Journal of Data Science 21 (2), 412–427
April 2023
DOI: 10.6339/23-JDS1099
Statistical Data Science

Building a Foundation for More Flexible A/B Testing: Applications of Interim Monitoring to Large Scale Data

Wenru Zhou[1,*], Miranda Kroehl[2], Maxene Meier[2], and Alexander Kaizer[1]

[1]*13001 E 17th Pl, Aurora, CO 80045, Department of Biostatistics and Informatics University of Colorado, USA*

[2]*6380 S Fiddlers Green Cir, Greenwood Village, CO 80111, Charter Communication, USA*

Abstract

The use of error spending functions and stopping rules has become a powerful tool for conducting interim analyses. The implementation of an interim analysis is broadly desired not only in traditional clinical trials but also in A/B tests. Although many papers have summarized error spending approaches, limited work has been done in the context of large-scale data that assists in finding the "optimal" boundary. In this paper, we summarized fifteen boundaries that consist of five error spending functions that allow early termination for futility, difference, or both, as well as a fixed sample size design without interim monitoring. The simulation is based on a practical A/B testing problem comparing two independent proportions. We examine sample sizes across a range of values from 500 to 250,000 per arm to reflect different settings where A/B testing may be utilized. The choices of optimal boundaries are summarized using a proposed loss function that incorporates different weights for the expected sample size under a null experiment with no difference between variants, the expected sample size under an experiment with a difference in the variants, and the maximum sample size needed if the A/B test did not stop early at an interim analysis. The results are presented for simulation settings based on adequately powered, under-powered, and over-powered designs with recommendations for selecting the "optimal" design in each setting.

Keywords *A/B testing; error spending function; interim monitoring; stopping rule*

1 Introduction

When presented with accumulating data over the course of an experiment, it is recognized that multiple testing during the experiment, for instance through interim monitoring, will lead to inflated type I error rates (Armitage et al., 1969). However, methodology for controlling type I error rates has been developed so that an experiment can be stopped early if there is strong evidence of some difference and/or futility during an interim analysis and is commonly applied to biomedical clinical trials. Ad hoc rules attempt to ensure that study operating characteristics (e.g., power and type I error rates) are maintained through the implementation of interim analyses (Friedman et al., 2015). Group sequential tests proposed by Pocock (1977), O'Brien and Fleming (1979); Wang and Tsiatis (1987); Demets and Lan (1994); Jennison and Turnbull (1999) have all been incorporated into clinical research and maintain the desired study operating characteristics while incorporating interim evaluations of the data to determine if a study should

*Corresponding author. Email: wenru.zhou@cuanschutz.edu.

Received December 15, 2022; Accepted April 17, 2023

stop early for futility (i.e., not detecting any effect), the difference (i.e., finding superiority or inferiority), or both.

Ongoing evaluation of accumulating data is not uncommon in an industry setting for A/B testing, where the ability to make rapid decisions is paramount to company success. Kohavi et al. (2013) comprehensively and thoroughly discussed online A/B testing at large scale. In addition, Miller (2010, 2015); Koning et al. (2022); Azevedo et al. (2020) discussed novel methods in A/B testing to preserve power, control the overall type I error rate, or even meet the need of some special data distribution. However, the application of interim monitoring methods for controlling type I error rates is still not widespread in A/B testing in many companies, and the use of standard inference tools that do not account for repeated looks at the accumulating data can lead to incorrect conclusions.

One factor which might limit the utility of more sophisticated monitoring methods is that, within a single company, several dozen or even hundreds of experiments may be running at any given time, and human bandwidth inhibits the ability to apply customized design and analysis practices to each of these experiments. Another factor is that, though many novel statistical methods for A/B testing are developed, they may be too complicated to be implemented on a large scale by non-statisticians. To overcome this hurdle and make recommendations for a scalable A/B testing framework with desired statistical properties, it is important to have a more complete understanding of the performance of standard methods on commonly encountered scenarios.

In this paper we first review frequently used sequential monitoring boundaries, statistical approaches to analyzing A/B tests, and general study design considerations in Section 2. Section 3 then presents the simulation set-up and a novel loss function to use in selecting the "optimal" A/B test design by considering 16 possible combinations of group sequential methods and stopping criteria. The results of the simulations with general recommendations are summarized in Section 4. We conclude with a brief discussion in Section 5.

2 Background

Sequential monitoring designs have been developed and applied in the context of clinical research studies where regulatory agencies require strict control of the type I error rate α (i.e., concluding an effect when there is none) while trying to achieve acceptable statistical power (i.e., the ability to detect an effect if one exists). In the following subsections, we discuss approaches developed for interim monitoring that we will further examine in simulation studies for optimal A/B test designs.

2.1 Reasons to Stop Early

There are many reasons one may wish to terminate a study early, including for safety and efficacy. In general, for studies that compare groups and wish to detect a difference (e.g., an A/B test) we consider three potential types of stopping rules to use in an interim analysis:

1. Only stop for some detectable difference: In this situation, at each interim analysis, we determine if we should stop the study because there is evidence of a difference between our two variants in the A/B test. This may be more descriptively presented as stopping either for superiority/benefit or inferiority/harm caused by one variant with respect to the other.

2. Only stop for futility: In this situation, at each interim analysis we determine if we should stop the study because there is evidence that we are unlikely to detect a difference between our

two variants in the A/B test were the experiment to continue enrolling to its planned maximum sample size.

3. Stop for either a detectable difference or futility: In this situation, at each interim analysis, we could stop for either detecting some difference between variants or for futility to detect a difference based on the accumulating data within the experiment.

2.2 Methods for Interim Monitoring

Once one has considered "why" one wishes to stop an experiment early, we must select stopping boundaries that identify "how" this decision is made. The different approaches to boundaries described below represent various trade-offs to study flexibility, the expected trial sample size, and the overall maximum trial sample size.

2.2.1 Ad Hoc Rules

Ad hoc rules attempt to ensure the conservative interpretation of interim results. For example, over a total of K analyses Haybittle (1971) uses a large critical value for all interim tests (such as the standard normal test statistic $Z_i = 3.0$ for any ith interim analysis) and uses the conventional critical value at the final Kth test. This specific method is ad hoc so that no precise type I error is guaranteed. This is a precursor for methods developed to explicitly control the overall type I error rate.

2.2.2 Group Sequential Boundaries

One such family of methods designed to control the overall type I error rate is known as group sequential tests, which have predetermined stages for evaluating the data for each desired interim analysis. For example, Pocock (1977) sets a constant and conservative critical value Z_{PO} for every interim analysis so that the overall significance level for the experiment will be α. Similarly, O'Brien and Fleming (1979) use critical value $Z_{OF}(\alpha, K)\sqrt{i/K}$ where $Z_{OF}(\alpha, K)$ is determined to control the overall type I error.

Wang and Tsiatis (1987) demonstrated that Pocock and O'Brien and Fleming are both special cases of a unified test where the critical value is defined as $Z_{WT}(\alpha, K, \delta)(i/K)^{\delta-0.5}$ where $Z_{WT}(\alpha, K, \delta)$ is determined to control the overall type I error. When $\delta = 0$, O'Brien and Fleming error spending function is produced. When $\delta = 0.5$, Pocock error spending function is produced. δ may also be set between the Pocock and O'Brien-Fleming boundaries, where intermediate shapes are produced.

2.2.3 Error Spending Functions

One major limitation of predetermined group sequential boundaries is that the number of interim analyses must be fixed in advance. If an additional interim analysis is requested or does not meet the predetermined analysis plan, the trial operating characteristics may not be maintained. To address this limitation, error spending functions were proposed by Demets and Lan (1994). In this approach, the type I error rate can be allocated flexibly across interim analyses throughout the study, so that at the end of the study the overall type I error is still controlled at the desired type I error rate, α. While it is still ideal to predetermine the expected number of interim analyses, error spending functions can facilitate unexpected interim looks at the data and unequal accrual throughout a study.

The error spending function $\alpha(t^*)$ is a function of t^*, the information fraction observed at the time of the interim analysis. t^* is generally defined as the ratio of the inverse of the variance of the test statistics at a particular interim analysis and at the final analysis (Gordon Lan et al., 1994). Practically, it is estimated by the fraction of participants enrolled at calendar time t divided by the maximum number of participants planned for at the end of the study. For example, when calendar time $t = 0$, the information fraction $t^* = \theta$ and the error spending function $\alpha(t^* = \theta) = 0$. When the study ends, the information fraction $t^* = 1$ and the error spending function $\alpha(t^* = 1) = \alpha$.

In the context of error spending function, Pocock boundaries can be approximated by the function $\alpha ln[1 + (e - 1)t^*]$, and for O'Brien-Fleming boundaries the approximate function is $2 - 2\theta(Z_{1-\alpha/2}/\sqrt{t^*})$. The power family of functions is another approach for interim monitoring proposed by Jennison and Turnbull (1999) that is defined as $\alpha t^{*\rho}$, where $\rho > 0$. For these error spending functions, they are equal to zero when $t^* = 0$ (i.e., no data has been observed) and equal to α when $t^* = 1$ (i.e., all data has been observed).

Examples of the boundaries of the different error spending functions discussed are presented in Figure 1 for a study that considers stopping for either futility or detecting some difference based on four total analyses with equal sample sizes enrolled in each stage. The statistical test statistic presented on the y-axis is on the standardized Z-scale (i.e., a normal distribution with mean 0 and standard deviation 1). To illustrate how these boundaries would be used in practice, assume we are comparing a binary outcome between two variants A and B so that $p_A - p_B$, where a positive difference indicates variant A performs better than variant B. At each interim stage of the A/B experiment, we may conclude one of four outcomes for a two-sided hypothesis test:

- if the Z-score falls in area 1 in Figure 1, the null hypothesis of no difference between variants is rejected, and we can conclude that we stop for the superiority of variant A.
- if the Z-score falls in area 5 in Figure 1, the null hypothesis of no difference between variants is also rejected, but this time we stop for the inferiority of variant A, concluding that variant B is better.

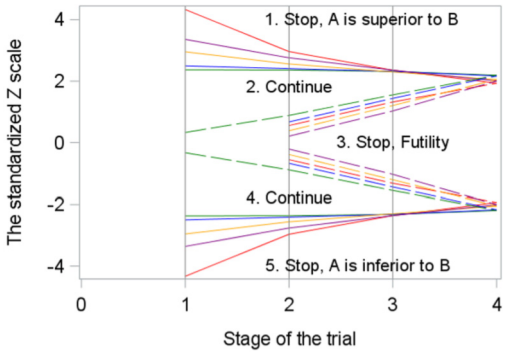


Figure 1: Example of the stopping boundary shapes for different error spending functions with three interim analyses. The red line represents O'Brien-Fleming boundary; the purple line represents Power function ($\rho = 3$); the yellow line represents Power function ($\rho = 2$); the blue line represents Power function ($\rho = 1$); The green line represents the Pocock boundary. The solid lines represent the boundaries of stopping for difference, and the dashed lines represent the boundaries of stopping for futility.

- if the Z-score falls in area 3 (the "inner wedge"), we fail to reject the null hypothesis and cannot conclude that variants A and B are different, and we still stop the study for futility.
- if the Z-score falls in area 2 or 4, we do not draw any conclusion and continue the study to the next stage.

2.3 Statistical Tests to Evaluate Outcomes

In many A/B tests, the outcomes may be represented as a binary variable (e.g., yes/no). When analyzing dichotomous outcomes between two variants, the chi-squared test without Yates' continuity correction (χ^2) or the chi-squared test with Yates' continuity correction (χ_c^2) would be natural choices. For large sample sizes, the chi-squared test becomes asymptotically equivalent to a two-sample Z-test. Based on this asymptotic equivalence with larger sample sizes, many A/B tests with dichotomous outcomes may instead apply the two-sample t-test to compare the two variants since the t-distribution becomes increasingly normal as the sample size increases. Others have previously discussed the different behaviors of statistical methods among various significant levels. D'agostino et al. (1988) concluded that for significance levels of 0.02 and 0.01, χ^2 test performs better than the t-test. For significant levels 0.1 and 0.05, the t-test performs better than the χ^2 test.

3 Simulation Study Design and Evaluation

3.1 A Proposed Loss Function to Identify "Optimal" Designs

To select among multiple candidate designs that facilitate various combinations of stopping boundaries (e.g., Pocock and power family) and rules (e.g., stopping for futility only, difference only, or both), in this subsection we propose a loss function to identify what is the "optimal" design. The loss function for each boundary is a linear combination of the weighted ratio of designs relative to the fixed sample designs based on their expected sample size under the null hypothesis ($ESS_{null,boundary}$), the expected sample size under the alternative hypothesis ($ESS_{alt,boundary}$), and the maximum sample size ($MSS_{boundary}$) if the study does not stop early:

$$L_1 = w_1 \frac{ESS_{null,boundary}}{SS_{fixed}} + w_2 \frac{ESS_{alt,boundary}}{SS_{fixed}} + w_3 \frac{MSS_{boundary}}{SS_{fixed}} \tag{1}$$

where $w_1+w_2+w_3 = 1$, and SS_{fixed} is the sample size of fixed sample size design. The "boundary" in the loss function refers to any of the fifteen stopping boundaries rather than the fixed design. The optimal design is one that minimizes L_1. With the fixed sample design as the comparator in the denominator based on sample size, any strategy with $L_1 < 1$ indicates an improvement over no interim monitoring. An advantage of this loss function is that the w_i can be customized for a given study to identify what the "optimal" design for the A/B test is based on the emphasis placed on the expected and maximum sample sizes.

One strength of this loss function is its adaptability to company goals. Departments may adjust the weights according to their objectives, such as reducing the maximum sample size by increasing w_3 to remain within budget constraints or increasing w_2 to decrease the expected sample size under the alternative hypothesis when testing a new variant. If a department does not have a specific preference for minimizing a particular type of sample size, they may select the stopping boundary with the minimum loss function value for most weight combinations. By leveraging our loss function, companies can experiment with various weight combinations

and determine the stopping boundary that minimizes the loss function value, allowing them to achieve their desired outcome.

To illustrate the use and interpretation of this loss function more clearly we provide two examples. When designing an A/B test, if one wants to minimize the maximum sample size, they can set $w_3 = 1$, and $w_1 = w_2 = 0$. In this context, the loss function becomes:

$$L_1 = \frac{MSS_{boundary}}{SS_{fixed}} \tag{2}$$

This means the design with the smallest maximum sample size among all maximum sample sizes from other designs will minimize the loss function and be selected as "optimal" by implementing this weighting for the loss function.

Another example can be shown when one is agnostic on how to split the weights. In this situation, the weights can be equally assigned to each component and the loss function becomes:

$$L_1 = \frac{1}{3}\frac{ESS_{null,boundary}}{SS_{fixed}} + \frac{1}{3}\frac{ESS_{alt,boundary}}{SS_{fixed}} + \frac{1}{3}\frac{MSS_{boundary}}{SS_{fixed}} \tag{3}$$

In this situation, the design with the smallest sum of expected sample sizes under the null and alternative hypothesis, and maximum sample size among all other designs will minimize the loss function, and thus be selected as the "optimal design". From the two examples, it can be seen that the optimal design may change when different weights are specified in the loss function depending on what the study team believes is important.

3.2 Simulation Design

A common A/B scenario is simulated to compare the proportion responding in the "A" variant (θ_A) and the "B" variant (θ_B) under the null hypothesis of no difference versus the alternative hypothesis that there is some difference in variants. In our motivating industry context, t-tests are most frequently used for A/B testing experiments, regardless if the outcome is continuous or binary. In addition, Zhou et al. (2023) demonstrates that when the sample size per arm is at or above 500, the t-test and the chi-squared test for two proportions comparison have nearly identical power, type I error rates, and expected sample sizes, even when interim analyses are incorporated. Therefore, a two-sample two-sided t-test is used as our primary benchmark. We also considered the chi-squared test χ^2 and Yates's chi-squared test χ^2_c, however nearly similar results to the t-test were observed and are not presented here.

Five different stopping boundaries (O'Brien-Fleming, Pocock, and power families with $\rho =$ 1, 2, and 3) are evaluated under three stopping strategies (futility only, difference only, or both for some difference or futility), for a total of fifteen combinations. A sixteenth approach is considered with no interim monitoring to reflect that some contexts may not optimally benefit from stopping early. The effects of increasing the number of interim looks at the data are examined across simulations with 1-, 3- or 19-interim analyses for a maximum number of 2-, 4-, or 20-looks at the data, respectively.

Assuming a constant response in variant "A" of $\theta_A = 0.5$, five different effect sizes are simulated for θ_B: 0.589 (large effect), 0.528 (moderate effect), 0.509 (small effect), 0.504 (tiny effect), and 0.500 (no effect). These effects were driven to reflect A/B tests that would enroll approximately 500, 5,000, 50,000, or 250,000 per variant to detect the decreasing effect sizes, respectively, in a fixed sample design without interim monitoring. However, in practice, stakeholders may either request a larger sample size than deemed necessary by a statistical power

Table 1: Simulation settings.

$\theta_A - \theta_B$ SS per arm	Alternative scenarios: $\theta_B > \theta_A = 0.5$				Null scenarios: $\theta_B = \theta_A$
	0.089	0.028	0.009	0.004	0
500	Adequately Powered	Under-powered	Under-powered	Under-powered	For type I error
5000	Over-powered	Adequately Powered	Under-powered	Under-powered	For type I error
50,000	Over-powered	Over-powered	Adequately Powered	Under-powered	For type I error
250,000	Over-powered	Over-powered	Over-powered	Adequately Powered	For type I error

calculation or alternatively be limited by external factors and are unable to enroll the necessary sample size. To address these potential settings, we also examine the choice of optimal interim monitoring strategy when a study is under- or over-powered. The simulation design and evaluation are shown in Table 1 and Figure S7 in the supplementary material.

We conducted a total of 10,000 simulated studies in R v4.2.0 (Vienna, Austria) for each combination of effect size and stopping boundary, assuming equal accrual between each interim analysis. We determined the stopping boundaries and sample size required to detect a given effect size for sequential designs using PROC SEQDESIGN in SAS (Cary, North Carolina). Subsequently, we calculated key statistics, including the effective sample size under the null hypothesis ($ESS_{null,boundary}$), effective sample size under the alternative hypothesis ($ESS_{alt,boundary}$), maximum sample size ($MSS_{boundary}$), power, and type I error.

3.2.1 Approaches to Determine Early Stop

For example, in 2-total analysis, approximately 500 participants per arm, O'Brien-Fleming with early stop for both has stopping boundary at the 1st analysis (260 per arm): 0.00154, 0.28149, 0.71851, 0.99846. This means that, if the one-sided p-values from simulated studies fall below 0.00154 or above 0.99846, those studies will stop early and claim a difference between B and A. If the one-sided p-values fall between 0.28149 and 0.71851, those studies will stop early and claim that there is a lack of evidence to show that B and A are different. For other p-values, studies will continue to the final analysis.

The boundary at the final analysis (519 per arm) is 0.02651 and 0.97349. If the p-values from simulated studies fall below 0.02651 or above 0.97349, those studies will stop early and claim a difference between B and A. If the p-values fall between 0.02651 and 0.97349, those studies will claim that there is a lack of evidence to show that B and A are different.

3.2.2 Approaches to Calculate Key Statistics

To calculate the $ESS_{null,boundary}$, we extracted scenarios with $\theta_B = \theta_A = 0.5$ for 16 stopping boundaries and computed the average sample sizes for all 16 stopping boundaries among 10,000 simulated datasets. Similarly, to calculate $ESS_{alt,boundary}$, we extracted scenarios with

$\theta_B = 0.589, 0.528, 0.509, 0.504$ for 16 stopping boundaries and computed the average sample sizes for all 16 stopping boundaries among 10,000 simulated datasets for each effect size.

$MSS_{boundary}$ was determined as the maximum sample size that could be attained if no early stop occurred during the study with interim analysis. We calculated the power for each simulated study by extracting scenarios with $\theta_B = 0.589, 0.528, 0.509, 0.504$ for 16 stopping boundaries and computing the proportion of studies that successfully claimed B was different from A (either superior or inferior since we used two-sided t-test) among 10,000 simulated datasets for all 16 stopping boundaries. Similarly, we determined the type I error for each simulated study by extracting scenarios with $\theta_B = \theta_A = 0.5$ for 16 stopping boundaries and computing the proportion of studies that claimed B was different from A (either superior or inferior) among 10,000 simulated datasets for all 16 stopping boundaries.

For each simulation scenario, we identify what would be chosen as the "optimal" design for 5151 unique combination of weights across settings where our restriction $w_1 + w_2 + w_3 = 1$ is met with weights defined across a grid from 0 to 1 in increments of 0.01. Since there are three weight components, we present the results graphically in a 2-D plot that is colored by the design considered optimal for each weight combination. To further generalize the optimal stopping rules, we also present a 2-D plot where we ignore the specific stopping boundary type and present if the optimal design recommends no interim stopping, stopping for futility only, stopping for difference only, or stopping for either futility or difference. The step-by-step process of how plots are generated is presented in the supplementary materials: An example to illustrate how loss functions are calculated and plotted.

4 Results

In this section, we present the results for what is selected as the "optimal" design based across our different simulation scenarios. Given that the conclusions are similar across scenarios (adequately-, under-, or over-powered) and number of total analyses (i.e., 2-, 4-, and 20-total looks), we present a subset of scenarios with 4-total looks in this section with complete results in the supplementary materials. The $ESS_{null,boundary}$, $ESS_{alt,boundary}$, and $MSS_{boundary}$ of each stopping boundary are summarized in Table S1-S12 in the supplementary.

4.1 Adequately Powered Simulation Scenarios

4.1.1 Optimal Boundaries

The stopping boundaries that minimized the loss function for each set of loss function weights w_1, w_2, and w_3 were plotted for each of the four adequately powered effect size scenarios with the percentage of every boundary selected as optimal among the 5151 weight combinations is presented in Figure 2. To illustrate how an "optimal" design is chosen for each combination of weights, Table 2 provides the estimated loss function value if we set $w_1 = w_2 = 0.33$ and $w_3 = 0.34$ for the scenario with a small effect size ($n = 50,000$ per variant in the fixed sample design). In this example, the O'Brien-Fleming design that allows stopping for both futility or a difference had the smallest loss function value ($L_1 = 0.876$), therefore it was selected as optimal based on this weight combination (i.e., see the $\oplus$ in Figure 2). Figure 2 also showed that: Near the area $w_1 = w_2 = 0.33$, there was a large black-colored region, which means that O'Brien-Fleming was also selected as optimal for other combination of w_1 and w_2 near 0.33. Specifically, the O'Brien Fleming boundary that allows stopping for both is selected as optimal 24.75% among

Table 2: Equal weights for the adequately powered scenario with small effect size (50,000 per variant), 4-total analysis.

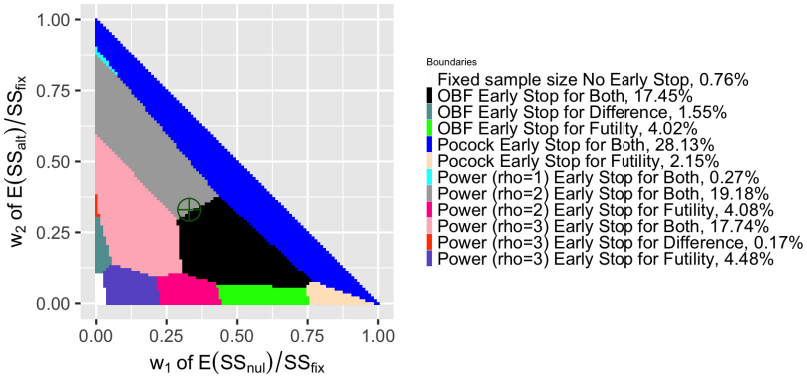
Error spending function	Stopping rule	ESS_{null}	ESS_{alt}	MSS	L_1
O'Brien-Fleming	Stop for Both	36702	40789	53661	0.876
Power (rho = 2)	Stop for Both	38927	38601	54088	0.879
Power (rho = 3)	Stop for Both	40401	39996	52049	0.884
Power (rho = 1)	Stop for Both	38185	37808	59251	0.904
Power (rho = 2)	Stop for Futility	38115	49485	51456	0.927
Power (rho = 1)	Stop for Futility	36470	50025	52801	0.929
O'Brien-Fleming	Stop for Futility	36655	50068	52827	0.931
Pocock	Stop for Both	35345	37150	68026	0.940
Power (rho = 3)	Stop for Futility	42143	49430	50340	0.946
Power (rho = 3)	Stop for Difference	51017	41066	51390	0.956
O'Brien-Fleming	Stop for Difference	50724	41954	51024	0.958
Power (rho = 2)	Stop for Difference	52288	40133	52837	0.968
Pocock	Stop for Futility	34397	53413	58933	0.979
Fixed sample size	No Early Stopping	50042	50042	50042	1.000
Power (rho = 1)	Stop for Difference	55859	39948	56900	1.018
Pocock	Stop for Difference	58570	40477	59865	1.060

Note: $w_1 = w_2 = 0.33$, $w_3 = 0.34$. Take O'Brien-Fleming with stop for both for example: (0.33 x 36702 + 0.33 x 40789 + 0.34 x 53661)/50042 = 0.876

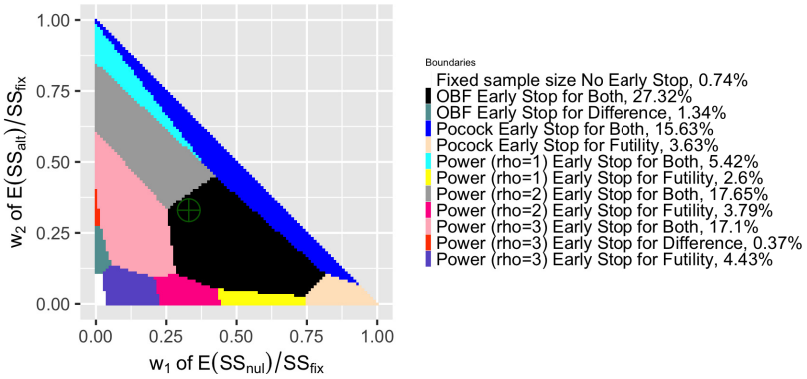
all 5151 weight combinations.

More generally, from Figure 2 the fixed sample size is only the "optimal" design if most of the weight is placed on the maximum sample size (i.e., a large w_3 value) across all effect sizes. If the weight of $ESS_{alt,boundary}$ was set near 0 (e.g., $w_2 < 0.05$), the optimal designs for various weights on $ESS_{null,boundary}$ (w_1) favor designs that only stop for futility. In contrast, if $w_1 < 0.05$ and $w_2 < 0.4$, many optimal designs favor stopping for the difference. As w_2 increases, many optimal designs start to favor stopping for both futility and detecting any difference. If similar values were given to all three weights, all optimal designs favor stopping for both futility and difference, with the O'Brien-Fleming boundary being optimal with the power boundary with $\rho = 2$ and $\rho = 3$ also near this weight combination. As shown in Table 2, $w_1 = w_2 = 0.33$ and $w_3 = 0.34$, the loss function values of O'Brien-Fleming, power boundary $\rho = 2$ and 3 were all between 0.87 to 0.89. While there are subtle differences across the scenarios in Figure 2, the general trends are largely the same for each adequately powered study design.

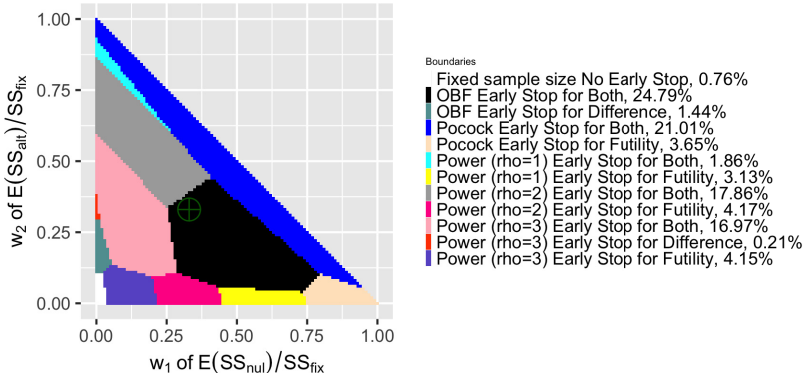
It is worth noting, some designs were never or rarely chosen as optimal in some scenarios. For example, the Pocock boundary and Power ($\rho = 1$) stopping for only a difference were never selected as optimal across all scenarios. Results, as noted previously, were similar if we had 2- or 20-total looks at the data. On exception is that the power ($\rho = 2$) that allows stopping only for a difference was selected as optimal for a very small range of w_1, w_2, and w_3 when there are 2-total analyses, but not for any adequately powered design with 4- or 20-total analyses.



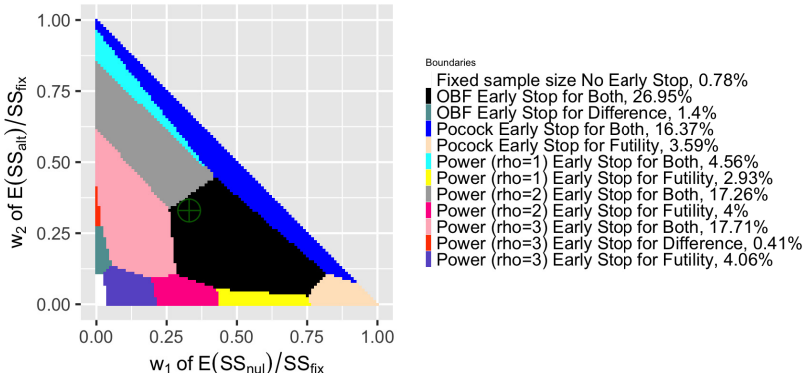
(a) large effect size



(b) moderate effect size



(c) small effect size



(d) tiny effect size

Figure 2: Optimal boundaries for four effect sizes from the 4-total analysis. Upper left: large effect size ($n = 500$ per variant in a fixed sample design); Upper right: moderate effect size ($n = 5000$ per variant in a fixed sample design); Lower left: small effect size ($n = 50,000$ per variant in a fixed sample design); Lower right: tiny effect size ($n = 250,000$ per arm in a fixed sample design). The $\oplus$ represents the point when the three loss function weights are equally allocated, OBF is O'Brien-Fleming.

4.1.2 Optimal Stopping Rules

While the specific design boundaries are important in selecting the truly "optimal" design based on the chosen loss function weights, it is also helpful to generalize the results in Figure 2 to summarize the broad stopping rules (i.e., no early stopping, stopping for only futility, stopping for only difference, or stopping early for both) to understand the potential reasons to stop a study early. In Figure 3, the optimal stopping rules for the four different adequately powered effect size scenarios are presented.

Among the four considered stopping rules, early stopping for both was selected as optimal in over 80% of weight combinations. Only when $w_2 < 0.1$, is early stopping for futility favored. Stopping for the difference was the least optimal except for small w_1 and w_2 around 0.25. Only when w_3, the maximum study sample size, was given the most weight was a fixed sample size design with no early stopping favored. These findings suggest that most adequately powered studies would be optimal when allowing stopping for both futility or a difference, except under A/B studies with fairly imbalanced loss function weights.

4.2 Over-powered Simulation Scenarios

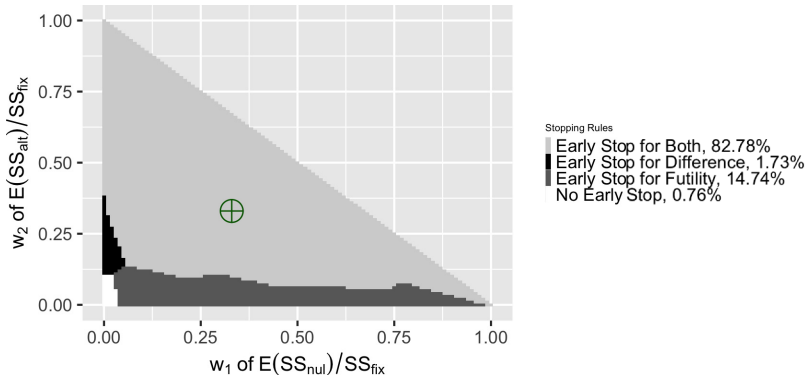
In some contexts, a stakeholder may wish to implement an intentionally over-powered design if sufficient resources are available. Figure 4a and 4c presents the results for an overpowered study to detect our moderate effect where we enroll 50,000 per variant instead of the 5000 needed for the fixed sample design to be adequately powered.

The patterns for overpowered scenarios were very similar to the patterns in adequately powered scenarios. The fixed sample size would still only be the best option when most of the weight was put on the maximum sample size (w_3). If the weight of $ESS_{alt,boundary}$ was set near 0 (e.g., $w_2 < 0.05$), the optimal designs for various weights on $ESS_{null,boundary}$ (w_1) still favored designs that only stop for futility. Further, if similar values were given to all three weights, all optimal designs favored stopping for both futility and difference, with the O'Brien-Fleming (black, 20.38 %), and power boundaries with $\rho = 2$ (grey, 26.17 %) and $\rho = 3$ (light pink, 32.91 %), selected as optimal. For overpowered scenarios, designs that stop only for detecting a difference were selected if $w_1 < 0.05$, which is less often than in adequately powered designs.

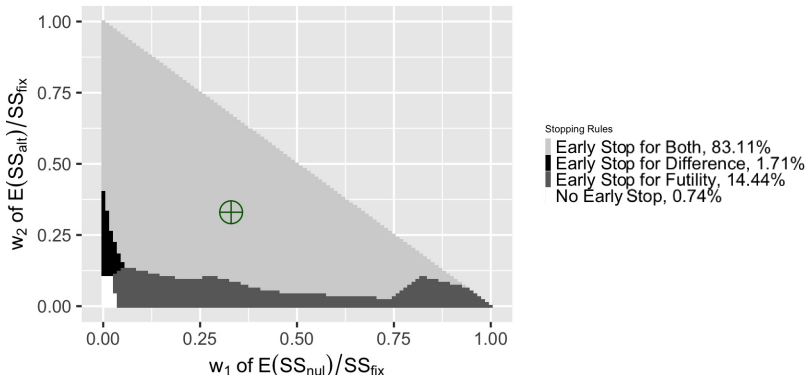
4.3 Under-powered Simulation Scenarios

In other contexts, a stakeholder may not be able to enroll the necessary sample size for an A/B test but still desires to implement an under-powered experiment. Figure 4b and 4d present the results for the tiny effect size if only 50,000 per variant are enrolled instead of the needed 250,000.

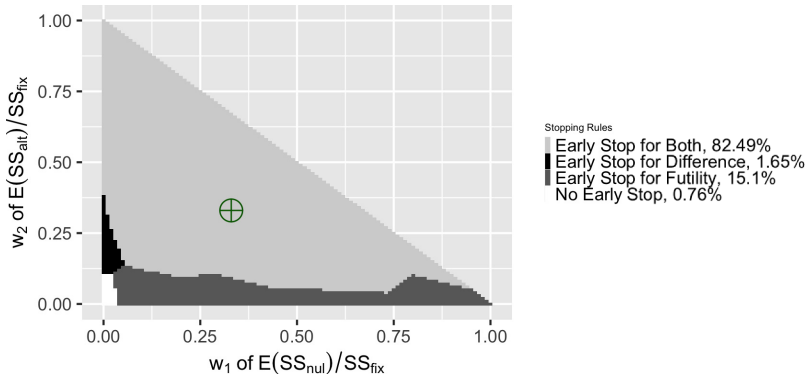
The pattern of stopping boundaries in Figure 4b and 4d were very different from the pattern in the adequately powered design. Given that we are intentionally running an under-powered A/B test, approximately 60% of the weight combinations identify stopping for futility only. This intuitively makes sense, because we are implementing an intentionally under-powered study that is unlikely to detect the desired difference. However, as w_2 increased above 0.5, weight combinations began favoring optimal designs the stop for both futility and a difference. In practice, the choice of "optimal" designs for under-powered studies may require additional considerations about not stopping for futility since the design is expected to be futile.



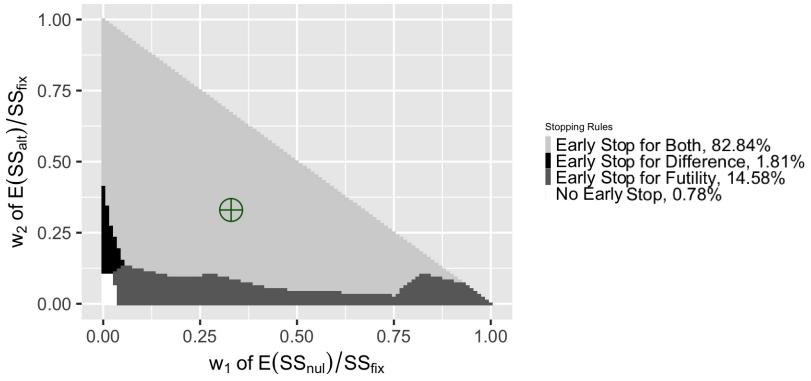
(a) large effect size



(b) moderate effect size

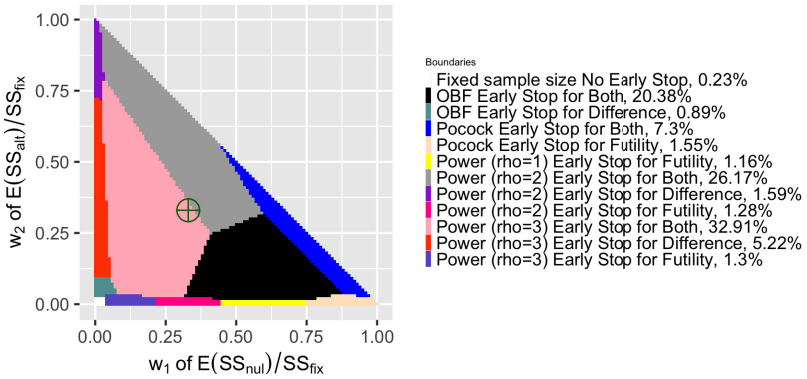


(c) small effect size

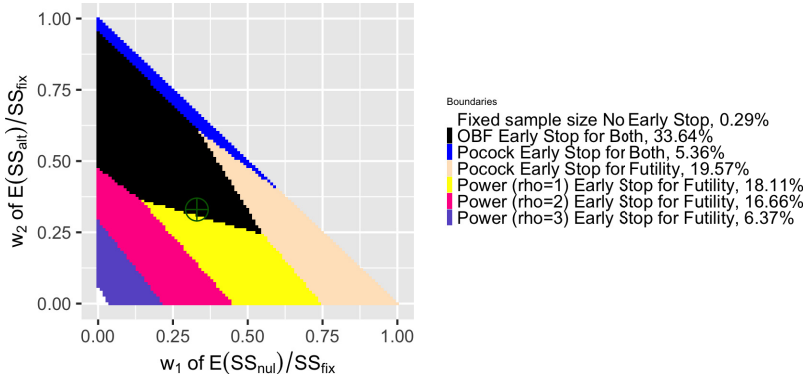


(d) tiny effect size

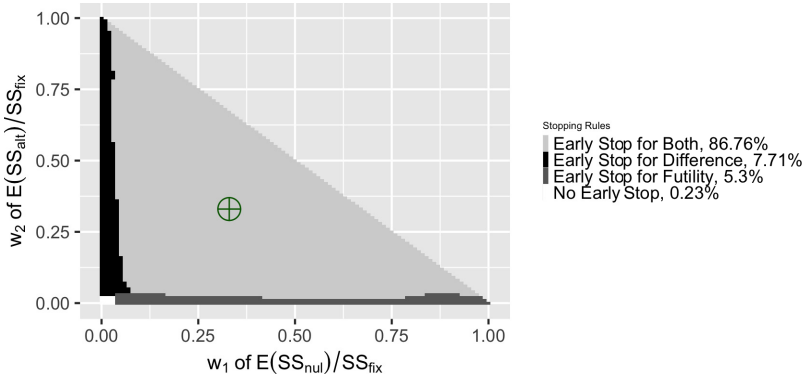
Figure 3: Optimal stopping rules for four effect sizes from the 4-total analysis. Upper left: large effect size ($n = 500$ per variant in a fixed sample design); Upper right: moderate effect size ($n = 5000$ per variant in a fixed sample design); Lower left: small effect size ($n = 50,000$ per variant in a fixed sample design); Lower right: tiny effect size ($n = 250,000$ per arm in a fixed sample design). The $\oplus$ represents the point when the three loss function weights are equally allocated.



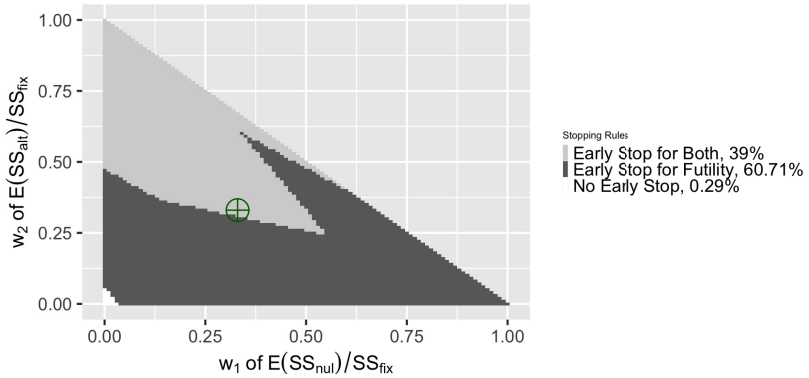
(a) over-powered design, optimal boundaries



(b) under-powered design, optimal boundaries



(c) over-powered design, optimal stopping rules



(d) under-powered design, optimal stopping rules

Figure 4: Optimal stopping boundaries and rules for 4-total analysis. (a), (c) Over-powered design, 50,000 per variant in the fixed sample design for moderate effect size which only needs 5000 per variant in the fixed sample design. (b), (d) Under-powered design, 50,000 per variant in the fixed sample design for tiny effect size which needs 250,000 per variant in the fixed sample design. The $\oplus$ represents the point when the three loss function weights are equally allocated, OBF is O'Brien-Fleming.

5 Discussion

The intention of this article was to evaluate the feasibility of applying the existing sequential monitoring methodology developed primarily for use in clinical trials to the setting of A/B experimentation in large-scale environments. After first reviewing some of these methods for interim monitoring, we implemented a rigorous simulation study to determine if general guidance could be given for A/B experimentation. The effect of decreasing or increasing the number of interim analyses on the data was examined as well. Given the large number of potential designs one could consider, we also proposed a novel loss function that evaluated the sample size demands under expected and maximum values.

In terms of general approaches to "optimal" designs in A/B test designs, we recommend based on our simulation results for adequately powered studies that designs with sequential monitoring that allow stopping for both some detectable difference between variants or for futility could be used across most combinations of weights for the loss function. When no strong preference exists for minimizing either the expected sample size or maximum sample size we recommend that it may be most efficient to use O'Brien-Fleming boundaries that allow stopping for both. Since the power boundary with $\rho = 2$ or $\rho = 3$ have loss function values similar to the O'Brien-Fleming, they could also be the other two good choices, but in practice we have seen more familiarity with O'Brien-Fleming boundaries when presented to stakeholders. While not strictly "optimal" in all scenarios, recommending a single boundary choice could facilitate easier implementation and scalability in A/B testing environments to the choice of design based on the proposed loss function.

When considering designs that were implemented while being intentionally over-powered, the conclusion is similar to adequately-powered scenarios. However, for under-powered designs, it is more challenging to provide a general conclusion based on the proposed loss function. It is not unexpected that our simulation results suggest stopping early for futility is optimal in approximately 60% of the presented scenario's weight combinations, given that an under-powered design is naturally a "futile" study that is unlikely to detect the desired difference. In practice, it may be ideal to choose a fixed sample design to ensure other data, such as safety signals, may be collected in the presence of an underpowered primary outcome.

While we proposed a single loss function based on sample size parameters, others may also think about developing more kinds of loss functions. For example, loss functions could be proposed to include the type I error rate or power. If these terms are added to our existing loss function, there would be more than three weights and the results would not be easily plotted in a static 2-D figure. Further, as long as each stopping boundary accounts for the corrections to multiple interim looks, the power and type I error rates should already be similar across each approach with minimal difference.

This research has limitations worth discussing with room for further research. The simulations included only binary outcomes. While commonly used in A/B testing, other types of outcomes would be worth considering. However, given the large sample sizes simulated, it is likely that continuous outcomes would have similar results since a t-test was used in our simulation studies. A second limitation is that we considered only one outcome in an A/B test, but many experiments have multiple metrics that may be of interest. This would represent an additional layer of multiple testing that is not examined in our simulations. A third limitation is that large-scale data environments may have multiple, competing experiments running simultaneously that may not be independent. Our methods and simulations do not consider the case for potentially correlated experiments that are occurring during overlapping time periods.

It is worth noting that many novel methods have already been developed for sequential monitoring, and many of them are specifically designed for A/B testing. Johari et al. (2022, 2017, 2015) came up with always valid p-values and confidence intervals that are robust to the inflated type I error rate from continuous monitoring, which let users try to take advantage of data as fast as it becomes available. Balsubramani and Ramdas (2015) proposed a novel algorithmic framework for sequential hypothesis testing. Sample size can even be boosted at the penultimate stage in the sequential monitoring that achieves specified power against an alternative hypothesis (Gao et al., 2008). Tamburrelli and Margara (2014) investigated a novel approach to automate A/B test on a large scale. Those methods, however, may still not be easily scaled or implemented for hundreds of ongoing A/B tests, since involve what may be perceived as intensive mathematical background and complicated algorithms. Conversely, our recommended designs and stopping rules in this paper are simpler and easy to be implemented on a large scale and build off a rich history in biomedical clinical trials research.

This article provided an overview of fundamental concepts and a reference of choice of optimal study designs with interim monitoring for A/B testing. Future work will extend the proposed design and loss function to non-inferiority and equivalence studies, as well as experiments with multiple outcomes. Additional considerations will be given to the design of flexible platform trials that have emerged in biomedical research, to see if adaptations can facilitate the design of an optimal sequence of studies to arrive at an optimal product via sequential and potentially simultaneous A/B experimentation.

Supplementary Material

All tables and Figures are uploaded as Supplementary Materials.

Funding

AMK and WZ supported by NHLBI K01 HL151754.

References

Armitage P, McPherson C, Rowe B (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A. General*, 132(2): 235–244. https://doi.org/10.2307/2343787

Azevedo EM, Deng A, Montiel Olea Rao JL, Rao J Weyl EG (2020). A/b testing with fat tails. *Journal of Political Economy*, 128(12): 4614–000. https://doi.org/10.1086/710607

Balsubramani A, Ramdas A (2015). Sequential nonparametric testing with the law of the iterated logarithm. arXiv preprint: https://arxiv.org/abs/1506.03486.

D'agostino RB, Chase W, Belanger A (1988). The appropriateness of some common procedures for testing the equality of two independent binomial populations. *American Statistician*, 42(3): 198–202. https://doi.org/10.1080/00031305.1988.10475563

Demets DL, Lan KG (1994). Interim analysis: The alpha spending function approach. *Statistics in Medicine*, 13(13–14): 1341–1352. https://doi.org/10.1002/sim.4780131308

Friedman LM, Furberg CD, DeMets DL, Reboussin DM, Granger CB (2015). *Fundamentals of Clinical Trials*. Springer.

Gao P, Ware JH, Mehta C (2008). Sample size re-estimation for adaptive sequential design in clinical trials. *Journal of Biopharmaceutical Statistics*, 18(6): 1184–1196. https://doi.org/10.1080/10543400802369053

Gordon Lan K, Reboussin DM, DeMets DL (1994). Information and information fractions for design and sequential monitoring of clinical trials. *Communications in Statistics. Theory and Methods*, 23(2): 403–420. https://doi.org/10.1080/03610929408831263

Haybittle J (1971). Repeated assessment of results in clinical trials of cancer treatment. *British Journal of Radiology*, 44(526): 793–797. https://doi.org/10.1259/0007-1285-44-526-793

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

Johari R, Koomen P, Pekelis L, Walsh D (2017). Peeking at a/b tests: Why it matters, and what to do about it. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1517–1525.

Johari R, Koomen P, Pekelis L, Walsh D (2022). Always valid inference: Continuous monitoring of a/b tests. *Operations Research*, 70(3): 1806–1821. https://doi.org/10.1287/opre.2021.2135

Johari R, Pekelis L, Walsh DJ (2015). Always valid inference: Bringing sequential analysis to a/b testing. arXiv preprint: https://arxiv.org/abs/1512.04922.

Kohavi R, Deng A, Frasca B, Walker T, Xu Y, Pohlmann N (2013). Online controlled experiments at large scale. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1168–1176.

Koning R, Hasan S, Chatterji A (2022). Experimentation and start-up performance: Evidence from a/b testing. *Management Science*.

Miller E (2010). *How Not to Run an A/B Test*. URL: http://www.evanmiller.org/how-not-to-run-an-ab-test.html

Miller E (2015). *Simple Sequential A/B Testing*. URL http://www.evanmiller.org/sequential-abtesting.html, blog post.

O'Brien PC, Fleming TR (1979). A multiple testing procedure for clinical trials. *Biometrics*, 549–556. https://doi.org/10.2307/2530245

Pocock SJ (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2): 191–199. https://doi.org/10.1093/biomet/64.2.191

Tamburrelli G, Margara A (2014). Towards automated a/b testing. In: *International Symposium on Search Based Software Engineering*, 184–198. Springer.

Wang SK, Tsiatis AA (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, 193–199. https://doi.org/10.2307/2531959

Zhou W, Kroehl M, Meier M, Kaizer A (2023). Approaches to analyzing binary data for large-scale A/B testing. *Contemporary Clinical Trials Communications*, 101091–101091. https://doi.org/10.1016/j.conctc.2023.101091