where f_{SC} is a proportionality constant set by the user. By default, the constant f_{SC} is set to be negative (so the trial is only stopped for futility if the z-statistic is below the corresponding negative threshold), although this is not required. In the K^{th} stage of the trial, **interAdapt** sets the futility boundary to be equal to the efficacy boundary. This ensures that the final z-statistic $Z_{C,K}$ crosses either the efficacy boundary or the futility boundary.

The decision boundaries for the design SS are defined analogously as for the design SC, except using z-statistics $Z_{1,k}$. **interAdapt** makes the simplification that the number of participants n_k enrolled in each stage k of SS is constant, denoted by n_{SS}, and set by the user. The efficacy boundary for the k^{th} stage is set equal to $e_{SS}(N_{1,k}/N_{1,K})^{\delta}$, where e_{SS} is the constant computed by **interAdapt** to ensure the Type I error rate is at most α. The first $K-1$ futility boundaries for H_{01} are set equal to $f_{SS}(N_{1,k}/N_{1,K})^{\delta}$, 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.

Consider the adaptive design AD. **interAdapt** allows the user to a priori specify a final stage at which there will be a test of the null hypothesis for the combined population, denoted by stage k^*. Regardless of the results at stage k^*, AD always stops enrolling from subpopulation 2 at the end stage k^*. This reduces the maximum sample size of AD compared to allowing enrollment from both subpopulations through the end of the trial. The futility boundaries $l_{2,k}$ are not defined for $k > k^*$, since subpopulation 2 is not enrolled after stage k^*. The user may effectively turn off the option described in this paragraph by setting $k^* = K$, the total number of stages; then the combined population may be enrolled throughout the trial.

For the AD design, the user can specify the following two types of per-stage sample sizes: one for stages where both subpopulations are enrolled ($k \leq k^*$), and one for stages where only participants in subpopulation 1 are enrolled ($k > k^*$). We refer to these two sample sizes as $n^{(1)}$ and $n^{(2)}$, respectively.

Because AD simultaneously tests H_{0C} and H_{01} it has two sets of decision boundaries. 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 $u_{C,k}$ are set equal to $e_{AD,C}(N_{C,k}/N_{C,K})^{\delta}$ for each $k \leq k^*$; the boundaries $u_{1,k}$ are set equal to $e_{AD,1}(N_{1,k}/N_{1,K})^{\delta}$ for each $k \leq K$. The constants $e_{AD,C}$ and $e_{AD,1}$ are set such that the probability of rejecting one or more null hypotheses under the global null hypothesis is α (ignoring futility boundaries). It is proved by Rosenblum *et al.* (2013) that this strongly controls the familywise Type I error rate at level α. The algorithm for computing the proportionality constants $e_{AD,C}, e_{AD,1}$ is described later in this section.

The boundaries for futility stopping of enrollment from certain population in the AD design, at the end of stage k, are denoted by $l_{1,k}$ and $l_{2,k}$. These stopping boundaries are defined relative to the test statistics $Z_{1,k}$ and $Z_{2,k}$, respectively. The boundaries $l_{1,k}$ and $l_{2,k}$ are set equal to

$f_{AD,1}(N_{1,k}/N_{1,K})^\delta$ (for $k \leq K$) and $f_{AD,2}(N_{2,k}/N_{2,K})^\delta$ (for $k < k^*$), respectively, where $f_{AD,1}$ and $f_{AD,2}$ can be set by the user. In stage k^*, the futility boundary l_{2,k^*} is set to "Inf" (indicating ∞), to reflect that we stop enrollment in subpopulation 2. At the end of each stage, AD may decide to continue enrolling from the combined population, enroll only from subpopulation 1 for the remainder of the trial, or stop the trial entirely. Specific decision rules based on these boundaries for the z-statistics are described below.

As discussed in (Rosenblum *et al.* 2013), the decision rules in AD consist of the following steps carried out at the end of each stage k:

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

2. (Assess Futility of 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 any null hypothesis.

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

   3a. If $Z_{1,k} > u_{1,k}$, reject H_{01} and stop all enrollment.

   3b. If $Z_{1,k} \leq l_{1,k}$ or if this is the final stage of the trial, fail to reject any null hypothesis and stop all enrollment.

   3c. Else, continue enrolling from only subpopulation 1. If $k < k^*$ then $\pi_1 n^{(1)}$ participants from subpopulation 1 should be enrolled in the next stage. If $k \geq k^*$, then $n^{(2)}$ participants from subpopulation 1 should be enrolled in the next stage. In all future stages, ignore steps 1, 2, 4, and use steps 3a–3c.

4. (Continue Enrollment from Combined Population) Else, continue by enrolling $\pi_1 n^{(1)}$ participants from subpopulation 1 and $\pi_2 n^{(1)}$ participants from subpopulation 2 for the next stage.

The motivation for Step 2 is that there is assumed to be prior evidence that if the treatment works, it will work for subpopulation 1. Therefore, if subpopulation 1 is stopped for futility, the whole trial is stopped. It is an area of future research to consider modifications to this rule, and to incorporate testing of a null hypothesis for only subpopulation 2.

A consequence of the rule in Step 3 is that Steps 1, 2, and 4 are only carried out for stages $k \leq k^*$. This occurs since Step 3 restricts enrollment to subpopulation 1 if $Z_{2,k} \leq l_{2,k}$ or $k \geq k^*$, and if so runs Steps 3a–3c through the remainder of the trial.

We next describe the algorithm used by **interAdapt** to compute the proportionality constants $e_{AD,C}, e_{AD,1}$ that define the efficacy boundaries $u_{C,k}, u_{1,k}$. These are selected to ensure the familywise Type I error rate is strongly controlled at level α. By Theorem 5.1 of (Rosenblum *et al.* 2013), to guarantee such strong control of the familywise Type I error rate, it suffices to set $u_{C,k}, u_{1,k}$ such that the familywise Type I error rate is at most α at the global null hypothesis defined in Section 1.1. The algorithm takes as input the following, which are set by the user as described in Section 4.1.1: the per-stage sample sizes $n^{(1)}, n^{(2)}$, the study-wide (i.e., familywise) Type I error rate α, and a value a_c in the interval $[0, 1]$. Roughly speaking, a_c represents the fraction of the study-wide Type I error α initially allocated to testing H_{0C}, as described next.

The algorithm temporarily sets $e_{AD,1} = \infty$ (effectively ruling out rejection of H_{01}) and computes (via binary search) the smallest value $e_{AD,C}$ such the probability of rejecting H_{0C} is $a_c\alpha$ under the global null hypothesis defined in Section 1.1. This defines $e_{AD,C}$. Next, **interAdapt** computes the smallest constant $e_{AD,1}$ such that the probability of rejecting at least one null hypothesis under the global null hypothesis is at most α.

All of the above computations use the approximation, based on the multivariate central limit theorem, that the joint distribution of the z-statistics is multivariate normal with covariance matrix as given, e.g., by Jennison and Turnbull (1999); Rosenblum *et al.* (2013).

2. Related software

The most comparable available software tools are AptivSolutions ADDPLAN PE (Participant Enrichment), and the **asd** R package (Parsons, Friede, Todd, Marquez, Chataway, Nicholas, and Stallard 2012). Both have features that our software does not have. Conversely, there are features of our software that ADDPLAN PE and **asd** do not have.

ADDPLAN PE is a versatile, commercial software tool that implements many types of adaptive enrichment designs. One limitation is that 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 prior uncertainty as to when sufficient information will have accrued to make such a decision. In contrast, **interAdapt** is more flexible, in that one can select designs in which the decision to change enrollment criteria may occur at any stage (by setting k^* to the maximum number of stages K). **interAdapt** implements the class of designs from (Rosenblum *et al.* 2013), while ADDPLAN PE does not. However, ADDPLAN implements a wide variety of other decision rules and testing procedures not available in **interAdapt**. Finally, **interAdapt** is cross-platform and open-source, while ADDPLAN PE is commercial software that is only compatible with the Windows OS.

The **asd** R package allows users to generate two-stage adaptive designs, which can be used to combine phase II and phase III clinical trials into a seamless design (Parsons *et al.* 2012). Unlike **interAdapt**, the **asd** package can generate not only adaptive enrichment designs, but also adaptive designs that test multiple different treatments. Also in **asd**, decision rules at interim analyses can be based on short-term outcomes for each subject enrolled, if the long-term outcome is not yet available. However, **asd** does not allow more than two stages, unlike **interAdapt** which allows up to 20 stages (though in practice fewer stages will probably be used, e.g., 5 stages). Using **asd** requires a working knowledge of R, while the GUI for **interAdapt** can be run in a web browser, with little to no interaction with the R console (see Section 3), and so does not require knowledge of the R language.

3. Running interAdapt

The **interAdapt** R package contains an interactive web browser application, as well as a command line function which performs the same computations. The browser application is built on the **shiny** (RStudio and Inc. 2013) and **RCurl** (Lang 2013) packages, with the back-end calculations done in R.

To access the **shiny** application, **interAdapt** requires the user's default web browser to be set to either Firefox (`http://www.mozilla.org`) or Chrome (`http://www.google.com/chrome/`). Users can then run the **shiny** application locally by installing R and the **interAdapt** package from CRAN. Loading the **interAdapt** package will automatically prompt the user on whether the **shiny** application should be loaded.

Users can also access the software online, without installing R, simply by following the link below. However, because the online application will slow down noticeably if accessed by multiple users simultaneously, we encourage heavy users to install **interAdapt** locally.

`http://spark.rstudio.com/mrosenblum/interAdapt`

The same calculations that the **shiny** application performs can also be done directly in the R console, with the `compute_design_performance` function. Further details are provided in the

interAdapt package documentation. The arguments for `compute_design_performance` are the same as the parameters available in the **shiny** application (see Section 4.1). The function's value contains the output tables of the **shiny** application, which can be used to generate the plots shown by the application (see Section 4.2).

4. User interface for the shiny application

In this section, and in Section 5, we will generally use the term **interAdapt** to refer to the **interAdapt shiny** application, although the inputs and outputs of the `compute_design_performance` function have the same interpretation.

Inputs to **interAdapt** can be entered in the side panel on the left, with outputs shown in the main panel on the right (Figures 1 and 2). The parameters in the input panel let the user describe characteristics of their study populations, such as the proportion of participants in each subpopulation. The user can also input design requirements, such as the familywise Type I error rate. Also, the user can input conjectured rates of success under treatment and control, to determine how well different designs perform at a given set of such values. Specifically, the user can input values for p_{1t}, p_{1c}, and p_{2c}, and **interAdapt** will compare the performance of different designs over a range of values of p_{2t}, as further described in Section 4.1.

The main panel displays the decision boundaries and trial designs computed by **interAdapt** to satisfy the requirements specified by the user (Figure 1). It also compares the performance of the three designs, AD, SC and SS (Figure 2). Performance is compared in terms of power, expected sample size, and expected trial duration.

All tables generated by **interAdapt** can be downloaded as csv files by clicking on the "Download" button beneath the table. Users can also download a printable, html-based report of the results by clicking the "Generate Report" button at the bottom of the main panel (Figures 1 and 2). This report is generated with the **knitr** package for R (Xie 2013). Citations in the report are created using the **knitcitations** package (Boettiger 2013).

4.1. Inputs

Parameters in the input panel are organized into the following 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 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 a single 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, click the dropdown menu titled "Show Basic Parameters" and select "Show All Parameters and Save/Load Option". You can then save the current parameters as a csv file, or load a previously saved csv file of inputs (Figure 2). 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 csv format (e.g., from a previous trial or study) to use in setting population parameters for the current trial. For example, if one is planning a Phase III trial, one might upload Phase II trial data that is already available. The purpose of this feature is to allow the data generating mechanisms in the **interAdapt** simulations

to mimic properties of real datasets relevant to the study being planned. The uploaded dataset must be structured to have one row for each participant. 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 the empirical values π_1, p_{1c}, p_{1t}, p_{2c}, and p_{2t}, and adjust the input sliders accordingly. The user can then modify these parameter settings to determine how robust a given design is to differences from what was observed in previous studies.

A detailed explanation of each input is given below.

Basic Parameters (with corresponding variables in parentheses, where applicable)

- 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 a successful outcome for subpopulation 1 under assignment to the control arm. 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 a successful outcome for subpopulation 2 under assignment to the control arm. This is used in estimating power and expected sample size of each design.
- Probability outcome = 1 under treatment for subpopulation 1 (p_{1t}): The probability of a successful outcome for subpopulation 1 under assignment to the treatment arm. Note that the user does not specify p_{2t}; instead, **interAdapt** considers a range of possible values of p_{2t} that can be set through the Advanced Parameters described below.
- Per stage sample size, combined population, for adaptive design ($n^{(1)}$): Number of participants enrolled per stage in AD, whenever both subpopulations are being enrolled.
- Per stage sample size for stages where only subpopulation 1 is enrolled, for adaptive design ($n^{(2)}$): The number of participants required for each stage in AD after stage k^* (only used if $k^* < K$). For stages up to and including stage k^*, the number of participants enrolled from subpopulation 1 is equal to $\pi_1 n^{(1)}$.
- Alpha (FWER) requirement for all designs (α): The familywise Type I error rate defined in Section 1.3.
- Proportion of Alpha allocated to H0C for adaptive design (a_C): This is used in the algorithm in Section 1.4 to construct efficacy boundaries for the design AD.

Advanced Parameters (with corresponding variables in parentheses, where applicable)

- Delta (δ): This parameter is used as the exponent in defining the efficacy and futility boundaries as described in Section 1.4.
- # of Iterations for simulation: This is the number of simulated trials used to approximate the power, expected sample size, and expected trial duration. In each simulated trial, z-statistics are simulated from a multivariate normal distribution (determined by the input parameters). The greater the number of iterations, the more accurate the simulation results will be. It is our experience that a simulation with 10,000 iterations takes about 7-15 seconds on a commercial laptop.

- 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 avoid this, either the number of iterations can be reduced, or the time limit for the simulation can be extended. **interAdapt** does not allow for the time limit to exceed 90 seconds in the online version; there is no such restriction on the local version.

- Total number of stages (K): The total number of stages, which is used in each type of design. The maximum allowed number of stages is 20.

- Last stage subpopulation 2 is enrolled under adaptive design (k^*): In the adaptive design, no participants from subpopulation 2 are enrolled after stage k^*.

- Participants enrolled per year from combined population: This is the assumed enrollment rate (per year) for the combined population. It impacts the expected duration of the different trial designs. The enrollment rates for subpopulations 1 and 2 are assumed to equal the combined population enrollment rate multiplied by π_1 and π_2, respectively. I.e., enrollment rates are proportional to the relative sizes of the subpopulations. This reflects the reality that enrollment will likely be slower for smaller subpopulations. Active enrollment from one subpopulation is assumed to have no effect on the enrollment rate in the other subpopulation. This implies that each stage of the AD design up to and including stage k^* takes the same amount of time to complete, regardless of whether enrollment stops for subpopulation 2. Also, each stage after k^* takes the same amount of time to complete.

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

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

- Stopping boundary proportionality constant for subpopulation 2 enrollment for adaptive design ($f_{AD,2}$): This is used to calculate the futility boundaries ($l_{2,k}$) for the z-statistics calculated in subpopulation 2 ($Z_{2,k}$) as defined in Section 1.4.

- H_{01} futility boundary proportionality constant for the adaptive design ($f_{AD,1}$): This is used to calculate the futility boundaries ($l_{1,k}$) for the z-statistics calculated in subpopulation 1 ($Z_{1,k}$) as defined in Section 1.4.

- H_{0C} futility boundary proportionality constant for the standard design (f_{SC}): This is used to calculate the futility boundaries for H_{0C} in SC as defined in Section 1.4.

- H_{01} futility boundary proportionality constant for the standard design (f_{SS}): This is used to calculate the futility boundaries for H_{01} in SS as defined in Section 1.4.

- Lowest value to plot for treatment effect in subpopulation 2: **interAdapt** does simulations under a range of treatment effect sizes $p_{2t} - p_{2c}$ for subpopulation 2. This sets the lower bound for this range. This also effectively sets the lower bound for p_{2t}, since p_{2c} is set by the user as a Basic parameter.

- Greatest value to plot for treatment effect in subpopulation 2: **interAdapt** does simulations under a range of treatment effect sizes $p_{2t} - p_{2c}$ for subpopulation 2. This sets the upper bound for this range.

4.2. Outputs

The output panel on the right side of the user interface is split into the following three sections: "About interAdapt", "Designs", and "Performance." Users can navigate between these sections using the radio buttons at the top of the panel. The About interAdapt section gives a brief introduction to the software, and a link to the full software documentation. The Designs section describes the design parameters for each of the three trials: SC, SS, and AD. This includes the efficacy and futility boundaries, and the maximum cumulative number of participants enrolled by the end of each stage (under no early stopping). The Performance section compares the three designs in terms of their power, expected sample size, and expected duration.

Designs

The Designs section gives design features that result from the user's inputs. Tabs at the top of the page can be used to navigate between the different designs. Each of the first three tabs corresponds to one of the designs, and the fourth tab shows all three designs together.

In the "Adaptive" tab, the table at the bottom of the page shows the maximum cumulative number of participants enrolled by the end of each stage for the design AD. This is broken down by subpopulation. 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 AD stops enrolling subpopulation 2 after stage k^*, futility boundaries $l_{2,k}$ (for statistics $Z_{2,k}$) in stages $k > k^*$ are not given, and l_{2,k^*} is set to "Inf" (indicating ∞). Efficacy boundaries for $Z_{C,k}$ are not given for stages $k > k^*$ since by construction (see Section 1.4) the AD design does not test H_{0C} after stage k^*. (It is an area of future research to consider designs that continue to test H_{0C} even after enrollment for subpopulation 2 has stopped.) A plot at the top of the page displays the efficacy and futility boundaries over all stages of the trial.

The two tabs for the standard designs SC and SS have a comparable layout to that for AD. Note that the efficacy boundaries for SS and SC are identical. This is because the efficacy boundaries for SC and SS are both proportional to $(k/K)^\delta$, with proportionality constants set to achieve Type I error α, which leads to identical proportionality constants for SC and SS.

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

Performance output

interAdapt displays the performance of each of the three designs in terms of three metrics: power, expected sample size, and expected duration. 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** calculates performance metrics (using simulations as described in the section on Advanced Parameters) for the specified range of treatment effects, and generates plots of each metric versus the treatment effect in subpopulation 2. The plot showing each metric can be accessed via the tabs at the top of the page. The table at the bottom of the Performance section shows all three metrics, with each column of the table denoting a different treatment effect in subpopulation 2.

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}. Since the standard design SC only tests H_{0C}, **interAdapt** only shows its power to reject H_{0C}. Similarly, **interAdapt** only shows the power of SS to reject H_{01}. The power of SS is constant with respect to the true treatment effect in subpopulation 2. This is as expected, since SS does not enroll any participants from subpopulation 2.

For the standard designs SC and SS, the expected duration is proportional to the expected sample size. However, for the AD design, this does not hold; this is because the total trial

duration is not necessarily proportional to the total sample size. E.g., stopping subpopulation 2 will reduce the sample size but not necessarily reduce the trial duration if subpopulation 1 is not stopped.

5. Example of entering input and interpreting output

The default inputs to **interAdapt** come from the motivating example of planning the MISTIE Phase III trial. We next summarize the design goals of the investigators, based on (Rosenblum *et al.* 2013). The MISTIE III trial aims to assess a new surgical treatment for stroke. The primary outcome is based on each participant's disability score on the modified Rankin Scale (mRS). A successful outcome was defined as a mRS score less than or equal to 3.

At the time of planning the Phase III MISTIE trial, the previous Phase II trial had only enrolled participants with little or no intraventricular hemorrhage (IVH). More specifically, participants 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, participants were classified as "large IVH." The Phase II trial only recruited small IVH participants, and yielded 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 participants, but very little data was available to assess this. We refer to those with small IVH as subpopulation 1, since there was more prior evidence of treatment efficacy in this subpopulation; those with large IVH are subpopulation 2.

The study designers focused on the following three scenarios of special interest:

(a) The average treatment effect (on the risk difference scale) is 12.5% for both small and large IVH participants;

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

(c) The treatment effect is zero for both subpopulations.

The goals were as follows:

(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) Familywise Type I error rate (α) of 0.025.

Furthermore, the familywise Type I error rate was to be strongly controlled at level 0.025.

Based on prior research by Hanley (2012), the proportion of participants with small IVH (π_1) was projected to be 0.33, the probability of a positive outcome under control was projected to be 0.25 for small IVH participants (p_{1c}), and the probability of a positive outcome under control was projected to be 0.2 for large IVH participants (p_{2c}). If the true average treatment effect in subpopulation 1 is 12.5%, then the probability of a positive outcome under treatment for participants in subpopulation 1 (p_{1t}) is projected to be 12.5%+25%=37.5%.

Parameters for the adaptive design AD were computed to achieve all three goals (i)-(iii), as fully described by (Rosenblum *et al.* 2013). The corresponding standard designs, used for comparison, only had to satisfy subsets of these goals. This is to show the cost of achieving all three goals (since the adaptive design generally requires greater expected sample size, in return for achieving all three goals instead of a subset of the goals). The standard design SC was set to achieve goals (i) and (iii), and the standard design SS was set to 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 mean treatment effect in subpopulation 2 to be [-0.2, 0.2]. This includes scenarios (a) and (b) since in scenario (a) the mean treatment effect in subpopulation 2 is 0.125, and in scenario (b) the mean treatment effect is 0.

The remaining default input parameters for the *AD* design are based on the adaptive enrichment design in Section 5.2 of (Rosenblum *et al.* 2013). They constructed this design by first setting $K = 5$ and $\delta = -.5$, and then searching over a large class of parameter values with the goal of minimizing the average expected sample size over scenarios (a)-(c), while still achieving goals (i)-(iii). They found a minimum average expected sample size at $k^* = 3$, $n^{(1)} = 280$, $n^{(2)} = 148$, $a_C = .09$, and $f_{AD,2} = f_{AD,1} = 0$.

Now we turn to the output of **interAdapt** that results from the default parameters, and show that each of the three designs achieves its corresponding 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) (Figure 2). Although it is not shown, the familywise Type I error rate is at most .025, as this was the specified value of α input to **interAdapt**, and the designs are guaranteed to strongly control the familywise Type I error rate at level α by Theorem 5.1 of (Rosenblum *et al.* 2013).

Summary

We described the **interAdapt** R package and **shiny** 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.

Current limitations of the software include that the outcome is assumed to be binary. We also currently only consider the case where outcomes are measured without delay, immediately after participants are enrolled. Relaxing both of these requirements is a goal of future work.

Acknowledgements

This research was supported by U.S. National Institute of Neurological Disorders and Stroke (grant numbers 5R01 NS046309-07 and 5U01 NS062851-04), the U.S. Food and Drug Administration through the "Partnership in Applied Comparative Effectiveness Science," (contract HHSF2232010000072C), and the National Institute of Environmental Health Sciences (grant number T32ES012871). This publication's contents are solely the responsibility of the authors and do not necessarily represent the official views of the above agencies.

References

Boettiger C (2013). ***knitcitations**: Citations for **knitr** Markdown Files*. R package version 0.4-7, URL http://CRAN.R-project.org/package=knitcitations.

Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T (2013). ***mvtnorm**: Multivariate Normal and t Distributions*. R package version 0.9-9996, URL http://CRAN.R-project.org/package=mvtnorm.

Hanley D (2012). http://braininjuryoutcomes.com/studies/mistie/entry/mistie/international-stroke-conference-2012-mistie-phase-2-results.

Jennison C, Turnbull BW (1999). *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC Press.

Lang DT (2013). ***RCurl**: General Network (HTTP/FTP/...) Client Interface for R*. R package version 1.95-4.1, URL http://CRAN.R-project.org/package=RCurl.

Morgan T, Zuccarello M, Narayan R, Keyl P, Lane K, Hanley DF (2008). "Preliminary Findings of the Minimally-Invasive Surgery plus rtPA for Intracerebral Hemorrhage Evacuation (MISTIE) Clinical Trial." *Acta Neurochirurgica Supplementum*, **105**, 147–51.

O'Brien P, Fleming T (1979). "A Multiple Testing Procedure for Clinical Trials." *Biometrics*, **35**, 549–556.

Parsons N, Friede T, Todd S, Marquez EV, Chataway J, Nicholas R, Stallard N (2012). "An R Package for Implementing Simulations for Seamless Phase II/III Clinical Trials Using Early Outcomes for Freatment Selection." *Computational Statistics & Data Analysis*, **56**(5), 1150–1160.

Rosenblum M, Thompson RE, Luber BS, Hanley DF (2013). "Adaptive Group Sequential Designs that Balance the Benefits and Risks of Expanding Inclusion Criteria." *Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 250.* URL http://biostats.bepress.com/jhubiostat/paper250.

RStudio, Inc (2013). ***shiny**: Web Application Framework for R*. R package version 0.8.0, URL http://CRAN.R-project.org/package=shiny.

Wang SJ, O'Neill RT, Hung H (2007). "Approaches to Evaluation of Treatment Effect in Randomized Clinical Trials with Genomic Subsets." *Pharmaceutical Statistics*, **6**, 227–244.

Xie Y (2013). ***knitr**: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.4.1, URL http://yihui.name/knitr/.

Affiliation:

Michael Rosenblum
Department of Biostatistics
Assistant Professor
Johns Hopkins Bloomberg School of Public Health
615 N. Wolfe St. Room E3616
E-mail: mrosenbl@jhsph.edu
URL: http://people.csail.mit.edu/mrosenblum/

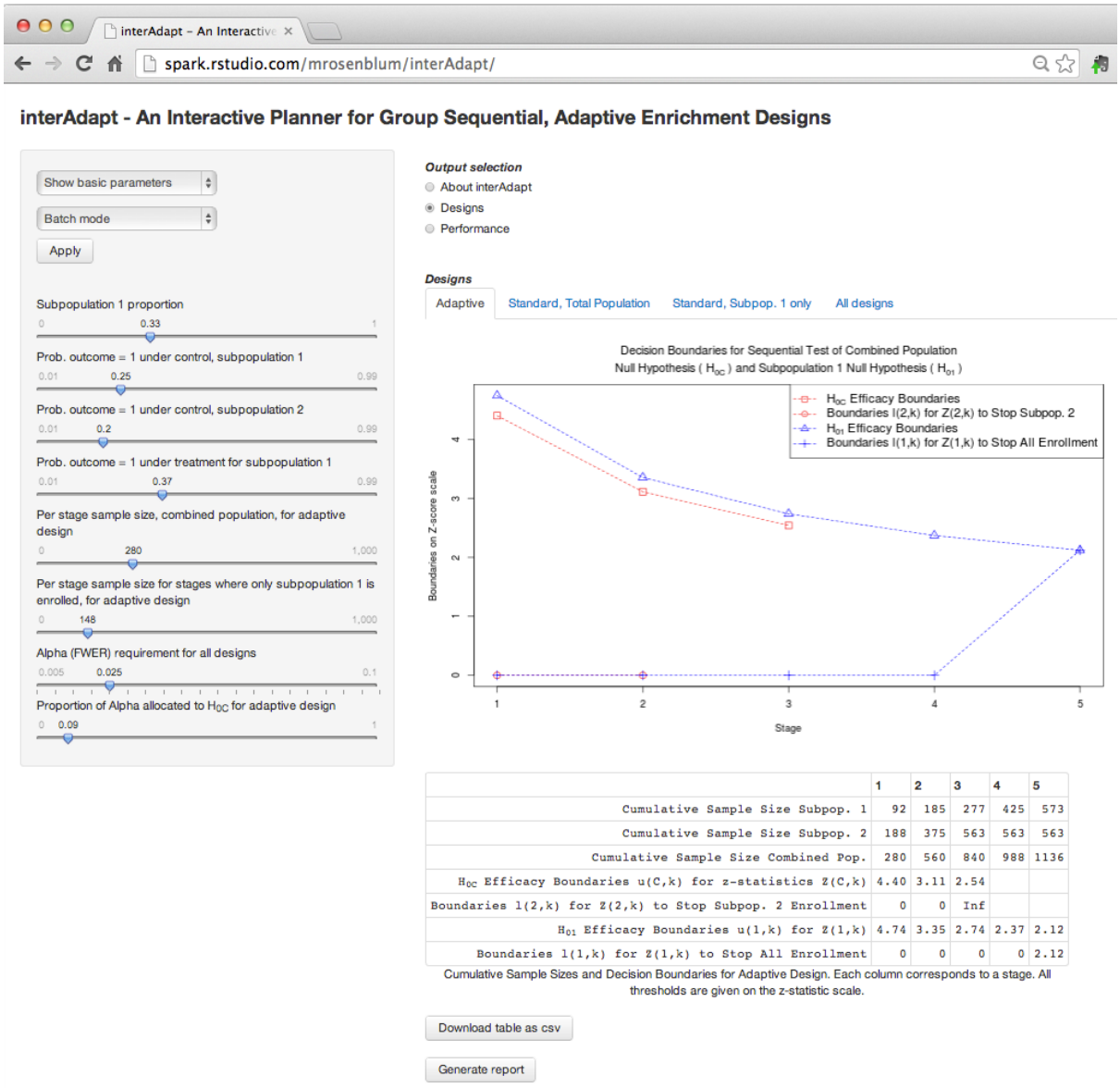
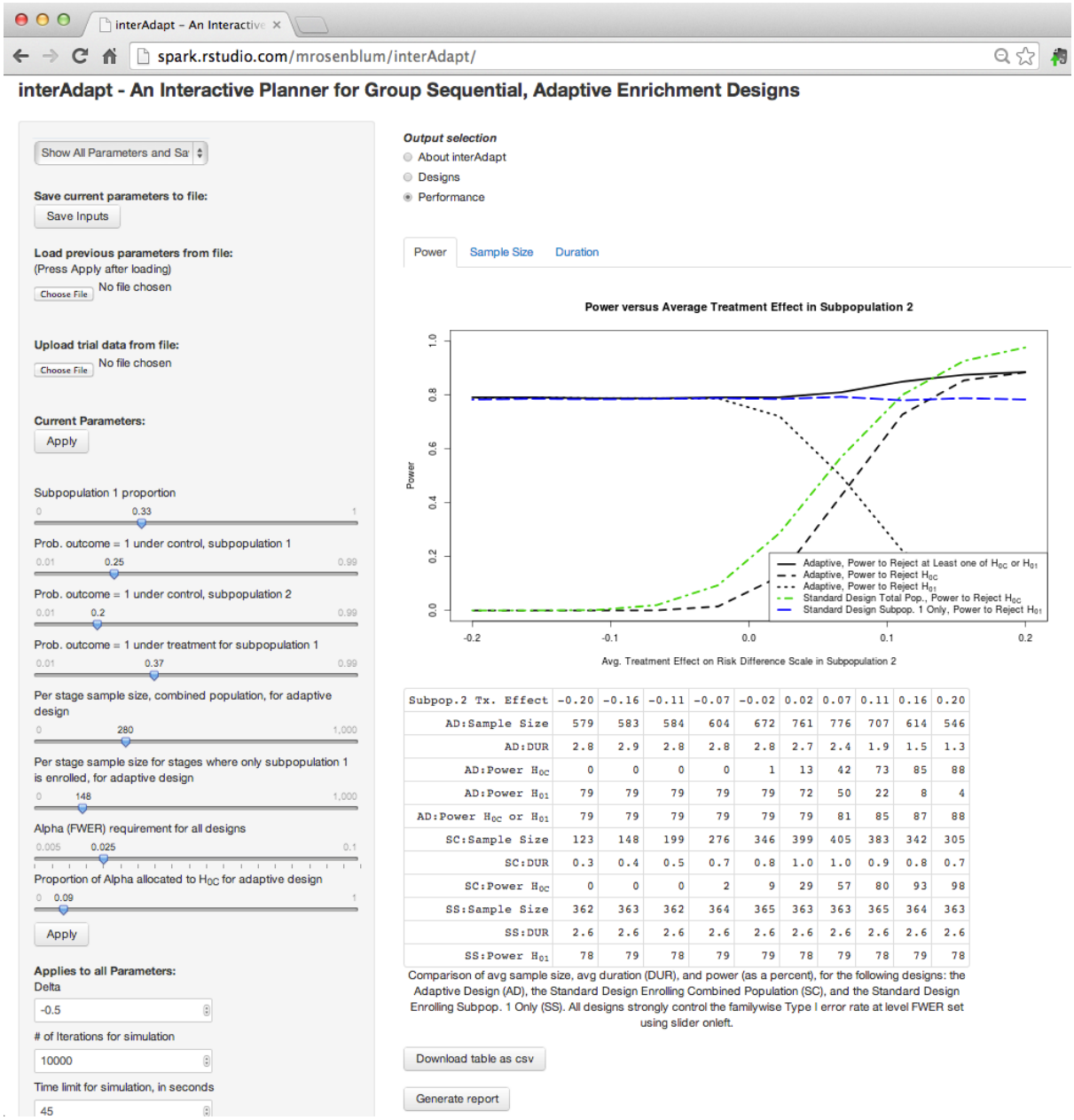


Figure 1: Designs Screenshot: Inputs can be entered in the side panel on the left, with results visible in the main panel on the right. The drop down menus at the top of the side panel can be used to navigate different interfaces to input parameters. Here we show the "Basic parameter" inputs, in "Batch mode," where the Apply button must be pressed to update the results in the main panel. The radio buttons at the top of the main panel can be used to navigate between design outputs describing the decision rules for each trial, and performance summaries for each trial. In this figure we show the design for the adaptive trial (AD), based on the default input parameters. Boundaries for the z-statistics $Z_{1,k}$, $Z_{2,k}$ and $Z_{C,k}$ are shown both in the plot, and in the table. The table also contains information on how many participants are enrolled in each stage. The scroll bar on the right of the web browser has been cropped out of this figure for the sake of increased screenshot resolution.



Subpop.2 Tx. Effect	-0.20	-0.16	-0.11	-0.07	-0.02	0.02	0.07	0.11	0.16	0.20
AD:Sample Size	579	583	584	604	672	761	776	707	614	546
AD:DUR	2.8	2.9	2.8	2.8	2.8	2.7	2.4	1.9	1.5	1.3
AD:Power H_{0C}	0	0	0	0	1	13	42	73	85	88
AD:Power H_{01}	79	79	79	79	79	72	50	22	8	4
AD:Power H_{0C} or H_{01}	79	79	79	79	79	79	81	85	87	88
SC:Sample Size	123	148	199	276	346	399	405	383	342	305
SC:DUR	0.3	0.4	0.5	0.7	0.8	1.0	1.0	0.9	0.8	0.7
SC:Power H_{0C}	0	0	0	2	9	29	57	80	93	98
SS:Sample Size	362	363	362	364	365	363	363	365	364	363
SS:DUR	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6
SS:Power H_{01}	78	79	78	79	79	78	79	78	79	78

Comparison of avg sample size, avg duration (DUR), and power (as a percent), for the following designs: the Adaptive Design (AD), the Standard Design Enrolling Combined Population (SC), and the Standard Design Enrolling Subpop. 1 Only (SS). All designs strongly control the familywise Type I error rate at level FWER set using slider onleft.

Figure 2: Performance Screenshot: Here the main panel shows performance output based on the default parameter inputs. The tabs at the top of the Performance section can be used to navigate between displays of power, expected sample size, and expected trial duration for all three designs. In the side panel, we show the interface for saving and loading sets of parameters (section 4.1). Users can save the current set of inputs, load a previously used set of inputs, or upload a datafile containing results from a previous trial. If results from a previous trial are uploaded, **interAdapt** will automatically compute relevant input parameters based on this file. Additional input parameters in the side panel are available by scrolling down. As in Figure 1, the scroll bar on the right of the web browser has been cropped out of this figure for the sake of increased screenshot resolution.