EAGLE – An Interactive Tool for Designing Randomized Trials with Adaptive Enrollment Criteria

Abstract

We consider the problem of designing a randomized trial when there is prior evidence that the experimental treatment might work better in certain subpopulations. We implement novel trial designs with built in rules for whether to continue enrolling patients from each subpopulation based on data accrued at interim analyses. In order for the type I error and the power of the trial to be calculable, the decision rules for changing enrollment are all set before the trial starts. We introduce the EAGLE software, a tool for generating pre-determined decision rules for trial designs with adaptive enrollement criteria. The application compares the performance of the resulting adaptive designs to the performance of comparable standard group sequential designs. Performance is compared in terms of expected sample size, expected trial duration, and power, with family-wise type I error rate set to be constant (e.g. 0.05) for all trials compared. Unlike existing software, EAGLE is open-source and cross-platform.

1. Introduction

A group sequential trial design is one that incorporates pre-determined decision rules for stopping the trial early based on preliminary results. The treatment effect is analyzed at several stages throughout the trial, and each interim analysis may lead the trial to either stop early for treatment efficiacy, if there is strong evidence that the treatment is beneficial, or to stop early for treatment futility, if we have evidence the treatment is ineffective or harmful. In this paper, we will generally use the term "adaptive design" to refer to a group sequential design where enrollment criteria may also change based on these interim analyses. We use the term "standard design" to refer to a group sequential design where the enrollment criteria is fixed.

The motivating example for our work comes from the planning of a Phase III trial for a new surgical treatment of stroke (Rosenblum 2013)[4]. The new treatment is known as Minimally-Invasive Surgery Plus rt-PA for Intracerebral Hemorrhage (MISTIE), and is described in more detail in (Morgan 2008)[3]. was believed to be more effective patients under age 65. In this context, an adaptive design would first recruit patients from all ages, and then decide whether to continue recruiting patients over 65 based on the results of those currently enrolled. Comparable standard designs would either enroll patients from all ages throughout the trial, or only ever enroll patients under age 65.

In section 2 we give a formal definition of the problem addressed by these trial designs. In section 3 we present a brief overview of the currently available comparable software, none of which are open-source, cross-platfrom, or able generate these types of adaptive designs as efficiently as EAGLE does. In Section 4 we describe how to install EAGLE on a personal computer, as well as how to access it online through a web browser. Section 5 describes the inputs available when using EAGLE, and discusses the interpretation of the application's output. Section 6 presents an example of how an adaptive design can be created and analyzed with EAGLE.

2. Formal Description of the Problem

Consider the case where we have two subpopulations, referred to as subpopulation A and subpopulation B. Let subpopulation A be the subpopulation where we have prior evidence of a stronger treatment effect. Let p_A and p_B denote the proportion of patients in each of the two subpopulations.

Both the adaptive and standard designs discussed here consist of ongoing enrollment, and include rules for stopping the trial early based on interim analyses of currently enrolled patients. We discretize each trial into K stages, and say that the k^{th} stage will be completed once a pre-specified number of additional patients (n_k) have been enrolled. In stages when both subpopulations are being recruited,

we assume that $p_A n_k$ of the patients recruited in stage k are from subpopulation A, and $p_B n_k$ are from subpopulation B. An interim analysis is done at the end of each stage, which may lead us to stop the trial early if there is either strong evidence of treatment efficacy, or strong evidence of treatment futility.

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

For subpopulation A, denote the probability of a success under treatment as π_{At}, and the probability of a success under control as π_{Ac}. Similarly for population B, let π_{Bt} denote the probability of a success under treatment, and π_{Bc} denote the probability of a success under control. We define the average treatment effect for a given population as difference in the probability of a successful outcome between the treatment and control groups.

In the remained of this section we give an overview of the relevant concepts and notation needed to understand and use EAGLE. A more detailed discussion of the theoretical context, and of the parameter calculation procedure, can be found in (Rosenblum 2013)[4].

Hypotheses

We focus on testing the null hypothesis for a treatment effect in subpopulation A, and the null hypothesis for a treatment effect in the combined population. The two hypotheses are defined respectively as

- H_{0A}: $\pi_{At} - \pi_{Ac} \leq 0$
- H_{0C}: $p_A(\pi_{At} - \pi_{Ac}) + p_B(\pi_{Bt} - \pi_{Bc}) \leq 0$

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

Whenever any of the trials AD, SC or SA is stopped early, there will be some patients who have been enrolled but who's outcomes have not yet been measured. These patients are sometimes referred to as "overrunning" or "pipeline" patients.

Test Statistics

We calculate two z-scores at the end of each stage, one for the treatment effect in the combined population And one for the treatment effect in subpopulation A.

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

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

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

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

$$Z_{A,k} = \left[\frac{\sum_{k'=1}^{k} \sum_{i=1}^{n'_k} Y_{i,k'} T_{i,k'} A_{i,k'}}{\sum_{k'=1}^{k} \sum_{i=1}^{n'_k} T_{i,k} A_{i,k'}} - \frac{\sum_{k'=1}^{k} \sum_{i=1}^{n'_k} Y_{i,k'} (1 - T_{i,k'}) A_{i,k'}}{\sum_{k'=1}^{k} \sum_{i=1}^{n'_k} (1 - T_{i,k}) A_{i,k'}} \right] \times \left\{ \left(\frac{2}{\sum_{k'=1}^{k} \sum_{i=1}^{n'_k} A_{i,k'}} \right) (p_A[\pi_{Ac}(1 - \pi_{Ac}) + \pi_{At}(1 - \pi_{At})]) \right\}^{-1/2}$$

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

Family-wise Type I Error

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

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

Decision Rules for Stopping the Trial Early

In the SC trail, our decision rules consist of efficacy and futility boundaries for H_{0C}. At each stage k, we calculate the test statistic $Z_{C,k}$. If $Z_{C,k}$ is above the efficacy boundary for stage k, we reject H_{0C} and end the trial. If the $Z_{C,k}$ is between the efficacy and futility boundaries for stage k, we make no conclusion and continue the trial. If $Z_{C,k}$ is below the futility boundary for stage k, we end the trial with the conclusion that we have failed to reject H_{0C}.

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 $c^e_{S,C}\{(K-1)/k\}^{-\delta}$, where K is the total number of stages, δ is a constant in the range $[-.5, .5]$, and $c^e_{S,C}$ is the constant that ensures the desired family-wise Type I error rate. In order to calculate $c^e_{S,C}$, we make use of the fact that the random vector of test statistics $(Z_{C,1}, Z_{C,2}, \dots Z_{C,K})$ follows a multivariate normal distribution with a known covariance structure (Jennison and Turnbull, 1999, Chapter 3)[2]. Under H_{0C} we assume this vector has mean zero. Using 'mvtnorm' package in R to evaluate the multivariate normal distribution function, we find the proportionality constant $c^e_{S,C}$ such that the null probability of $Z_{C,k}$ exceeding $c^e_{S,C}\{(K-1)/k\}^{-\delta}$ for any stage k is less than or equal to α.

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

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

The decision boundaries for $Z_{A,k}$ in the SA design are defined by exactly the same form. The efficacy boundary for the k^{th} stage is set equal to $c^e_{S,a}\{(K-1)/k\}^{-\delta}$, where $c^e_{S,a}$ is the constant that ensures the appropriate type I error rate. The first $K-1$ futility boundaries for H_{0A} are again set equal to $c^f_{S,a}\{(K-1)/k\}^{-\delta}$, where $c^f_{S,a}$ 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.

Decision boundaries for AD vary from those of the standard designs two ways. First, because AD simultaneously tests H_{0C} and H_{0A} it must have two sets of decision boundaries rather than one. For the k^{th} stage of AD, let $u_{C,k}$ and $u_{a,k}$ denote the efficacy boundaries for H_{0C} and H_{0A} respectively, and let $l_{C,k}$ and $l_{a,k}$ denote the corresponding futility boundaries. The boundaries $u_{C,k}$ and $u_{a,k}$ are set equal to $c^e_{AD,C}\{(K-1)/k\}^{-\delta}$ and $c^e_{AD,S}\{(K-1)/k\}^{-\delta}$ respectively, where $c^e_{AD,C}$ is the constant that ensures a type I error rate less than $a_C\alpha$ for the test of H_{0C}, and $c^e_{AD,S}$ is the constant that ensures a type I error rate less than $(1-a_C)\alpha$ for the test of H_{0A}. The futility boundaries $l_{C,K}$ and $l_{a,K}$ set equal to $c^f_{AD,C}\{(K-1)/k\}^{-\delta}$ and $c^f_{AD,S}\{(K-1)/k\}^{-\delta}$ respectively, where $c^f_{AD,C}$ and $c^f_{AD,S}$ can be set by the user.

Our decision rules in AD consist of the following steps for each stage k:

(1) Assess Efficacy in the Combined Population: If $Z_{C,k} > u_{C,k}$, reject H_{0C} and stop all enrollment. If $Z_{A,k} > u_{a,k}$, reject H_{0A} as well.

(2) Assess Futility in the Combined Population: Else, if $Z_{C,k} \leq l_{C,k}$, stop enrolling from subpopulation B in all future stages. In this case when $Z_{C,k} > u_{C,k}$, the following additional steps must be done:

(a) If $Z_{A,k} > u_{a,k}$, we reject H_{0A} and stop all enrollment.

(b) If $Z_{A,k} \leq l_{a,k}$, we fail to reject either H_{0C} or H_{0A}, and stop all enrollment.

(c) Else, we continue to enroll from subpopulation A, and re-evaluate steps (2)-(3) at the end of the next stage.

(3) If $l_{C,k} < Z_{C,k} \leq u_{C,k}$, continue enrolling from both subpopulations.

The second way that the decision boundaries of AD differ from those of the standard designs is that we allow for more flexibility in the futility boundaries. EAGLE allows the user to specify a final stage in which we test for an effect in the total population. We denote this stage by k^*. If we reach k^* before stopping the trial, we will only always stop enrolling from subpopulation B in all future stages. We represent this in the context of our decision rules defined above by setting the futility boundary l_{C,k^*} equal to the efficacy boundary u_{C,k^*}. This ensures that Z_{C,k^*} will always be either greater than u_{C,k^*} or less than or equal to l_{u,k^*}.

EAGLE allows the user to specify two stage specific sample sizes, one for stages when both populations are enrolled ($k \leq k^*$), and one for stages where only patients in subpopulation A are enrolled ($k > k^*$). We refer to these two sample sizes as n^*_1 and n^*_k respectively.

3. ADDPLAN

4. Running EAGLE

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

Running EAGLE Over the Web

EAGLE is currently hosted on the RStudio webserver, and can be accessed simply visiting the link below.

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

Running EAGLE Locally

To run EAGLE locally, one must first install the R programming language, as well as the R packages "Shiny" and "mvtnorm." R runs on both Windows & MacOS. The most current version 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. To install the required versions of Shiny and mvtnorm, type the lines below into the R Console while connected to

the Internet. The return key must be pressed after each line of code. Both lines may cause R to give feedback on the installation progress, which we do not show here.

```
install.packages("shiny")
install.packages("mvtnorm")
```

To download the EAGLE application for offline use, first download the EAGLE zip file from the following link.

`https://github.com/aaronjfisher/eagle_stable_solo/archive/master.zip`

Unzip the archive to the directory where you want to store the app. Once EAGLE has been downloaded, the application can be run without an internet connection by the opening the R Console and typing the code below. In place of the characters {"Full/Path/Name/To/EAGLE_stable_solo"}, you will need to enter the full path name, in quotes, of the unzipped EAGLE folder.

```
#Change Working Directory
setwd("Full/Path/Name/to/EAGLE_stable_solo")

#Load Shiny Package
library(shiny)

#Run EAGLE
runApp()
```

Alternatively, to download EAGLE as a temporary file and open it for a single use, you can type the following code into the R console.

```
library(shiny)
runGitHub(repo="eagle_stable_solo",user="aaronjfisher")
```

5. User Interface

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

Inputs

Parameters in the input panel are organized into two sections, basic parameters and advanced parameters. To view 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, EAGLE 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 EAGLE to recalculate the results inbetween each individual change. In Interactive mode, EAGLE will automatically recalculate the results after each change, which can make it easier to quickly see the effect of changing one specific input parameter.

Switching between Batch mode and Interactive mode can be done using the dropdown menu at the top of the Basic Parameters section. Interactive mode is not available when entering advanced parameters.

To save the current set of inputs, select the dropdown menu titled "Show basic parameters" and select "Show all parameters". From here, you can save the current parameters as a csv file, or load a previously saved csv file of inputs. Regardless of whether EAGLE is being run online or locally, these csv files are always stored on the user's computer.

A detailed explanation of each input is given in Box 5.1

Box 5.1.1: Basic Parameters

- Subpopulation A proportion (p_A): The proportion of the population in subpopulation A. This is the subpopulation in which we have prior evidence of a stronger treatment effect.
- Probability outcome = 1 under control, subpopulation A (π_{Ac}): The probability of experiencing a successful outcome for control patients in subpopulation A. This is used in estimating power and expected sample size of each design.
- Probability outcome = 1 under control, subpopulation B (π_{Bc}): The probability of experiencing a successful outcome for control patients in subpopulation B. This is used in estimating power and expected sample size of each design.
- Probability outcome = 1 under treatment for sub-population A (π_{At}): The probability of experiencing a successful outcome for treated patients in subpopulation A. Note that a specific effect size is not specified for subpopulation B. Instead, EAGLE generates the relevant performance metrics for a range of several possible effect sizes in subpopulation B. This range can be specified in the Advanced Parameters section.
- Alpha (FWER) Requirement (α): The family-wise Type I error rate for all hypotheses in the trial. In AD, this is the probability of falsely rejecting either H_{0C} or H_{0A}. In SC it is the probability of falsely rejecting H_{0C}. In SA it is the probability of falsely rejecting H_{0A}.
- Proportion of Alpha allocated to H0C (a_C): In the AD design, the alpha level of the test for H_{0C} is set equal to $a_C\alpha$, and the alpha level for the test of H_{0A} is equal to $(1 - a_C)\alpha$.
- Per stage sample size, combined population (n_1^*): In SC, this is the number of patients that must be recruited before we end each stage to do an interim analysis (???). In AD it is the number of patients required for stages 1 through k^*.
- Per stage sample size for stages where only sub-population 2 is enrolled (n_k^*): In SA this is the number of patients that must be recruited before we end each stage to do an interim analysis (???). In AD, it is the number of patients required for each stage after k^*.

Box 5.1.2: Advanced Parameters

- Delta (δ): This parameter defines the curvature of the efficacy and futility boundaries, which are all proportional to $\{(K-1)/k\}^{-\delta}$.
- Number of Iterations for simulation: Z-statistics are simulated generate the power, expected sample size, expected trial duration, and expected number of overrunning patients. Generally, about 10,000 simulations are needed for reliable results. On our systems, a simulation with 10,000 iterations takes about 15 seconds.
- Time limit for simulation, in seconds: If the simulation time exceeds this threshold, calculations will stop and the user will get an error message saying that the application has "reached CPU time limit". To remove the error, either the number of iterations can be reduced, or the time limit for simulation can be extended. EAGLE does not allow for this time limit to exceed 90 seconds.
- Total number of stages (K): The total number of stages for all three designs.
- Recruitment rate in sub-population A: The number of patients recruited per year in sub-population A. This will affect the expect duration of the trial.

- Recruitment rate in sub-population B: The number of patients recruited per year in sub-population B. This will affect the expect duration of the trial.
- Lower bound for treatment effect in sub-population B: EAGLE simulates performance metrics under a range of treatment effect sizes for subpopulation B. This sets the lower bound for this range.
- Upper bound for treatment effect in sub-population B: EAGLE simulates performance metrics under a range of treatment effect sizes for subpopulation B. This sets the upper bound for this range.
- Last stage sub-population B is enrolled under an adaptive design (k^*): In the adaptive design, we don't enroll any patients from subpopulation B after stage k^*.
- H0C futility boundary proportionality constant for the adaptive design ($c^f_{AD,C}$): This is used to calculate the futility boundary for H_{0C} in the adaptive design, which is set to $c^f_{AD,C}\{(K-1)/k\}^{-\delta}$ in stage k.
- H0S futility boundary proportionality constant for the adaptive design ($c^f_{AD,S}$): This is used to calculate the futility boundary for H_{0A} in the adaptive design, which is set to $c^f_{AD,S}\{(K-1)/k\}^{-\delta}$ in stage k.
- H0C futility boundary proportionality constant for the standard design ($c^f_{S,C}$): This is used to calculate the futility boundary for H_{0C} in SC, which is set to $c^f_{S,C}\{(K-1)/k\}^{-\delta}$ in stage k.
- H0S futility boundary proportionality constant for the standard design ($c^f_{S,a}$): This is used to calculate the futility boundary for H_{0C} in SA, which is set to $c^f_{S,a}\{(K-1)/k\}^{-\delta}$ in stage k.

Outputs

The output section in the main panel of the user interface is split into two categories, "Design" output and "Performance" output. Design output gives a road plan for how to conduct each of the three trials: FA, AD and SC. This includes the efficacy boundaries; user specified non-binding futility boundaries, and number of patients to recruit by the end of each stage. Performance output compares the three designs in terms of their power, expected sample size, expected duration, and expected number of overrunning patients. The radio buttons at the top of the output section can be used to switch between a display of either the Design output or the Performance output.

Design Output

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

In the "Adaptive" tab, the table at the top of the page shows the required number of patients that must be recruited by the end of each stage. The table also gives efficacy and futility boundaries for H_{0C} and H_{0A}. Efficacy and futility boundaries for H_{0C} are not given however for stages after stage k^*, as we will have always either rejected or failed to reject H_{0C} by that point in the trial. A plot at the bottom of the page shows these efficacy and futility boundaries for H_{0C} and H_{0A} over the course of the trial.

The two tabs for the standard designs have a comprable layout. Note that since our efficacy boundaries are based on the null distributions of the z-statistics, which is the same regardless of whether the z-statistics come from a subpopulation or from the combined population, the efficacy boundaries for SA and SC will always be identical.

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

Performance Output – Layout & Interpretation

EAGLE shows performance of each of the three designs in terms of four metrics: power, expected sample size, expected duration, and expected number of overrunning patients. These metrics all depend, among other things, on the true treatment effect in each subpopulation. A treatment effect for subpopulation

A can be specified in the Basic Parameters section, and a range of values for the treatment effect in subpopulation B can be specified in the Advanced Parameters section. EAGLE 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 B. These four plots can be accessed via the tabs at the top of the page. The table at bottom of the output section shows all four metrics side by side, with each column of the table denoting a different treatment effect in subpopulation B.

When the true treatment is stronger, trials will be able to detect the treatment effect more easily, and will be more likely to stop early. This translates to an overall increase in power, a decrease in expected sample size, and a decrease in expected trial duration. Conversely, the number of overrunning patients actually increases when the underlying treatment effect is stronger. This is because it is precisely the process of stopping early that generates overrunning patients. Note that a significantly harmful underlying treatment effect will also correspond with a smaller expected sample size, a shorter expected duration, and a larger expected number of overrunning patients. These patterns are reflected in the plots shown by EAGLE.

The power plot shows the power of AD to reject H_{0C}, to reject H_{0A}, and to reject at least one of H_{0C} or H_{0A}. As the standard design SC only tests H_{0C}, we only show it's power to reject H_{0C}. Likewise, we only show the power of SA to reject H_{0A}. Note that the power of SC and AD to reject H_{0C} both increase as the treatment effect for subpopulation B increases. The power of AD to reject H_{0A} decreases as the treatment effect in subpopulation B increases, but this is only because AD does not bother to test H_{0A} 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^* and n_k^*), increasing the number of stages (K), lowering the futility boundaries ($f_{C,F}$, $f_{S,F}$, $f_{C,AD}$, or $f_{S,AD}$), or relaxing the required type I error rate (α).

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

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

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

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

6. Example of Entering Input and Interpreting Output

The default inputs to EAGLE come from the motivating example of the MISTIE Phase III trial. This trial studied a new surgical treatment for stroke, and measured patient's outcomes by their disability score on the modified Rankin Scale (mRS) 180 days after enrollment. A successful outcome was defined as a mRS score less than or equal to 3.

Treatment effect estimates from Phase II were 12.4% [95% CI: (-11.9%, 36.7%)] in patients under 65, and 7.7% [95% CI: (-12.8%, 28.2%)] for patients over 65. Thus, we refer to the subpopulation of patients under age 65 as subpopulation A, as they appeared more likely to benefit from the treatment.

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 patients both under age 65 and over age 65;

(b) The average treatment effect is 12.5% for patients under age 65, and zero for patients over 65;

(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_{0A} in scenario (b);

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

Prior research by (Hanley 2012)[1] indicated that the proportion of patients under age 65 (p_A) was .61, and that the probability of a positive outcome for control patients was .33 for those under 65 (π_{Ac}), and .12 for those under 65 (π_{Bc}). If the true treatment effect in subpopulation A is 12.5%, this suggests that the probability of a positive outcome under treatment for patients in subpopulation A (π_{At}) is approximately 12.5%+33%=45.5%.

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

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

Now we turn to the output of EAGLE 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_{0A} in scenario (b). SC has 80% power to reject H_{0C} in scenario (a), and SA has 80% power to reject H_{0A} in scenario (b). Although it is not shown, we know that the family-wise type I error rate is less than .025, as this was specified as an input to EAGLE.

Note that if we change k^*, holding all other inputs fixed, we're no longer able to satisfy the desired power goals for AD. Setting $k^* > 4$, means that we will tend to test H_{0C} for a longer period of time. Because H_{0A} is only ever tested after H_{0C} has been accepted or rejected, setting $k^* > 4$ decreases the power of the test for H_{0A} to below 80%. Setting $k^* < 4$ means that we tend to do fewer tests of H_{0C}, which lowers the power of the test for H_{0C} below 80%.

In general, EAGLE can be used to tweak the study design parameters to see which set of inputs will best meet the study coordinators goals. In our future work, we hope to create tools that automate this search process.

Summary

In this paper we introduce the EAGLE application for designing and simulating trials with adaptive enrollment criteria. We provide an overview of the theoretical problem the application addresses, and give an explanation of the application's inputs and outputs.

EAGLE improves on the commercially available software for design of adaptive enrichment trials by generating trials with lower expected sample sizes and equally controlled type I error rates. Not only that, but it is also cross-platform and open-source.

Acknowledgements

References

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

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

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

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