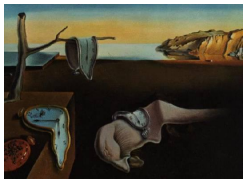

Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue 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

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 participants, such as those with a certain biomarker or risk score at baseline. Randomized trial designs have been proposed that dynamically adapt enrollment criteria based on accrued data, with the goal of learning if the treatment benefits the overall population, only a certain subpopulation, or neither. We introduce the **interAdapt** software tool, a densely featured **shiny** application 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 (Rosenblum, Thompson, Luber, and Hanley 2013).

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 may be restricted to a certain subpopulation if there is strong early evidence that the complementary subpopulation is not benefiting from treatment. We focus on the class of designs introduced by Rosenblum *et al.* (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 the entire trial may be stopped early for efficacy or futility).

We introduce the **interAdapt** software tool, a densely featured **shiny** application which provides a user-friendly interface for exploring certain types of adaptive enrichment designs, and for

comparing these to standard designs. The software can either be run locally as an R package, or accessed online through a web browser. **interAdapt** is designed to be used by statisticians and clinical investigators to plan randomized trials. The software provides information that can help users determine if certain adaptive designs offer tangible benefits compared to standard designs, in the context of their specific trial goals and constraints. Calculations typically require less than 1 minute on a standard commercial laptop. Several user inputs are available to allow the user to describe the context of his/her trial. Alternatively, users can upload data from previous studies, and **interAdapt** will automatically compute the relevant parameters for the trial being planned. Once entered, the full set of input parameters can be saved to the user's computer for use in future sessions. Results of the design comparisons can be immediately downloaded in the form of either csv-tables, or printable, html-based reports.

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 different trial designs. In Section 2, we compare our software to the most similar, currently available commercial software, AptivSolutions ADDPLAN PE (Participant Enrichment). In Section 3, we describe how to install **interAdapt** on a personal computer, and how to access it online through a web browser. Section 4 describes the inputs available when using interAdapt, and discusses the interpretation of the application's output. In Section 5, we present an example demonstrating how an adaptive design is 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). 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 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 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 parameter 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}: $\pi_1(p_{1t} - p_{1c}) + \pi_2(p_{2t} - p_{2c}) \leq 0$.

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.

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, and the decision to entirely stop the trial early can be made at the end of any stage using 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 hypothesis simultaneously. Implementing standard designs such as those discussed in (Rosenblum *et al.* 2013) into the **interAdapt** software is an area of future research.

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:

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

The term in square brackets is the difference in sample means between the treatment and control groups. The term in curly 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 the probability of rejecting at least one true null hypothesis (among H_{0C}, H_{01}) is at most α, regardless of the true values of $p_{1c}, p_{1t}, p_{2c}, p_{2t}$. To be precise, we mean such strong control holds asymptotically, as sample sizes in all stages goes 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 reached). 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). This means that the efficacy boundary for the k^{th} stage is set to $e_{SC}\{(\sum_{k'=1}^{K} n_{k'})/(\sum_{k'=1}^{k} n_{k'})\}^{-\delta}$, where K is the total number of stages, δ is a constant in the range $[-.5, .5]$, and e_{SC} is the constant 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, this boundary reduces to $e_{SC}(k/K)^{\delta}$. 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** calculates the proportionality constant e_{SC} such that the probability of $Z_{C,k}$ exceeding $e_{SC}\{(\sum_{k'=1}^{K} n_{k'})/(\sum_{k'=1}^{k} n_{k'})\}^{-\delta}$ 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 constants. All three designs are calibrated such that 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. However, it is possible to assess the impact of ignoring futility boundaries by setting them to $-\infty$ (entered in the software as "-Inf") and recomputing these quantities.

Futility boundaries for the first $K-1$ stages of SC are set equal to $f_{SC}\{(\sum_{k'=1}^{K} n_{k'})/(\sum_{k'=1}^{k} n_{k'})\}^{-\delta}$, 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 patients n_k enrolled in each stage of SS is constant, denoted n_{SS}, input by the user. The efficacy boundary for the k^{th} stage is set equal to $e_{SS}\{(\sum_{k'=1}^{K} n_{k'})/(\sum_{k'=1}^{k} n_{k'})\}^{-\delta}$, 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 $f_{SS}\{(\sum_{k'=1}^{K} n_{k'})/(\sum_{k'=1}^{k} n_{k'})\}^{-\delta}$, 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.

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 $u_{C,k}$ and $u_{1,k}$ are set equal to $e_{AD,C}\{(\sum_{k'=1}^{K} n_{k'})/(\sum_{k'=1}^{k} n_{k'})\}^{-\delta}$ and $e_{AD,1}\{(\sum_{k'=1}^{K} n_{k'})/(\sum_{k'=1}^{k} n_{k'})\}^{-\delta}$ respectively, where $e_{AD,C}$ and $e_{AD,1}$ are constants 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 below.

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 $f_{AD,1}\{(\sum_{k'=1}^{K} n_{k'})/(\sum_{k'=1}^{k} n_{k'})\}^{-\delta}$ and $f_{AD,2}\{(\sum_{k'=1}^{K} n_{k'})/(\sum_{k'=1}^{k} n_{k'})\}^{-\delta}$, respectively, where $f_{AD,1}$ and $f_{AD,2}$ can be set by the user. 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.

interAdapt allows the user to a priori specify a final stage at which there will be a test of in 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^*. The motivation is that since information may be expected to accrue more quickly for the combined population than for a single subpopulation, it may be desirable to stop enrolling the combined population when sufficient information has accrued; 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 when both populations are enrolled ($k \leq k^*$), and one for stages where only patients in subpopulation 1 is enrolled ($k > k^*$). We refer to these two sample sizes as $n^{(1)}$ and $n^{(2)}$ respectively.

As described in (Rosenblum *et al.* 2013), the decision rule in AD consists 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 $\pi_1 n^{(1)}$ patients should be enrolled in the next stage. If $k \geq k^*$, then $n^{(2)}$ patients should be enrolled in the next stage. In all future stages, ignore steps (1–2) 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.

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 distribution with covariance matrix as given, e.g., by Jennison and Turnbull (1999); Rosenblum *et al.* (2013).

2. Related software

The most comparable available software is AptivSolutions ADDPLAN PE (Participant Enrichment). It is versatile, 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 *et al.* 2013). 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 prior 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 in which the decision to change enrollment criteria may occur at any stage (by setting k^* to the maximum number of stages K).

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

3. Running interAdapt

interAdapt is an interactive application built on the **shiny** package (RStudio and Inc. 2013) for the R programming language (http://www.r-project.org/). The user interface is shown in the user's web browser, while the back-end calculations are done in R.

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 **interAdapt** either by installing R and the **interAdapt** R package locally on their computer, or by simply using Firefox or Google Chrome to view **interAdapt** online. Both options are free and quick to set up. However, because the online application will slow down noticeably if accessed by multiple users simultaneously, we encourage heavy users to install **interAdapt** locally.

3.1. Running interAdapt over the web

interAdapt is currently hosted on the RStudio webserver, and can be accessed at:
http://spark.rstudio.com/mrosenblum/interAdapt

3.2. Running interAdapt locally

To run **interAdapt** locally, one must first install the R programming language, which is free. R runs on both Windows & MacOS, is 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's 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

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

4. User interface

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 below.

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 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 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 planning the current trial. The 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 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 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 patients 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 patients 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 patients 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

- 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.

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

- 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. Active enrollment from one subpopulation is assumed to have no affect on the enrollment rate in the other subpopulation. This implies that all stages of the AD design up to and including stage k^* take the same amount of time to complete, regardless of whether we cease enrollment in subpopulation 2. Stages after stage k^* may require a longer or shorter amount time to complete, relative to the earlier stages, depending on whether the number of patients enrolled per stage from subpopulation 1 increases or decreases after stage k^* (i.e., whether $n^{(2)} > \pi_1 n^{(1)}$ or $n^{(2)} < \pi_1 n^{(1)}$).

- 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 boundary ($l_{2,k}$) for the z-statistics calculated in subpopulation 2 ($Z_{2,k}$). The boundary for stage k is set equal to $l_{2,k} = f_{AD,2}\{(\sum_{k'=1}^{K} n_{k'})/(\sum_{k'=1}^{k} n_{k'})\}^{-\delta}$. If $Z_{2,k} \leq l_{2,k}$, we stop enrollment of subpopulation 2 (see section 1.4).

- H_{01} futility boundary proportionality constant for the adaptive design ($f_{AD,1}$): This is used to calculate the futility boundary ($l_{1,k}$) for the z-statistics calculated in subpopulation 1 ($Z_{1,k}$). The boundary for stage k is set to $l_{1,k} = f_{AD,1}\{(\sum_{k'=1}^{K} n_{k'})/(\sum_{k'=1}^{k} n_{k'})\}^{-\delta}$. If $Z_{1,k} \leq l_{1,k}$, we stop all enrollment (see section 1.4).

- 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 $f_{SC}\{(\sum_{k'=1}^{K} n_{k'})/(\sum_{k'=1}^{k} n_{k'})\}^{-\delta}$ 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 $f_{SS}\{(\sum_{k'=1}^{K} n_{k'})/(\sum_{k'=1}^{k} n_{k'})\}^{-\delta}$ in stage k.

- Lowest value to plot for treatment effect in subpopulation 2: **interAdapt** simulates performance metrics under a range of treatment effect sizes for subpopulation 2. This sets the lower bound for this range.

- Greatest value to plot for treatment effect in subpopulation 2: **interAdapt** simulates performance metrics under a range of treatment effect sizes for subpopulation 2. This sets the upper bound for this range.

4.2. Outputs

The output panel on the right the user interface is split into three sections, "About interAdapt", "Designs" output and "Performance" output. Users can navigate between these sections using the radio bottoms 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 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 participants to recruit by the end of each stage. The Performance section compares the three designs in terms of their power, expected sample size, and expected duration.

Designs

The Designs section 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 participants 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 $l_{2,k}$ (for statistics $Z_{2,k}$) in stages $k > k^*$ are not given, and l_{2,k^*} is set to Inf (indicating ∞) which means guaranteed stopping of subpopulation 2 for futility at stage k^*. Efficacy boundaries for $Z_{C,k}$ are not given for stages $k > k^*$. (However, 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 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 comparable 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 is unaffected by the choice of study population.

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

Performance output

interAdapt shows 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** 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 three plots can be accessed via the tabs at the top of the page. The table at bottom of the Performance section shows all three 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, and a decrease in expected trial duration. Conversely, if the true underlying treatment effect is significantly harmful, the trials will be more likely to stop early for futility. This also leads to small expected sample sizes, and shorter expected durations. 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 its 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 ($n^{(1)}$, $n^{(2)}$, n_{SS} and n_{SC}), increasing the number of stages (K), lowering the futility boundaries (f_{SC}, f_{SS}, $f_{AD,2}$, or $f_{AD,1}$), 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. This is as we expect, since SS does not take any data from subpopulation 2. The expected sample size and expected duration 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 participants 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}, $f_{AD,2}$, or $f_{AD,1}$), 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 participant is enrolled. Like expected sample size, the expected duration can be decreased by lowering K or k^*, increasing the futility bounds (f_{SC}, f_{SS}, $f_{AD,2}$, or $f_{AD,1}$), or relaxing α. Increasing the recruitment rate can also shorten the expected duration of a trial.

5. Example of entering input and interpreting output

The default inputs to **interAdapt** come from the motivating example of planning a 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). The MISTIE trial studied a new surgical treatment for stroke, and measured participant's outcomes by their 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 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, patients 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 patients, but no data had yet been collected to test this. Thus, we refer to the subpopulation of small IVH participants 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 patients;

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

participants;

(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 familywise Type I error rate (α) of .025.

Prior research by Hanley (2012) indicated that the proportion of participants with small IVH (π_1) was .33, that the probability of a positive outcome under control was .25 for small IVH participants (p_{1c}), and that the probability of a positive outcome under control was .2 for large IVH participants (p_{2c}). If the true treatment effect in subpopulation 1 was 12.5% then the probability of a positive outcome under treatment for participants 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 for the AD design come from the analysis section of (Rosenblum *et al.* 2013). There, the authors first fixed $K = 5$ and $\delta = -.5$, and then searched over a large class of values for remaining parameters in order minimize 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 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) (Figure 2). Although it is not shown, we know that the familywise 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.

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 patients 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.

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 Neurochir Suppl.*, **105**, 147–51.

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

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." *Pharmaceut. Statist.*, **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/

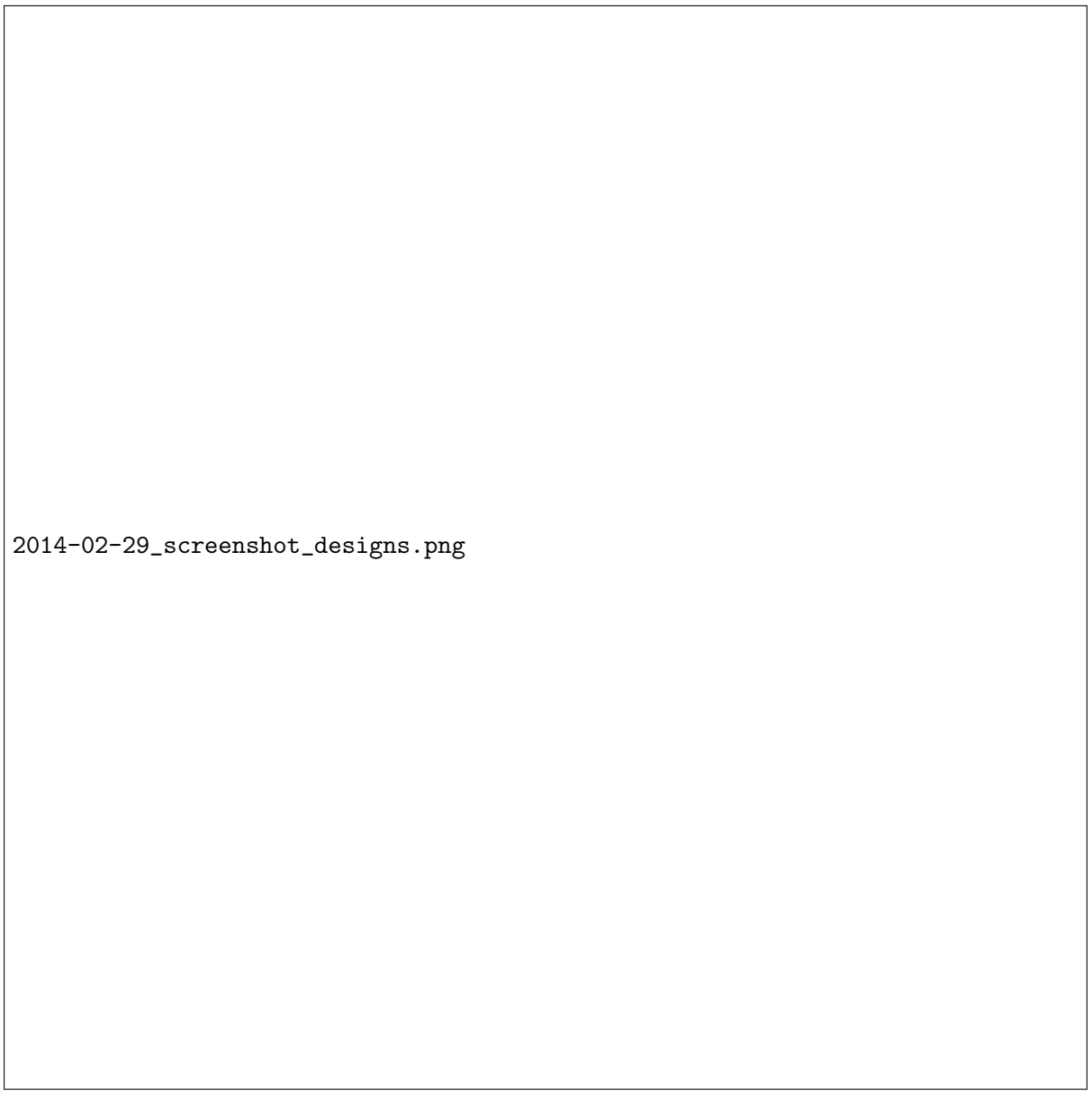

Figure 1: Designs Screenshot: Inputs can be entered in the side panel on the left, with results visible in the main panel on the right. The drop down menus at the top of the side panel can be used to navigate different interfaces for inputting parameters. Here we show the "Basic parameter" inputs, in "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 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 patients should be enrolled in each stage. Note that the scroll bar on the right of the web browser has been cropped out of this figure for the sake of increased screenshot resolution.

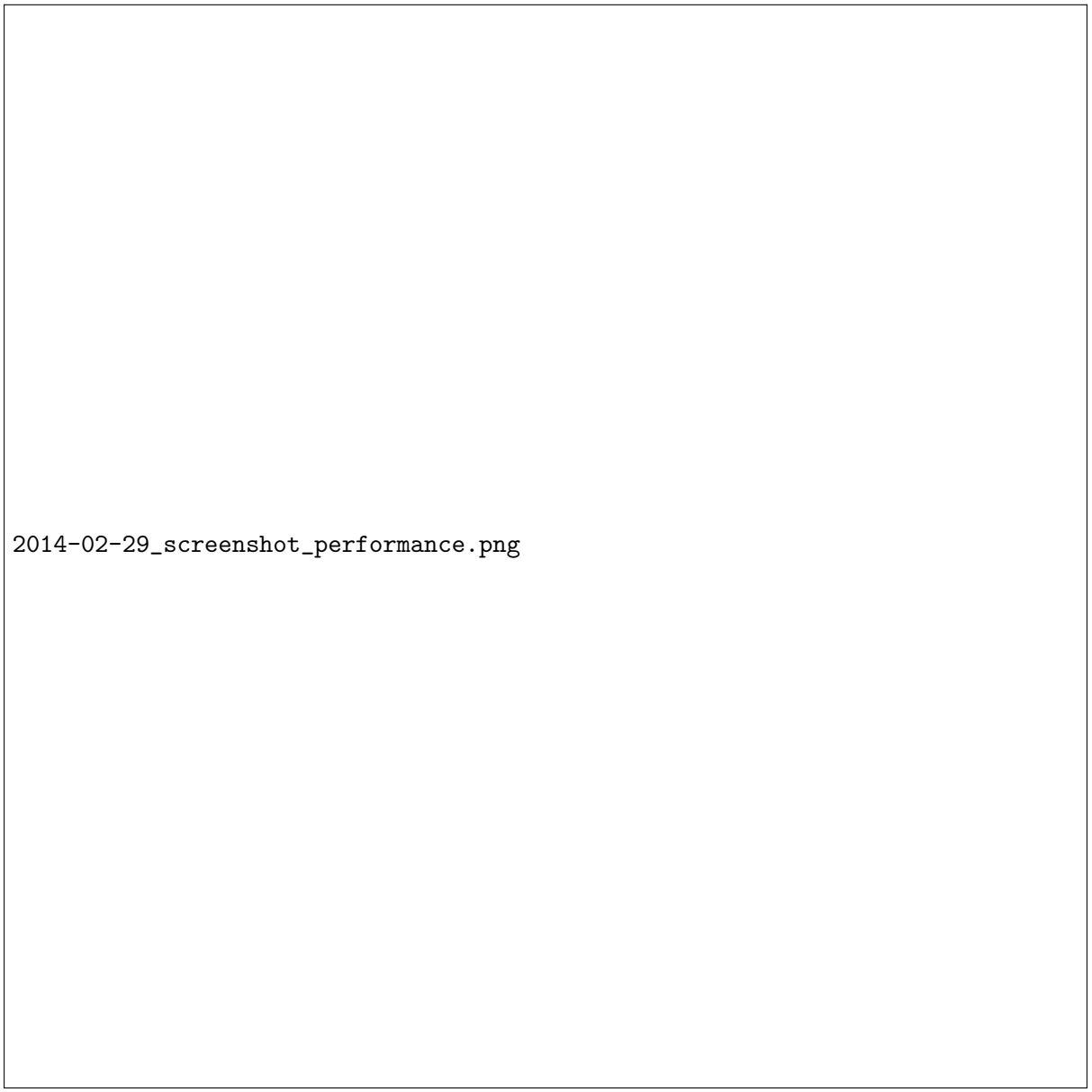

Figure 2: Performance Screenshot: Here the main panel shows performance output based on the default parameter inputs. The tabs at the top of the Performance section can be used to navigate between displays of power, expected sample size, and expected trial duration for all three designs. In the side 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.