Know When to Fold: Futility-Aware Early Termination in Online Experiments

Yu Liu*
Amazon
Seattle, USA
liuyu0jlu@gmail.com

Runzhe Wan*
Amazon
Seattle, USA
runzhe.wan@gmail.com

Yian Huang
Columbia Univeristy
New York, USA
huang.yian@columbia.edu

James McQueen
Amazon
Seattle, USA
jmcq@amazon.com

Doug Hains
Amazon
Seattle, USA
dhains@amazon.com

Jinxiang Gu
Amazon
Seattle, USA
gjinxian@amazon.com

Rui Song
Amazon
Seattle, USA
ruisong@amazon.com

Abstract

As the demand for online A/B testing continues to rises for tech companies, the opportunity cost of conducting these experiments becomes increasingly significant. Consequently, there is a rising need for an efficient continuous monitoring system capable of early terminating experiments when necessary. Existing literature and tools primarily focuses on early terminating experiments with evidently significant results (*demonstrated efficacy*). However, for example, among the tens of thousands of online experiments conducted every year in Amazon, only a small proportion will meet launch criterion. To improve innovation efficiency and allow terminating experiments for *futility*, in this paper, we present a comprehensive literature review, propose new methods, and conduct a large-scale meta-analysis using historical online experiments in Amazon. This is the first such kind of study in the literature. We also delve into empirical challenges and explore various empirical strategies to handle them that we met while deploying these methods at Amazon. This paper is based on our work to develop the first such service for the largest online experiment platform at Amazon. Launched in 2024, this product is now available to thousands of labs on the platform each year and sends automatic notifications to experimenters with early termination recommendations. The product saves time for around 10% of labs, cuts about 2 weeks for each terminated lab, and reduces negative impact by several dozen basis points for ineffective or negative treatments.

*Equal contribution

WWW Companion '25, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/25/04
https://doi.org/10.1145/3701716.3715241

CCS Concepts

• **General and reference** → *General literature*; **Experimentation**;
• **Applied computing** → *Business process monitoring*.

Keywords

Sequential Decision Making, A/B testing, Experimentation

ACM Reference Format:
Yu Liu, Runzhe Wan, Yian Huang, James McQueen, Doug Hains, Jinxiang Gu, and Rui Song. 2025. Know When to Fold: Futility-Aware Early Termination in Online Experiments. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3701716.3715241

1 Introduction

Online A/B testing is widely used in online service companies (e.g., Amazon [25], Microsoft [17], Netflix [35], Didi Chuxing [30], etc) to compare multiple versions of one feature, such as the existing one (control) versus new one (treatment) [18], by tracking specific business metrics over a defined period, such as purchases or clicks. Subsequently, statistical inference is drawn to determine which version performs better.

A/B experiments are typically conducted using *fixed-horizon* hypothesis testing. This predetermined fixed horizon, commonly referred to as the *recommended duration* [25], is established through power analysis [5, 27]. Power analysis ensures that a sufficient sample size is collected during this recommended duration so as to achieve a minimal power. However, power analysis is a static approach: it is determined before experiment data are observed or when only initial periods' data are available, making it unable to reflect the dynamic status of the experiment. This limitation can lead to experiments running longer than necessary, thereby slowing down the innovation cycle and consuming more hardware and human resources. To address this issue, *continuous monitoring* (also known as *early termination*) is widely adopted to make early termination decisions based on interim data.

Usually, one may consider to terminate earlier in two cases: (1) experiments with *super efficacy*, indicating early evidence for a significant treatment effect (2) experiments with *futility*, indicating it is unlikely to achieve a significant treatment effect even if we continue. The futility can be due to low true treatment effect, high noise level, or the sample size is too low. To illustrate our motivations and also the difference between these two concepts, we cluster the trajectories of probability of positive return (PPR) for certain business metric in thousands of experiments over four weeks in Amazon, as depicted in Figure 1. The dashed line represents the threshold of successful experiments. Cluster 3 consistently shows positive effects over weeks, suggesting that the treatment is superior to the control and should be terminated early for efficacy. In contrast, Cluster 0 and 2 are overall less likely to achieve positive statistical significance. Therefore, we may consider terminating them for futility. We can also see that a significant portion of real online experiments exhibits very low power in Figure 2.

Contributions. This paper is the *first* systematic study on continuous monitoring for futility in the literature, and it is based on our work to develop and deploy the first futility-aware early termination method for the largest online experiment platform at Amazon. This system has been launched on July 2024, available to the thousands of labs running on the platform per year, and will send automatic notifications to experimenters on the early termination recommendation. While we cannot share the exact impact figures for confidentiality reasons, the product has been positively received by users. It saves time for about 10% of labs, reduces approximately 2 weeks for each terminated lab, and lowers the negative impact by several dozen basis points across all tested products for treatments with futile or negative effects (see Section 6 for details). Our contributions are multi-fold:

(1) We conduct a comprehensive literature review of approaches that can be applied for futility-aware early termination;
(2) We introduce two novel methods, including a data-based prediction method that leverages machine learning and an optimization-based method for utility maximization;
(3) We discuss real complexities and our practical solutions when developing these methods in a real large-scale experimentation platform.
(4) We conduct a large-scale meta-analysis on experiments running at Amazon and compare the performance of these methods to provide insights.

Outline. We organize the paper as follows. In Section 2, we introduce the related works. Section 3 gives the problem setting. Section 4 provides detailed guidance on applying literature review approaches to real-world applications, alongside proposing our own methods. Additionally, discusses guidance for fine-tuning hyperparameters, which is supported by our meta-analysis results in Section 6. Section 7 concludes this paper.

2 Related Work

Duration Recommendation. In the extensive research dedicated to determining sample sizes in clinical trials [1, 15, 16, 24, 33], power analysis stands out as one standard approach [10]. For online experiments, power analysis is also used to determine the minimal sample size required to achieve the specified level of statistical

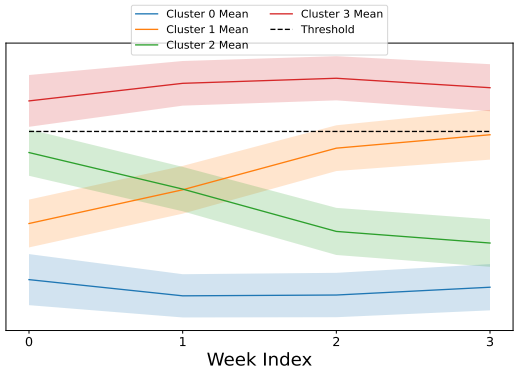

Figure 1: Trajectory clusters of PPR among thousands of real experiments. The dashed line represents the threshold of successful experiments.

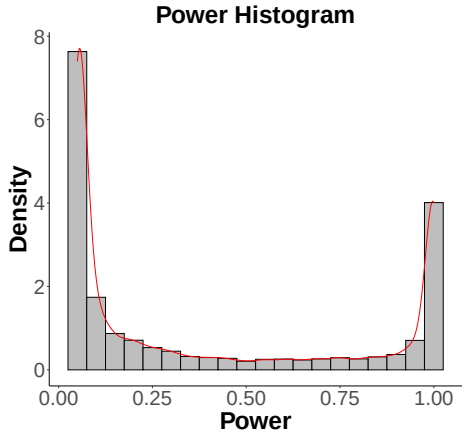

Figure 2: Histogram of the power values at the end of the experiments. Note this is not *post-hoc* power that uses observed effect size,

power. [25] proposes two methods for determining the suitable effect size in power analysis for online experiments. [27] proposed a hierarchical model using first week's traffic to predict the sample size.

Duration recommendation is static and doesn't incorporate observed experiment results. Thus continuous monitoring is used to make early decisions. **Continuous monitoring for efficacy** include Alpha-spending function approach [6, 11], which is proposed to control the overall type-I error across interim analysis. Sequential probability ratio test (SPRT) [31] and its variations, including MaxSPRT [19] and mixture SPRT [28] are proposed. Always valid inference [13, 14] controls the type-I error rate by converting any existing sequential testing into a sequence of p-values. In Bayesian hypothesis testing, the sequential Bayes factor (SBF) and its variation have been extensively discussed [8, 12, 29]. [32] innovates a reward-based approach using Reinforcement learning for continuous monitoring online experiment and conducts a comprehensive performance comparison with those methods for early termination for efficacy.

Approaches used in **continuous monitoring for futility** may overlap with early termination for efficacy. For instance, both SPRT and SBF can be applied to address both futility and efficacy. For example, SBF can be used to terminate a study for efficacy if the Bayes Factor (BF) value is very large. Conversely, if the BF value is too small, it can serve for terminating for futility. On the other hand, methods based on power are exclusively tailored for early termination for futility, e.g. condition power (CP) [3, 22], the predictive power (PP) [21] and probability of success (PoS) [9, 21].

3 Preliminary

We first introduce some notations and an assumption.

Setup. To simplify the exposition, we focus on a two-arm experiment and consider the one-sided hypothesis test. We will discuss the extensions later. Let J be the recommended duration, i.e., the maximum number of weeks for an experiment. Denote $n_{j,g}, g \in T, C$, as the number of units in group g at time j (i.e., the week j), respectively, where T denotes the treatment group and C denotes the control group. Define the set of customers who are initially *triggered* using an experiment-specific triggering mechanism (e.g., upon encountering the relevant features) in the experiment during week j as $I_{j,g}$. Let $X_{i,j}$ be the response value of unit i in week j. $\boldsymbol{X_{jg}} = \{X_{i,t} | i \in \bigcup_{k=1}^{j} I_{k,g}, t \in [\![1, j]\!]\}$ be the observed response values of units in group g up to week j, where the integer set $[\![a, b]\!]$ is defined as $\{a, a+1 \ldots, b\}$.

To extend the methods in the literature to our setting in online experiments, we introduce the concept of *effective sample size*. In online experiments, we typically follow the ongoing activities of customers from their initial engagement in the experiment until its end. The effective sample size up to week j is defined as $N(j, g) := \sum_{t=1}^{j} n_{t,g}$, which defines the observed unit as individual customer-weeks rather than individual customers. The total effect size at week j is defined as $N(j) := N(j, T) + N(j, C)$. This accounts for the uncertainty in predicting future responses. When we stop at one week, we can use all available data to run hypothesis testing, in either the frequentest or Bayesian setting.

Assumption 1. *We assume that each customer-week is independently and identically distributed (i.i.d.) after first triggering:*

$$For\ i \in I_{l,g}, \begin{cases} X_{ij} = 0, & if\ j \in [\![1, l-1]\!] \\ X_{ij} \sim F, & if\ j \in [\![l, J]\!] \end{cases}$$

where F represents some distribution with mean μ_g and variance σ_g^2, $\sigma_g^2 < \infty$.

Thus, at week j, according to the central limited theorem (CLT), the sample mean at the interim analysis for the arm g :

$$\bar{X}_{jg} = \frac{\sum_{i \in \bigcup_{k=1}^{j} I_{k,g}, t \in [\![1,j]\!]} X_{i,t}}{N(j, g)} \xrightarrow{d} \mathcal{N}(\mu_g, \frac{\sigma_g^2}{N(j, g)}), g \in \{T, C\}$$

We acknowledge that, in some experiments, violations of this assumption may occur, such as tracking the customer's activity over multiple weeks or encountering non-stationary trends over time. However, we emphasize that these assumption are only needed for appreciating the derivations of some methods; our final evaluation and performance guarantees are empirical without relying on them.

Hypothesis testing and Power Without loss of generality, we consider one-tailed tests and equal allocation between the treatment and control group. Define $\mu := \mu_T - \mu_C$, we want to test

$$H_0 : \mu \leq 0 \qquad vs \qquad H_1 : \mu > 0$$

In a *frequentist setting*, the Welch t-test [34] is commonly used to test the hypothesis mentioned above. The standard normal distribution approximates the t-distribution as sample sizes grow larger, which holds in online experiment setting. In this paper, we use z-scores for the sake of simplicity in explanation. Z-score at week j is computed as $\frac{\bar{X}_{jT} - \bar{X}_{jC}}{\sqrt{\sigma_T^2/N(j,T) + \sigma_T^2/N(j,C)}}$. In fixed-horizon testing, the null hypothesis is rejected if the p-value for the observed z-score at week J is smaller than the level of statistical significance α.

In a *Bayesian setting*, we define the probability of positive return (PPR) as $P\{\mu > 0 \mid \boldsymbol{X}_{JT}, \boldsymbol{X}_{JC}\}$, and reject the null hypothesis if this PPR exceeds a certain threshold η.

Power refers to the probability that a statistical test will correctly reject the null hypothesis. In fixed-horizon testing, the smallest sample size (or optimal duration) is determined as the one that achieves the minimal threshold for power, typically set at 80% or another specified level.

Performance Metrics. Before introducing our methods, it is important to define what is a good futility-aware early termination method. This is by nature a trade-off: the shorter we run the experiment, the more time we can save, but the lower accuracy we will have. Therefore, we introduce a few performance metrics to evaluate the performance in a holistic way. We note that since the ground truth of treatment effect (and hence correct decision) is not observable in real data, to not overly rely on assumptions, we mainly focus on empirical performance metrics that compare the decisions following different early termination rules with those made if we continue until the end of the duration J. The latter is regarded as a ground truth proxy. Such a practice yields intuitive explanations that are easy to convey to users. We define the following metrics, with detailed definitions of some metrics also available in Table 1:

(1) *% Early terminated experiments* is the percentage of experiments that terminated early.
(2) *Empirical False Termination Rate (eFTR)* is the probability of falsely terminating an experiment that would be launched if we wait until week J, among all launched experiments.
(3) *Empirical True Termination Rate (eTTR)* is the probability of correctly terminating an experiment that would not be launched even we wait until week J, among all experiments that were not launched.
(4) *Empirical False Discover Rate (eFDR)* is the frequency of falsely terminated cases among the experiments terminated early for futility, as determined by different methods.
(5) *Average Saved Weeks* is the average number of weeks we saved, over all experiments.
(6) *Opportunity Cost Saving*: We estimate the opportunity cost (OC; in certain unit D) being saved. Roughly speaking, we quantify the contributions from new features launched via experiments, and attribute the blocked opportunities due to

the limited experimentation resource we observed in historical data. The specific details are less relevant and we choose not to release for confidentiality.

	Terminated for futility (Te)	Not Terminated (Nte)	
Launch at the end (L)	falsely terminated (FT)		$eFTR = \frac{FT}{L}$
No Launch at the end (NL)	correctly terminated (CT)	missed opportunity	$eTTR = \frac{CT}{NL}$
	$eFDR = \frac{FT}{Te}$		

Table 1: eFRT, eTTR and eFDR definition Matrix

4 Methodology

From Section 4.1 to 4.4, we introduce a few different methods for early termination for futility. We propose four types of methods: the first type uses theoretical models to predict trial outcomes at the end (Sections 4.1), the second type trains a prediction model using historical data to predict trial outcomes at the end (Sections 4.1), the third approach focuses solely on making decisions based on the current observed results (Section 4.3), while the fourth solves an optimization problem to find the best early stopping rule (Section 4.4).

4.1 Theory-based Prediction Methods

The first approach uses probabilistic model to simulate and predict the end-of-horizon decisions. As these methods rely on theory-based model assumptions and the model itself does not involve training with historical data, we call them *Theory-based Prediction Methods.*

4.1.1 Conditional Power. Conditional power (CP) computes the Frequentist power of the experiment conditioning on the interim result and assuming a fixed effect size θ for the remainder of the trial. Denote T_J as the test statistic and $\mathcal{R}_J$ be the rejection region at the end of the experiment, which often using the Welch's t-test [34] and z-test. According to formula (4.7) of [20], CP at week j can be computed as:

$$\begin{aligned}\text{CP} &= P\{T_J \in \mathcal{R}_J \mid \mu = \theta, \boldsymbol{X}_{j,T}, \boldsymbol{X}_{j,C}\} \\ &= \Phi\Bigg(\frac{1}{r \cdot s_n}\sqrt{\frac{N(J)}{N(J)-N(j)}}\Bigg(\frac{1}{\sqrt{N(J)}} \\ &\quad \cdot \Big(N(j)\delta_j + \big(N(J)-N(j)\big)\theta\Big) - r \cdot s_n z_\alpha\Bigg)\Bigg) \end{aligned} \tag{1}$$

where s_n is the estimate of the pooled standard deviation, z_α is the upper α quantile of standard normal, and $\delta_j = \bar{X}_{jT} - \bar{X}_{jC}$ which is the estimated effect size during interim analysis at week j. For two-arm trial, the allocation ratio is $a:1$ and $r = \sqrt{\frac{(a+1)^2}{a}}$ which is 2 for equal allocation. At week j, the total sample size $N(J)$ is unknown. therefore, we employ sample size prediction methods to estimate its value [27].

There are two quantities that need to be prespecified: First, the assumed effect size θ: it is common practice for the experimenter to manually select the value of θ. Alternative approaches are discussed in [25]. Second, the early termination threshold: if the conditional power is smaller than this threshold, the experiment may be terminated early due to futility.

4.1.2 Predictive Power. In the conditional power calculation, we need to specify a θ for the post-interim data. Predictive power (PP) is an alternative method which adopts a prior for $\mu \sim N(\mu_0, \sigma_0^2)$ and average the CP over the posterior distribution of μ. According to formula (4.10) of [20], PP at week j can be computed as:

$$\begin{aligned}\text{PP} &= E_{\mu\sim P(\mu|\boldsymbol{X}_{jT},\boldsymbol{X}_{jC})} P\{T_J \in \mathcal{R}_J | \mu, \boldsymbol{X}_{j,T}, \boldsymbol{X}_{j,C}\} \\ &= \Phi\Bigg(\frac{1}{rs_n}\sqrt{\frac{N(j)}{N(J)-N(j)}} \cdot \frac{1}{\sqrt{\psi + (1-\psi)N(j)/N(J)}} \\ &\quad \Bigg[\frac{(1-\psi)[N(j)\delta_j + (N(J)-N(j))\mu_0]}{\sqrt{N(J)}} + \frac{\psi\delta_j}{1/\sqrt{N(J)}} - rs_n z_\alpha\Bigg]\Bigg)\end{aligned} \tag{2}$$

where $\psi = \frac{N(j)\sigma_0^2}{N(j)\sigma_0^2 + r^2 s_n^2}$. Note that a special case is when $\sigma_0 \to \infty$, and this corresponds to the noninformative prior. PP is a mixed Bayesian-Frequentist approach since we use the Frequentist rejection region and assume a prior for the effect size. We recommend early termination of the experiment when the PP falls below a pre-specified threshold.

4.1.3 Bayesian Predictive. Bayesian predictive (BP) is a purely Bayesian approach, where the Bayesian criterion is used at the end of the trial. According to formula (1) of [9], at week j, taking the expectation with respect to the posterior predictive distribution of future observations $\boldsymbol{X}_{(J-j)g}$ given $\boldsymbol{X}_{jg}$, BP at week j can be computed as:

$$\begin{aligned}\text{BP} &= E\mathbb{1}\{P\{\mu > 0 \mid \boldsymbol{X}_{jT}, \boldsymbol{X}_{jC}, \widetilde{\boldsymbol{X}}_{(J-j)T}, \widetilde{\boldsymbol{X}}_{(J-j)C}\} > \eta\} \\ &= \Phi\left(\frac{\sqrt{\frac{N(J)N(j)}{8}}\frac{\delta_j}{s_n}(1-a(j)) + \frac{b(j)}{s_n} + \sqrt{\frac{N(j)}{2}}z_\eta(1-c(J))}{\sqrt{\frac{N(J)-N(j)}{N(J)}(1-a(J))^2\left[(\frac{N(J)-N(j)}{2})(1-a(j)) + \frac{N(j)}{2}\right]}}\right)\end{aligned} \tag{3}$$

where $\widetilde{\boldsymbol{X}}_{(J-j)g}$ are the predicted values of future observations from week j up to week J, $\mathbb{1}\{.\}$ is the indicator function. $a(j) = (1 + \frac{N(j)}{2}(\frac{\sigma_0}{\sqrt{2}s_n})^2)^{-1}$, $b(j) = \sqrt{\frac{N(J)N(j)}{8}}\mu_0(1 + \frac{N(j)}{2}(\frac{\sigma_0}{\sqrt{2}s_n})^2)^{-1}$ and $c(j) = 1 - (1 + \frac{2}{N(j)}(\frac{\sqrt{2}s_n}{\sigma_0})^2)^{-1}$. We recommend early termination of the experiment when the BP falls below a pre-specified threshold.

4.2 Data-based Prediction Methods

The training-free approach in Section 4.1 relies on model assumptions. Alternatively, we can consider a pure empirical approach to train prediction models using historical data. The early termination problem can be unified under a regression (or classification) problem: at interim time j, experimenters aim to predict whether the online experiment will meet launch criterion at the end week J.

Features. A critical step in the proposed algorithm is the choice of the features. Given the trajectory of an experiment up to week j, features can be a list of summary statistics S_j, including estimated effect size δ_j, mean of control group $\bar{X}_{jC}$, sample size of treatment

and control group $N(j,.)$, pooled sample standard deviation $s_{n,j}$, statistical measures used to define the criteria for experiment launch z_j (e.g. the Z-score in frequentist test and the PPR in Bayesian test).

Algorithm. We can leverage modern machine learning (ML) algorithms to achieve this goal as depicted in Algorithm 1. This approach is under-explored in the literature.

Algorithm 1 Data-based Prediction

Input: Trajectories of m past experiments and its launch decision $\boldsymbol{L} = L_1, \ldots, L_m$.
Output: early termination decision for a new experiment.

for $j = 1$ **to** $J-1$ **do**
  For each experiment, extract features $\boldsymbol{S_j} = (S_1, S_2, \ldots, S_j)$
  Use a machine learning (ML) algorithm along with 5-fold cross-fitting to train the model.
  Predict the z_J for m past experiments using previous features up to week j, denoting the predicted values as $\hat{z}_{J,i,j}, i = 1, \ldots, m$.
end for
Set the early termination threshold K_{db} as the lowest predicted score $\hat{z}_{J,i,j}$ among launched experiment and over j.
$K_{db} := \min_{i:L_i=\text{Launch}} \min_{j=1,\ldots,J} \hat{z}_{J,i,j}$
Given a new experiment,
for $j = 1$ **to** $J-1$ **do**
  predict its z_J using trained model.
  if predicted $z_J < K_{db}$ **then**
    return Terminate due to futility at week j
  else
    return Continue
  end if
end for

Specifically, we use a gradient boosting model to predict the Z-score in Frequentist test (or PPR in Bayesian test) at week J, with above features. At week j, all features up to and including week j are utilized. For example, with $J = 4$, at week 3, the features from weeks 1, 2, and 3 are used to predict the z-score or PPR at week 4. If the predicted score is less than the threshold K_{db}, we suggest early terminating the experiment for futility.

4.3 Non-prediction methods

Alternatively, there are methods that do not (explicitly) rely on predicting future outcomes, but only the current observations.

4.3.1 Bayes Factor and Posterior Odds. We next introduce Bayes factor (BF) and posterior odds (PO)([7, 8]). According to Assumption 1,

$$\bar{X}_j = \bar{X}_{jT} - \bar{X}_{jC} \sim \mathcal{N}(\mu, \sigma_j^2), \tag{4}$$

where $\sigma_j^2 = \frac{\sigma_T^2}{N(j,T)} + \frac{\sigma_C^2}{N(j,C)}$. Further assume there is a prior: $\mu := \mu_T - \mu_C \sim \mathcal{N}(\mu_0, \sigma_0^2)$.

$$H_0 : \mu \overset{\text{d}}{=} \mathbb{1}\{U \le 0\}U \quad \text{vs.} \quad H_1 : \mu \overset{\text{d}}{=} \mathbb{1}\{U > 0\}U \tag{5}$$

where $U \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Following the calculation in Appedix A, we have

$$\text{Bayes Factor} := \frac{P(\bar{X}_j|H_1)}{P(\bar{X}_j|H_0)} = \frac{\Phi(-\mu_0/\sigma_0)}{1-\Phi(-\mu_0/\sigma_0)} \cdot \frac{1-\Phi(-\tilde{\mu}/\tilde{\sigma})}{\Phi(-\tilde{\mu}/\tilde{\sigma})} \tag{6}$$

where $\tilde{\mu} = \frac{\mu_0/\sigma_0^2 + \bar{x}/\sigma_j^2}{1/\sigma_0^2 + 1/\sigma_j^2}$ and $\tilde{\sigma^2} = \frac{1}{1/\sigma_0^2 + 1/\sigma_j^2}$. Following the Bayes formula, we have

$$\text{Posterior Odds} := \frac{P(H_1|\bar{X}_j)}{P(H_0|\bar{X}_j)} = \frac{P(H_1)}{P(H_0)} \frac{P(\bar{X}_j|H_1)}{P(\bar{X}_j|H_0)} \tag{7}$$

Bayes factor and posterior odds reflect our current belief about H_0 and H_1 after collecting the data. We can monitor the online experiment using Bayes factor or posterior odds. For example, if the Bayes factor $\frac{P(\bar{X}_j|H_1)}{P(\bar{X}_j|H_0)}$ or posterior odds $\frac{P(H_1|\bar{X}_j)}{P(H_0|\bar{X}_j)}$ is smaller than a pre-defined threshold K_{PO}, we may early terminate the experiment and accept H_0.

While [8] have discussed theoretical considerations for selecting K_{PO} to control false discovery rate, we provide additional empirical guidance on selecting K for real-world applications in Section 5, particularly when underlying assumptions are not met.

4.4 Optimization-based Approach

Finally, we propose another novel approach for early termination for futility, rooted in simulation optimization and optimal decision rule.

Specifically, we note two things: 1) we adopt a stance that we will pick the optimal method based on its empirical performance in meta-analysis instead of relying on theoretical assumptions, and 2) essentially we (and all aforementioned methods) want to find a series of time-dependent decision rule $\pi_j(x_j;\theta) \in \{\text{Terminate due to futility, Continue}\}$, where j is the week index, x_j is all information available so far (e.g., data points and the prior), and θ parameterizes the policy). Therefore, we can directly find the globally optimal decision rule by solving an optimization problem. Specifically, we can consider $\{\pi_j(\cdot;\theta)|\ \theta \in \Theta\}$ as a given Neural Network (NN) architecture (or tree-based model) with binary outputs, and the solve the one that maximizes the pre-specified objective function (e.g., the eTTR) under certain constraints (e.g., eFTR < 1% and eFDR < 5%). The optimization can be done via stochastic gradient descent, which is essentially policy gradient in sequential decision making, akin to [32]. We compare all previous methods except for the optimization-based approach, the engineering complexity of which sets it apart from other methods discussed in this section and hence is left for future study.

5 Tuning and Other Real Considerations

So far, we have introduced a few methods for early termination for futility. To deploy them in a large-scale real experimentation platform, there are still a few critical decisions to make.

First, our discussion above is about one business metric, but an experiment typically tracks multiple business metrics to construct the launch criteria (LC), i.e., under what conditions we should launch a new feature. While a straightforward approach is to terminate when it is futile to meet this LC (e.g., predicting this event as a whole), it is complicated because business metric combinations can vary experiment by experiments, and it is non-trivial to develop, test and maintain a uniformly good and adaptive procedure. In our meta-analysis with Amazon's experiment data, we study making decision for each business metric separately and aggregate them using an AND/OR logic: e.g., if the LC is to launch when all business

metrics meet the requirement ("AND" logic), then we terminate for futility when any of them cannot meet its requirement (i.e., with a "OR" logic); and vice versa, we will terminate with an "AND" logic when the LC is based on an "OR" logic. We analyze its impact in the numerical study.

Second, each method discussed above requires one *hyperparameter*, i.e., the pre-specified early termination threshold (e.g., the posterior ratio threshold K_{PO} for PO). How to choose the hyperparameter for each algorithm is critical. This essentially relates to how do we evaluate the performance. In our meta-analysis, we start from a conservative approach that is working-backwards and is free of model assumption: since both the error rate and the size of savings discussed in Section 3 are generally monotone with that parameter, we can find the hyper-parameter that makes our (empirical) error rates right below some pre-specified values (e.g., 1% for eFTR and 5% for eFDR). The next question is, under which setup do we tune the hyper-parameters? The hyper-parameter for a given business metric may change with the business metrics of interest among different experiments, We note that the only exception is that we do not tune the beta-spending method [23], as its hyper-parameter has the meaning as the assumed effect size which is used in power analysis. Instead, we try two different assumed effect sizes in our meta-analysis. Besides, we emphasize that, although our tuning and the following experiments are mainly empirical, for some methods we still have the theoretical guarantee if one is willing to make the technical assumptions. For example, for PO, suppose our tuning suggests $K = 9$, then we still control the false discovery rate ≤ 0.1. Moreover, instead of using a fixed threshold over time, we can allow the threshold to change over time. The optimal time-varying threshold can be solved from the the same optimization problem, e.g., maximizing opportunity cost saving, with eFTR < 1% and eFDR < 10%. Finally, we note, we only use the dataset to tune one parameter as the threshold, hence the impact of over-fitting is little. Indeed, in our initial tests, we observe the out-of-sample performance is almost the same with in-sample performance. Therefore we do not distinguish them below.

Third, in this paper we consider only two arms. With more arms, the methods can be easily extended with an "AND" logic, i.e., we terminate when all treatment arms show futility. However, this would be inefficient with many arms, since the probability of termination will decay exponentially. In that case, we recommend using best-arm identification algorithms [2].

6 Meta Analysis

In this section, we implement the previous discussed methods and compare their empirical performance in various settings with a large number of real Amazon experiments, so that to comprehensively evaluate their performance and choose the best configuration to launch in Amazon production.

Dataset. We use thousands of real experiments conducted in Amazon that last at least $J = 4$ weeks and use two common business metrics, labeled as "metric A" and "metric B", to demonstrate our approach to managing multiple metrics, as discussed in Section 4. This dataset enables us to study the performance of different techniques on different business metrics.

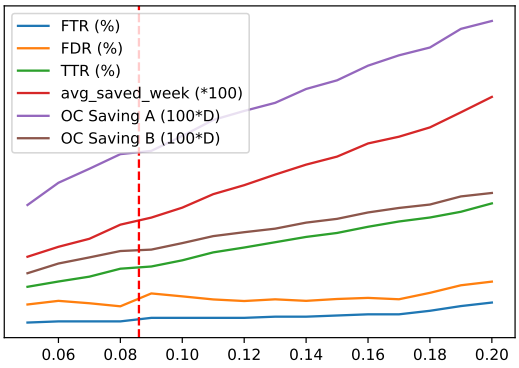

Figure 3: PO's performance trend with the threshold tuning parameter, with the setting in Table 9. The X-axis represents the threshold values, while the Y-axis displays the performance metric values. Each performance metric is plotted on a separate line. Due to confidentiality reasons, the values on the y-axis are not displayed. The dashed line represents the selected threshold, which maximizes OC savings and constrains eFTR errors below 1%.

Methods to Compare. We compare all methods discussed in Section 4 along with beta-spending approach [23], except for the optimization-based approach, the engineering complexity of which makes it not a viable now and hence is left for future study. We use the both Bayes factors (`BF`) defined in Equation 6 and Posterior Odds (`PO`) defined in Equation 7 [8], the latter of which also takes prior odds into consideration. We use the Predictive Power (`PP`) defined in Equation 2 and Bayesian Power (`BP`) defined in Equation 3. We use the Conditional Power (`CP`) defined in Equation 1. We use the Machine Learning (`ML`) method defined in Algorithm 1 using XGBoost [4]. We use the beta-spending function (`beta`) [23]. Given that `beta` needs a predefined fixed effect size, we explore two distinct assumed effect sizes in our meta-analysis. One, denoted as 'beta (small)', assumes a smaller effect size, yielding conservative results. The other, labeled as 'beta (large)', employs a larger effect size. The total sample size $N(J)$ across all methods is predicted using a Beta-Geometric model [27].

Our dataset encompasses experiments conducted in both Bayesian settings, where the PPR is used as statistical measure, and Frequentist settings, where Z-scores serve as statistical measure. We note that since experiments with the Bayesian setting dominate the dataset, we do not separately analyze their performance. Besides, since `BP` and `PP` share essentially the same idea, we merge their performance together, i.e., we use `BP` for experiments with Bayesian setting and `PP` for those with frequentest setting. The idea of `CP` is similar and its performance is significantly worse in this dataset than BP/PP, so we didn't include it in the results.

Parameter Tuning. As discussed in Section 5, methods such as `BF`, `BP`, `PP` and `ML` requires an early termination threshold. The empirical method for tuning thresholds follows these steps: We establish a range of thresholds and calculate the corresponding

Table 2: Results with Business Metric B using diff-of-means estimator.

	eFTR	eTTR	eFDR	% Terminated	Average Saved Weeks
PO	1.26%	7.83%	10.16%	5.13%	0.11
BF	1.26%	7.83%	10.16%	5.13%	0.11
BP/PP	1.26%	24.25%	3.52%	14.78%	0.24
ML	1.26%	8.45%	9.49%	5.49%	0.10
CP(observed)	1.26%	6.95%	11.30%	4.61%	0.10
CP(ASE)	1.26%	31.34%	2.75%	18.94	0.21
beta (small)	52.19%	95.44%	27.71%	77.61%	1.06
beta (large)	80.08%	97.41%	36.56%	90.27%	1.53

Table 3: Results with Business Metric B applying covariate adjustment.

	eFTR	eTTR	eFDR	% Terminated	Average Saved Weeks
PO	1.05%	10.53%	7.52%	6.26%	0.15
BF	1.05%	10.53%	7.52%	6.26%	0.15
BP/PP	1.05%	3.51%	19.61%	2.4%	0.03
ML	1.05%	14.9%	5.43%	8.67%	0.16
CP(observed)	1.05%	3.00%	22.22%	2.12%	0.04
CP(ASE)	1.05%	2.74%	23.81%	1.98%	0.02

Table 4: Results with business Metric A and Metric B, "AND" logic-based LC for terminating .

	eFTR	eTTR	eFDR	% Terminated	Average Saved Weeks	Metric A OC Saving (D)	Metric B OC Saving (D)
PO	1.2%	10.9%	4.1%	8.2%	0.19	0.344	0.15
BF	1.2%	10.9%	4.1%	8.2%	0.19	0.344	0.15
BP/PP	1.2%	22.4%	2.0%	16.6%	0.24	0.124	0.054
ML	1.8%	15.2%	4.2%	11.5%	0.21	0.341	0.148
CP(observed)	1.6%	7.7%	7.2%	6.01%	0.12	0.044	0.031
CP(ASE)	1.1%	18.79%	2.1%	13.95%	0.15	0.016	0.007

Table 5: Results under either the optimal fixed threshold or optimal time-varying threshold using Metric A for PO method.

	thresholds	eFTR	eTTR	eFDR	% Terminated	Average Saved Weeks	OC Saving (D)
Time-varying	[0.235, 0.235, 0.219]	1.0%	7.4%	7.1%	5.0%	0.10	0.25
Fixed	0.22	0.9%	7.0%	6.8%	4.7%	0.09	0.23

values for various performance metrics, as outlined in Section 3. Then, we identify the optimal thresholds that maximize savings while guaranteeing that the error rate remains below a specified threshold. For example, in Figure 3, we present the trend of different performance metrics with different values of threshold tuning parameter when using PO. we identify the optimal threshold for PO that maximize savings while ensuring the eFTR is lower than 1%, as indicated by the vertical dashed line. One possible threshold chosen method for the ML method was detailed in Algorithm 1, but empirical tuning method remain viable for determining it. Therefore, in this meta analysis, we use a consistent tuning strategy across all these methods.

Results. In Table 2 - Table 3 , We present the results of this dataset using a single metric B as the launch criterion, with different treatment effect estimators. Table 4 presents the results assuming all experiment adopting AND logic-based LC, hence we early terminate one experiment if either the business metric A or B shows indication of futility. The threshold tuning for each method follows the same procedure as depicted in Figure 3. In Table 5 and Table 6, we study the performance with time-varying thresholds instead of a fixed threshold. We present our main findings here, with more results deferred to the appendix.

(1) The performance of BF and PO is almost identical, because the only difference is the prior ratio of H1 over H0, which in our dataset is almost 1 with the priors being symmetric about 0.
(2) Beta-spending performs worst in our dataset, as it is a frequentest procedure and the real LCs in our applications are mostly Bayesian.
(3) BP/PP performs well when the results are estimated using the standard diff-of-means estimator, yet their performance is fairly poor for we apply them to some experiments using the

Table 6: Results under either the optimal fixed threshold or optimal time-varying threshold using Metric B for PO method.

	thresholds	eFTR	eTTR	eFDR	% Terminated	Average Saved Weeks	OC Saving (D)
Time-varying	[0.077, 0.131, 0.315]	1.1%	23.0%	3.6%	13.1%	0.22	0.13
Fixed	0.09	1.1%	10.5%	7.5%	6.3%	0.15	0.09

covariate adjustment-based treatment effect estimator [26]. This is because these methods were derived to predict the diff-of-means estimator. While one may consider extending BP/PP's derivation, that would lose their current nice closed forms and also make the engineering pipeline complex in real applications.

(4) As expected, using an OR logic, although is conservative and has a low error rate, reduces the opportunity size a lot (comparing Table 4 and Table 9 in Appendix).
(5) With a looser threshold, we will terminate more experiments, saving more time yet yielding higher error rate.
(6) For metric B, allowing time-varying thresholds looks promising - we can improve the saving by 50% while reducing FDR and maintaining FTR. This is achieved in a natural way by having a tighter threshold at first and making it looser over time. For metric A, the room for improvement (in this setting) is smaller.

Conclusion, Launch Decision and Business Impact. The best methods are characterized by higher eTTR, average saved weeks, OC Saving and low eFTR and eFDR. Overall, For our studied experiments at Amazon, based on studies above, we find that PO and BF consistently demonstrate superior performance compared to other methods. They do not suffer the performance drop with covariate adjustment as BP/PP does. Although the number of experiments (and average saved weeks) being terminated is less in some cases, the average opportunity cost is higher (due to those experiments are of larger size). Besides, they are very easy to implement. ML exhibits highly competitive performance, particularly when covariate adjustment is applied (see Table 3). However, the need for re-training for different metrics, experiment length and other factors, make it a less scalable approach. Therefore, we finally decided to launch the PO method in production, with the hyperparameter tuned as in Table 4. We intentionally keep the threshold high when we just launch the product, to be robust and err on the safe side to collect user feedback, and are considering to increase the threshold (and also deploy the time-varing thresholds) in the following iterations. Although we cannot disclose the absolute impact figures due to confidentiality, the product has been well received by users so far. It saves time for approximately 10% of labs, reducing around 2 weeks for each terminated lab, and minimizes negative impact by several dozen basis points (over the total impact of tested products) for treatments with futile or negative effects.

7 Conclusions and Discussions

This paper presents a comprehensive literature review and a large-scale meta-analysis on futility-aware early termination we did at Amazon, which lays down the foundation for us to launch such a service at Amazon in 2024. We also introduce two novel methods, including a data-based prediction method and an optimization-based method. This is the first such kind of study in the literature.

There are a few interesting next steps. First of all, due to data limitation, our meta-analysis is mainly in terms of the empirical performance compared with the historical decisions. With more granular data for each experiment available, we can apply sample splitting to compare our decision with the (imputed) true treatment effect. Second, we can extend the methods to efficacy-based early termination service. It would be of interest to study how do the two service interact and how we can design a consistent one. Third, we study terminating the whole experiment based on average treatment effects. It would be of interest to extend the methods to study heterogeneous treatment effect and hence terminate only for a subpopulation. Lastly, extending the theory-based methods to non-standard causal effect estimators (such as the covariate-adjusted one) and non-stationary models would be also of interest.

References

[1] CJ Adcock. 1997. Sample size determination: a review. *Journal of the Royal Statistical Society: Series D (The Statistician)* 46, 2 (1997), 261–283.
[2] Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. 2010. Best arm identification in multi-armed bandits.. In *COLT*. 41–53.
[3] Rebecca A Betensky. 1997. Early stopping to accept H0 based on conditional power: approximations and comparisons. *Biometrics* (1997), 794–806.
[4] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. ACM, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785
[5] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Academic press.
[6] David L Demets and KK Gordon Lan. 1994. Interim analysis: the alpha spending function approach. *Statistics in medicine* 13, 13-14 (1994), 1341–1352.
[7] Alex Deng. 2015. Objective bayesian two sample hypothesis testing for online controlled experiments. In *Proceedings of the 24th International Conference on World Wide Web*. 923–928.
[8] Alex Deng, Jiannan Lu, and Shouyuan Chen. 2016. Continuous monitoring of A/B tests without pain: Optional stopping in Bayesian testing. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*. IEEE, 243–252.
[9] Alexei Dmitrienko and Ming-Dauh Wang. 2006. Bayesian predictive approach to interim monitoring in clinical trials. *Statistics in medicine* 25, 13 (2006), 2178–2195.
[10] Han Du and Lijuan Wang. 2016. A Bayesian power analysis procedure considering uncertainty in effect size estimates from a meta-analysis. *Multivariate behavioral research* 51, 5 (2016), 589–605.
[11] KK Gordon Lan and David L DeMets. 1983. Discrete sequential boundaries for clinical trials. *Biometrika* 70, 3 (1983), 659–663.
[12] Harold Jeffreys. 1961. *The theory of probability*. Oxford University Press.
[13] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. 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.
[14] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. 2022. Always valid inference: Continuous monitoring of a/b tests. *Operations Research* 70, 3 (2022), 1806–1821.
[15] Prashant Kadam and Supriya Bhalerao. 2010. Sample size calculation. *International journal of Ayurveda research* 1, 1 (2010), 55.
[16] Ken Kelley and Joseph R Rausch. 2006. Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological methods* 11, 4 (2006), 363.
[17] Ronny Kohavi, Thomas Crook, Roger Longbotham, Brian Frasca, Randy Henne, Juan Lavista Ferres, and Tamir Melamed. 2009. Online experimentation at Microsoft. *Data Mining Case Studies* 11, 2009 (2009), 39.
[18] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online controlled experiments at large scale. In *Proceedings of the 19th*

ACM SIGKDD international conference on Knowledge discovery and data mining. 1168–1176.
[19] Martin Kulldorff, Robert L Davis, Margarette Kolczak, Edwin Lewis, Tracy Lieu, and Richard Platt. 2011. A maximized sequential probability ratio test for drug and vaccine safety surveillance. *Sequential analysis* 30, 1 (2011), 58–78.
[20] Madan Gopal Kundu, Sandipan Samanta, and Shoubhik Mondal. 2021. Conditional power, predictive power and probability of success in clinical trials with continuous, binary and time-to-event endpoints. (2021).
[21] Madan Gopal Kundu, Sandipan Samanta, and Shoubhik Mondal. 2023. Review of calculation of conditional power, predictive power and probability of success in clinical trials with continuous, binary and time-to-event endpoints. (2023).
[22] John M Lachin. 2005. A review of methods for futility stopping based on conditional power. *Statistics in medicine* 24, 18 (2005), 2747–2764.
[23] Daniel Lakens, Friedrich Pahlke, and Gernot Wassmer. 2021. Group sequential designs: A tutorial. (2021).
[24] Dennis V Lindley. 1997. The choice of sample size. *Journal of the Royal Statistical Society: Series D (The Statistician)* 46, 2 (1997), 129–138.
[25] Yu Liu, Runzhe Wan, James McQueen, Doug Hains, Jinxiang Gu, and Rui Song. 2024. Effect size estimation for duration recommendation in online experiments: Leveraging hierarchical models and objective utility approaches. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 14044–14051.
[26] Lorenzo Masoero, Doug Hains, and James McQueen. 2023. Leveraging covariate adjustments at scale in online A/B testing. In *The KDD'23 Workshop on Causal Discovery, Prediction and Decision*. PMLR, 25–48.
[27] Thomas S Richardson, Yu Liu, James McQueen, and Doug Hains. 2022. A Bayesian Model for Online Activity Sample Sizes. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1775–1785.
[28] Herbert Robbins. 1970. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics* 41, 5 (1970), 1397–1409.
[29] Felix D Schönbrodt, Eric-Jan Wagenmakers, Michael Zehetleitner, and Marco Perugini. 2017. Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological methods* 22, 2 (2017), 322.
[30] Chengchun Shi, Runzhe Wan, Ge Song, Shikai Luo, Hongtu Zhu, and Rui Song. 2023. A multiagent reinforcement learning framework for off-policy evaluation in two-sided markets. *The Annals of Applied Statistics* 17, 4 (2023), 2701–2722.
[31] Abraham Wald. 1992. Sequential tests of statistical hypotheses. In *Breakthroughs in statistics: Foundations and basic theory*. Springer, 256–298.
[32] Runzhe Wan, Yu Liu, James McQueen, Doug Hains, and Rui Song. 2023. Experimentation platforms meet reinforcement learning: Bayesian sequential decision-making for continuous monitoring. *KDD* (2023).
[33] Robert Weiss. 1997. Bayesian sample size calculations for hypothesis testing. *Journal of the Royal Statistical Society: Series D (The Statistician)* 46, 2 (1997), 185–191.
[34] Bernard L Welch. 1947. The generalization of 'STUDENT'S'problem when several different population varlances are involved. *Biometrika* 34, 1-2 (1947), 28–35.
[35] Huizhi Xie and Juliette Aurisset. 2016. Improving the sensitivity of online controlled experiments: Case studies at netflix. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 645–654.

A Calculation of Bayes Factors for One-Sided Tests

For one-side Bayesian hypotheses testing:

$$H_0 : \mu \overset{\mathrm{d}}{=} \mathbb{1}\{U \le 0\}U \quad \text{vs.} \quad H_1 : \mu \overset{\mathrm{d}}{=} \mathbb{1}\{U > 0\}U$$

where $U \sim \mathcal{N}(\mu_0, \sigma_0^2)$ and $\mathcal{I}[.]U$ represents the truncated normal. With $\bar{X} \sim \mathcal{N}(\mu, \sigma^2)$, $\mu' = \frac{\bar{x}\sigma_0^2+\mu_0\sigma^2}{\sigma_0^2+\sigma^2}$ and $\sigma_0' = \sqrt{\frac{\sigma^2\sigma_0^2}{\sigma^2+\sigma_0^2}}$, we have

$$\begin{aligned}
\int p(\bar{x}|\mu)p(\mu|H_0)d\mu &= \frac{1}{\Phi(\frac{0-\mu_0}{\sigma_0}) - \Phi(\frac{-\infty-\mu_0}{\sigma_0})} \frac{1}{2\pi\sigma\sigma_0} \\
&\quad \cdot \int_{-\infty}^{0} exp\left(-\frac{1}{2}\left(\frac{(\bar{x}-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{\sigma_0^2} \right)\right) d\mu \\
&= \frac{1}{\Phi(\frac{-\mu_0}{\sigma_0})} \frac{1}{2\pi\sigma\sigma_0} exp\left(-\frac{1}{2}\frac{(\mu_0-\bar{x})^2}{\sigma_0^2+\sigma^2} \right) \\
&\quad \cdot \int_{-\infty}^{0} exp\left(-\frac{1}{2}\frac{\sigma_0^2+\sigma^2}{\sigma_0^2\sigma^2}\left(\mu - \frac{\bar{x}\sigma_0^2+\mu_0\sigma^2}{\sigma_0^2+\sigma^2} \right)^2 \right) d\mu \\
&= \frac{1}{\sqrt{2\pi(\sigma_0^2+\sigma^2)}} exp\left(-\frac{1}{2}\frac{(\mu_0-\bar{x})^2}{\sigma^2+\sigma_0^2} \right) \frac{\Phi(-\mu_0'/\sigma_0')}{\Phi(-\mu_0/\sigma_0)}
\end{aligned}$$

Similarly, we can compute $\int p(\bar{x}|\mu)p(\mu|H_0)d\mu$. Thus Bayes Factor is computed as:

$$\begin{aligned}
BF[H_1 : H_0] : &= \frac{P(\bar{x}|H_1)}{P(\bar{x}|H_0)} \\
&= \frac{\int_{\mu>0} p(\bar{x}|\mu)p(\mu|H_1)d\mu}{\int_{\mu<0} p(\bar{x}|\mu)p(\mu|H_0)d\mu} \\
&= \frac{1-\Phi(-\mu_0'/\sigma_0')}{1-\Phi(-\mu_0/\sigma_0)} \cdot \frac{\Phi(-\mu_0/\sigma_0)}{\Phi(-\mu_0'/\sigma_0')}
\end{aligned}$$

B More Results

We present results for under other different settings.

Table 7: Results with Business Metric A using diff-of-means estimator.

	eFTR	eTTR	eFDR	% Terminated	Average Saved Weeks
PO	1.31%	8.59%	8.45%	5.85%	0.12
BF	1.31%	8.59%	8.45%	5.85%	0.12
BP/PP	1.31%	50.96%	1.53%	32.28%	0.55
ML	1.31%	12.49%	5.97%	8.29%	0.14
CP(observed)	0.88%	14.14%	3.60%	9.15	0.17
CP	0.88%	34.96%	1.49%	22.14%	0.27

Table 8: Results with Business Metric A applying covariate adjustment.

	eFTR	eTTR	eFDR	% Terminated	Average Saved Weeks
PO	0.87%	6.95%	6.78%	4.71%	0.09
BF	0.87%	6.95%	6.78%	4.71%	0.09
BP/PP	0.87%	25.23%	1.96%	16.33%	0.24
ML	0.87%	10.56%	4.57%	6.99%	0.11
CP(observed)	0.87%	6.13%	7.62%	4.19%	0.08
CP	0.87%	19.91%	2.48%	12.89%	0.13

Table 9: Results with Business Metric A and Metric B, OR logic-based LC for terminating, and 10% experiments indeed using OR logic-based LC.

	eFTR	eTTR	eFDR	% Terminated	Average Saved Weeks	Metric A OC Saving (D)	Metric B OC Saving (D)
PO	0.0%	3.0%	0.0%	2.1%	0.04	0.087	0.038
BF	0.0%	3.0%	0.0%	2.1%	0.04	0.087	0.038
BP/PP	0.2%	1.6%	4.2%	1.2%	0.01	0.003	0.001
ML	0.0%	4.4%	0.0%	3.1%	0.05	0.076	0.033
CP	0.0%	1.6%	0.0%	1.1%	0.01	0.0003	0.0001