SMART: SEQUENTIAL MONITORING WITH ADAPTIVE RISK TESTING FOR DYNAMIC A/B EXPERIMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Online experiment design, also known as A/B testing, is central to modern mobile health, yet existing work mainly compares policies through average treatment effects. Mean-based testing can overlook clinically important tail risks, particularly when outcomes are skewed, heavy-tailed, or safety-critical. Recent work on distributional treatment effects (DTE) highlights the value of evaluating policy impacts beyond the mean, offering a full characterization of policy effects and greater robustness to heavy-tailed outcomes. However, existing methods are largely designed for offline data or randomized experiments and do not directly address real-time monitoring of temporally dependent data streams. In this paper, we propose SMART, a scalable and interpretable framework for real-time A/B testing with adaptive risk in dynamic healthcare settings. SMART builds on a linear-quadratic dynamic model to characterize policy-induced return distributions over time while accounting for carryover effects. We establish the asymptotic joint normality of the test statistics, which underpins valid DTE inference, and introduce a sequential monitoring procedure with a bootstrap-based stopping criterion for efficient online experimentation. Experiments show that SMART can detect tail differences missed by mean testing and can stop substantially earlier than fixed-horizon inference, while also delineating its finite-sample and misspecification limitations.

1 INTRODUCTION

Online experiment design, also referred to as A/B testing, is a fundamental problem in policy evaluation, motivated by the need to rigorously compare novel policies with established baselines across diverse application domains. It plays a central role in areas such as medicine, pharmaceuticals, technology, and mobile health. For example, in clinical trials, A/B testing provides a principled framework for comparing two or more interventions, such as drug A versus drug B or alternative recruitment strategies. In web analytics, it is commonly used to compare variants of the same webpage by randomly assigning visitors to different versions and contrasting metrics of interest. In mobile health (mHealth), A/B testing is essential for evaluating Just-In-Time intervention delivery policies that continuously monitor patient states and deliver tailored treatments anytime and anywhere. These applications underscore the importance of rigorous A/B testing and motivate the need for methods that remain reliable in increasingly dynamic and data-rich settings.

The majority of the literature and many industry practices have focused on the Average Treatment Effect (ATE), designing hypothesis testing frameworks centered entirely on mean outcomesJohari et al. (2015); Kharitonov et al. (2015); Johari et al. (2017); Yang et al. (2017). While the mean is a critical statistic in risk-neutral settings, it inherently masks heterogeneity, particularly the tail risks. In safety-critical domains like mHealth, the average efficacy of a treatment is often secondary to its worst-case performance. For instance, an aggressive insulin dosing policy might yield a better average blood glucose level but occasionally trigger severe hypoglycemia—a catastrophic tail risk. Consequently, evaluating the Distributional Treatment Effect (DTE) offers a more comprehensive, robust, and risk-sensitive paradigm. Building on this, several distribution-based approaches, primarily focusing on the Quantile Treatment Effect (QTE), have been proposed and widely discussedChernozhukov & Hansen (2005); Firpo (2007); Callaway et al. (2018); Chernozhukov et al. (2020); Zhou et al. (2022).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Despite recent advancements, existing Distributional Treatment Effect methodologies struggle to address the complexities of modern real-world applications, particularly in mHealth. First, most approaches rely on the restrictive assumption of independence and require static, offline datasets. Consequently, they are ill-equipped for mHealth environments, which generate temporally correlated, online data streams requiring real-time sequential monitoring for a limited number of patients over extended decision trajectories. Second, current literature predominantly focuses on simple means or quantiles, largely ignoring broader risk measures like Conditional Value-at-Risk (CVaR) and expectiles. To provide tailored safety guarantees, a flexible A/B testing framework must accommodate these user-specified, comprehensive risk measures. Finally, while distributional reinforcement learning frameworks, such as C51 Bellemare et al. (2017), QR-DQN Dabney et al. (2018b), and IQN Dabney et al. (2018a), excel in complex environments, they face significant practical hurdles. Deriving analytical closed-form return distributions is inherently difficult, complicating the direct implementation of general risk measures. Moreover, their high computational cost restricts deployment on resource-constrained mobile devices, and their "black-box" nature fundamentally limits their utility in clinical settings where interpretability is a strict regulatory prerequisite.

1.1 Main Contributions

In this paper, we propose SMART (Sequential Monitoring with Adaptive Risk Testing for Dynamic A/B Experiments), a dynamic causal effect evaluation framework for online A/B testing, with mHealth serving as the primary motivating application. To the best of our knowledge, we are the first to formulate and solve the DTE hypothesis testing problem in an online data stream setting. Our main contributions are summarized as follows:

- We propose SMART, a distribution-aware framework for online A/B testing that evaluates policies beyond average treatment effects through distributional treatment effects, allowing inference on quantiles, variability, and tail-risk measures.
- We develop a linear quadratic dynamic formulation for sequential state–treatment–outcome data, providing a tractable mechanism for capturing carryover effects and aggregating information over time. Under this framework, we establish asymptotic joint normality, which provides the theoretical foundation for valid simultaneous inference on distributional treatment effects.
- We develop a sequential monitoring procedure with a pre-specified stopping rule, enabling sample-efficient online experimentation while preserving statistical validity.
- We evaluate SMART against baselines and under nonlinear, time-varying, and heavy-tailed misspecification; we further audit model fit, preprocessing, patient heterogeneity, and sequential-design sensitivity.

1.2 Related Works

A/B testing. There is a growing literature on A/B testing ((see Larsen et al., 2024; Quin et al., 2024, for recent reviews)), whose core idea is to leverage causal inference methods to estimate and make inference on the treatment effect of a new policy. Traditional A/B testing methods are designed for independent and identically distributed observations and typically assume the absence of the carryover effects. More recent proposals have focused on dynamic settings, where randomization is conducted over time in the presence of carryover effects. Naturally, existing methods from the causal inference literature designed to handle carryover effects are applicable (Robins (1986); Sobel & Lindquist (2014); Bojinov & Shephard (2019)). In parallel, a growing body of work has adopted an RL (Sutton et al. (1998)) framework for A/B testing, by leveraging the widely studied Markov decision process (Puterman (2014)) model in RL to explicitly capture the carryover effect (Glynn et al. (2020); Farias et al. (2022); Li et al. (2023); Shi et al. (2023); Chen et al. (2024); Wen et al. (2025)). Other related works have explored (i) sequential monitoring, which conducts A/B testing at multiple interim stages to enable early termination without inflating the overall type-I error (Johari et al. (2015); Waudby-Smith et al. (2024)); (ii) the careful design of experiments to enhance the efficiency of ATE estimators (Bojinov et al. (2023); Xiong et al. (2024); Li et al. (2023); Sun et al. (2024); Wen et al. (2025); Ni & Bojinov (2025)); (iii) methods for handling interference effects beyond temporal carryover effects, such as spatial, network, or marketplace interference (Ugander et al. (2013); Hu et al. (2022); Bajari et al. (2021); Leung (2022); Viviano et al. (2023)). Existing

sequential A/B testing methods mainly target mean effects or value functions, whereas our target is risk-sensitive distributional treatment effects.

Linear Quadratic Dynamic System. To address these challenges, we build upon the Linear Quadratic Regulator (LQR) framework Hambly et al. (2021); Toso et al. (2024). Compared to general MDPs, LQR provides a tractable structure that preserves essential temporal dependencies while facilitating rigorous, computationally efficient asymptotic analysis. Recent literature has established strong inferential foundations for LQR, exploring exact asymptotics Wang & Janson (2021) and online policy gradient inference Pan et al. (2025). Most critically, Wang et al. (2023) derives the closed-form distribution of random returns under LQR dynamics. Our contribution is not the LQR return representation itself: we propagate online OLS uncertainty through that representation, accommodate general regular risk functionals, derive a joint limit over multiple monitoring times, and calibrate a sequential stopping rule. Thus, Shi et al. (2023) studies dynamic mean-effect A/B testing but not DTE, whereas Wang et al. (2023) studies distributional LQR policy evaluation but not statistical testing, online parameter uncertainty, or sequential Type-I error control. We use LQR not as a general model for all mHealth systems, but as a tractable working model that enables closed-form return distribution approximation and interpretable inference.

Distributional Treatment Effect. Distributional and quantile treatment effects have long been recognized as important estimands beyond the mean effects. The quantile treatment effect was first introduced by Doksum (1974); Lehmann & D'Abrera (2006). Subsequently, estimation and inference methods for distributional and quantile treatment effects have been developed and applied in econometrics, statistics and machine learning communities, including Imbens & Rubin (1997); Abadie (2002); Chernozhukov & Hansen (2005); Koenker (2005); Firpo (2007); Chernozhukov et al. (2013); Callaway et al. (2018); Chernozhukov et al. (2020); Ge et al. (2020); Zhou et al. (2022); Park et al. (2021); Kallus & Oprescu (2023); Oka et al. (2026); Näf & Susmann (2024); Xu et al. (2025), among others. Most of this work explores conditional distributional and quantile treatment effects. Oka et al. (2026) and Byambadalai et al. (2024) consider the estimation of unconditional distributional treatment effects but under simple random sampling. Kallus et al. (2024) addresses problems in which nuisance parameters depend on the target parameter itself, as seen in cases like the quantile treatment effect (QTE) and local QTE. Recently, Li et al. (2024) introduced a testing procedure for dynamic conditional quantile treatment effects that effectively accounts for spatio-temporal dependence. However, their approach relies on fixed-horizon offline testing and is restricted solely to quantiles, lacking both a sequential monitoring mechanism for streaming data and the flexibility to incorporate a broader class of risk measures.

2 Problem Setting

Linear Quadratic Dynamic System. Consider a dynamic system where $x_t \in \mathbb{R}^d$ denotes the state variable and $u_t \in \mathbb{R}^p$ denotes the action variable. At each time index t, the next state x_{t+1} depends on both the current state x_t and an internal or external action u_t. Both adaptive control and reinforcement learning address the problem of selecting actions u_t when the system dynamics, namely the relationship between x_{t+1} and (x_t, u_t), are unknown.

In this paper, we consider a canonical model for such systems, the discrete-time linear dynamical system:

$$x_{t+1} = Ax_t + Bu_t + w_t,$$

where $A \in \mathbb{R}^{d\times d}$ and $B \in \mathbb{R}^{d\times p}$ characterize the system's linear dynamics, while the noise increments w_t are independently and identically drawn from an unknown distribution $\mathcal{D}$ with bounded moments. Our objective is defined by the instantaneous quadratic reward $r_t = x_t^\top \mathbf{Q} x_t + u_t^\top \mathbf{R} u_t$, where $\mathbf{Q} \in \mathbb{R}^{d\times d}$ and $\mathbf{R} \in \mathbb{R}^{p\times p}$ are positive-definite matrices. This formulation is closely related to the classical LQR problem, for which the optimal control law is known to yield a linear feedback policy of the form $u_t = Kx_t$. Motivated by this classical result, we restrict the treatment policy class in this paper to linear functions; specifically, the policy takes the form $\pi(x) = Kx$. This is a deliberate transparency–flexibility tradeoff: the structure captures carryover, yields an analyzable return law, and permits joint sequential inference. Adaptive LQR and iterative LQR provide routes to unknown or locally nonlinear systems, but extending SMART's inferential guarantees to those settings remains future work.

Exact Return Distribution. Assessing a policy's long-term performance under stochastic dynamics naturally necessitates analyzing its cumulative return. While standard reinforcement learning evaluates policies based on the expected return (i.e., the value function $V^K(x) = \mathbb{E}[G^K(x)]$), our risk-sensitive framework requires characterizing the exact return distribution. We define the conditional discounted return under a linear policy $\pi(x) = Kx$ as:

$$G^K(x) = \sum_{t=0}^{\infty} \gamma^t \left(x_t^\top \mathbf{Q} x_t + u_t^\top \mathbf{R} u_t \right), \quad x_0 = x.$$

Let F_x^K denote the cumulative distribution function of $G^K(x)$. To bypass intractable multi-step convolutions, we leverage the exact analytical expansion of this random return derived by Wang et al. (2023):

$$G^K(x) = x^\top P x + \sum_{k=0}^{\infty} \gamma^{k+1} w_k^\top P w_k + 2 \sum_{k=0}^{\infty} \gamma^{k+1} w_k^\top P A_K^{k+1} x + 2 \sum_{k=1}^{\infty} \gamma^{k+1} w_k^\top P \sum_{\tau=0}^{k-1} A_K^{k-\tau} w_\tau,$$

where $A_K = A + BK$ represents the closed-loop system matrix, P is the unique solution to the discrete Lyapunov equation $P = \mathbf{Q} + K^\top \mathbf{R} K + \gamma A_K^\top P A_K$, and $\{w_k\}_{k \geq 0}$ are i.i.d. noise increments drawn from the underlying distribution $\mathcal{D}$.

Risk-sensitive A/B Testing. The analytical characterization of the cumulative return enables us to evaluate risk profiles that are often obscured by expected returns. This distributional perspective motivates our risk-sensitive A/B testing framework, which compares baseline and new policies beyond their mean performance.

The primary goal of classical A/B testing is to evaluate the difference in average returns between two treatments, a concept closely related to the ATE. Formally, let matrices K_1 and K_0 represent the new and old policies, respectively, and let $\mathbb{G}$ denote a reference state distribution. The ATE is defined as:

$$\text{ATE} = \int_x \mathbb{E}[G^{K_1}(x) - G^{K_0}(x)] \mathbb{G}(dx).$$

However, simple expectations cannot capture the full distributional information or underlying risk profiles. Consequently, we shift our focus to the risk-sensitive A/B testing problem. To formalize this, we introduce a binary treatment indicator $a_t \in \{0, 1\}$, where $a_t = 1$ implies the execution of the new policy ($u_t = K_1 x_t$) at time t, and $a_t = 0$ corresponds to the old ($u_t = K_0 x_t$). For notational simplicity, let F_x^1 and F_x^0 denote the cumulative return distributions of $G^{K_1}(x)$ and $G^{K_0}(x)$, respectively. For a given risk measure κ, the Conditional Distributional Treatment Effect (CDTE) is defined as:

$$\text{CDTE}(x, \kappa) = \kappa(F_x^1) - \kappa(F_x^0).$$

The global DTE, denoted by δ, is then obtained by integrating the CDTE over the reference distribution, i.e., $\delta = \int_x \text{CDTE}(x, \kappa) \mathbb{G}(dx)$. The primary focus of this paper is to test the following hypotheses:

$$H_0 : \delta \leq 0 \quad \text{vs.} \quad H_1 : \delta > 0.$$

If the null hypothesis H_0 holds, the new policy does not improve upon the baseline with respect to the evaluated risk measure.

3 Test Procedure

The proposed methodology consists of three primary components: DTE estimation, variance estimation, and the determination of sequential monitoring thresholds, which we detail in the following subsections.

Estimation of DTE. First, we estimate the underlying transition dynamics of the environment. Given the observed trajectory data $\{(x_t, u_t, x_{t+1})\}_{t=1}^T$, we estimate the system matrices $(\widehat{A}, \widehat{B})$ via ordinary least squares regression:

$$(\widehat{A}_T, \widehat{B}_T) = \arg\min_{A,B} \frac{1}{T} \sum_{t=1}^{T} \|x_{t+1} - Ax_t - Bu_t\|^2 .$$

Using these estimates, we extract the residual sequence $\widehat{w}_t = x_{t+1} - \widehat{A}x_t - \widehat{B}u_t$. These empirical residuals serve as a nonparametric approximation of the underlying, unknown noise distribution $\mathcal{D}$.

Second, we employ a residual bootstrap procedure to evaluate the return distributions. Fixing the estimated dynamics $(\widehat{A}, \widehat{B})$, we sample with replacement from the empirical residuals to generate Q bootstrap noise sequences, denoted as $\{\widehat{w}_t^{(q)}\}_{t\geq 0}$ for $q = 1, \ldots, Q$. For each bootstrap sample and a given policy K, we substitute the simulated noise sequence into the analytical expansion of the random return to derive the following estimator:

$$\widehat{G}_T^{K,(q)}(x) = x^\top \widehat{P}x + \sum_{k=0}^{\infty} \gamma^{k+1} \widehat{w}_k^{(q)\top} \widehat{P} \widehat{w}_k^{(q)} + 2\sum_{k=0}^{\infty} \gamma^{k+1} \widehat{w}_k^{(q)\top} \widehat{P} \widehat{A}_K^{k+1} x + 2\sum_{k=1}^{\infty} \gamma^{k+1} \widehat{w}_k^{(q)\top} \widehat{P} \sum_{\tau=0}^{k-1} \widehat{A}_K^{k-\tau} \widehat{w}_\tau^{(q)},$$

where $\widehat{A}_K = \widehat{A}_T + \widehat{B}_T K$ represents the closed-loop system matrix, and $\widehat{P}$ is the solution to the discrete Lyapunov equation $P = \mathbf{Q} + K^\top RK + \gamma \widehat{A}_K^\top P \widehat{A}_K$.

This expansion enables us to simulate a comprehensive set of target returns $\{\widehat{G}_T^{K,(q)}(x)\}_{q=1}^Q$. Using these samples, we construct the empirical cumulative distribution functions, $\widehat{F}_{x,T}^1$ and $\widehat{F}_{x,T}^0$, evaluated under the new policy K_1 and the old policy K_0, respectively. Finally, the CDTE and the global DTE can be estimated by evaluating the chosen risk measure κ on these empirical distributions:

$$\widehat{\mathrm{CDTE}}_T(x, \kappa) = \kappa(\widehat{F}_{x,T}^1) - \kappa(\widehat{F}_{x,T}^0) \quad \text{and} \quad \widehat{\delta}_T = \int_x \widehat{\mathrm{CDTE}}_T(x, \kappa)\mathbb{G}(dx), \tag{1}$$

where $\mathbb{G}$ denotes the reference distribution of the initial state x_0. The point estimator alone is insufficient for inference; we next quantify its uncertainty via a multiplier bootstrap.

Estimation of Variance. To conduct valid statistical inference, it is crucial to account for the estimation uncertainty inherent in the initial approximation of the transition dynamics (A, B). We address this by employing a multiplier bootstrap method Wu (1986); Chen & Zhou (2020). Specifically, let $z_t = (x_t^\top, u_t)^\top$, and $\widehat{\Sigma}_T = \frac{1}{T}\sum_{t=1}^T z_t z_t^\top$,

$$\widehat{\eta}_T^{(q)} = \frac{1}{T}\sum_{t=1}^{T} \zeta_t^{(q)} z_t x_{t+1}^\top, \quad \zeta_t^{(q)} \sim N(0,1),$$
$$(\widehat{A}_T^{(q)}, \widehat{B}_T^{(q)}) = \widehat{\Sigma}_T^{-1} \widehat{\eta}_T^{(q)}. \tag{2}$$

For each bootstrap iteration $q \in \{1, \ldots, Q\}$, we obtain a perturbed parameter estimate $(\widehat{A}_T^{(q)}, \widehat{B}_T^{(q)})$. We then repeat the aforementioned residual bootstrap and CDTE estimation procedures using these updated transition matrices to compute the corresponding estimate $\widehat{\mathrm{CDTE}}_T^{(q)}$, and subsequently, $\widehat{\delta}_T^{(q)}$. Using these Q bootstrap estimates, the variance of the DTE estimator can be empirically approximated by:

$$\widehat{\mathrm{Var}}(\widehat{\delta}_T) = \frac{1}{Q}\sum_{q=1}^{Q} \left(\widehat{\delta}_T^{(q)} - \frac{1}{Q}\sum_{j=1}^{Q} \widehat{\delta}_T^{(j)} \right)^2. \tag{3}$$

This variance estimator enables confidence interval construction and formal hypothesis testing.

Finally, to evaluate the distributional treatment effects, we construct a t-type test statistic grounded in studentized bootstrap principles, defined as $Z_T = \frac{\widehat{\delta}_T}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\delta}_T)}}$. Under the null hypothesis $H_0 : \delta \leq 0$, the statistic Z_T converges in distribution to a standard normal variable, $\mathcal{N}(0,1)$, as the sample size $T \to \infty$. This construction ensures that the estimate is properly normalized by its empirical standard deviation, a property that will be formally justified in the subsequent theoretical analysis. For a given significance level α, the rejection region is defined as $\mathcal{R}_\alpha = \{Z_T : Z_T > c\}$, where c denotes a critical threshold that depends on the sequential monitoring horizon T. The explicit derivation of this threshold is deferred to the next section. If the observed statistic Z_T falls within $\mathcal{R}_\alpha$, we reject the null hypothesis, concluding that the new policy yields a statistically significant improvement in the system's long-term risk profile.

Sequential Monitoring. To facilitate early termination and improve resource efficiency, we adopt a sequential monitoring procedure based on the α-spending function approach Johari et al. (2015); Kharitonov et al. (2015); Shi et al. (2023). This framework enables adaptive decision-making as data batches arrive while rigorously controlling the overall Type-I error rate.

We partition the experimental timeline into M pre-determined interim analysis points, denoted by $1 \leq T_1 < T_2 < \cdots < T_M = T$. To maintain the nominal significance level α over multiple comparisons, we define a monotonically increasing spending function $\alpha(\cdot)$ satisfying $\alpha(0) = 0$ and $\alpha(T) = \alpha$. At each monitoring stage T_m, we utilize the cumulative data history to re-estimate the value functions and compute the following standardized test statistic:

$$Z_{T_m} = \frac{\widehat{\delta}_{T_m}}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\delta}_{T_m})}}. \tag{4}$$

The procedure adopts a stopping rule based on a sequentially determined critical boundary $\widehat{c}_m$: if $Z_{T_m} > \widehat{c}_m$, we reject the null hypothesis H_0 and terminate the experiment immediately. Otherwise, the trial proceeds to the next scheduled monitoring point T_{m+1}. If H_0 cannot be rejected at the final stage T_M, we fail to reject the null hypothesis.

Unlike classical sequential analysis, the exact joint distribution of $\{Z_{T_m}\}_{m=1}^M$ and its asymptotic covariance matrix are analytically intractable due to temporal data correlations and potential adaptive treatment designs. To resolve this, we approximate the joint distribution using a scalable multiplier bootstrap procedure Shi et al. (2023). Specifically, we introduce independent standard normal multipliers $\zeta_t^* \sim \mathcal{N}(0,1)$ into Equation 2 to obtain the perturbed score $\widehat{\eta}_{T_m}^*$. The bootstrapped system parameters are then estimated as $(\widehat{A}_{T_m}^*, \widehat{B}_{T_m}^*) = \widehat{\Sigma}_{T_m}^{-1}\widehat{\eta}_{T_m}^*$. This allows us to recompute the residuals $\widehat{w}_t^*$ and derive the bootstrapped DTE estimate $\delta_{T_m}^*$. We then construct the corresponding bootstrapped test statistic:

$$Z_{T_m}^* = \frac{\delta_{T_m}^*}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\delta}_{T_m})}}, \tag{5}$$

where the empirical variance $\widehat{\mathrm{Var}}(\widehat{\delta}_{T_m})$ is computed as in Equation 3. Crucially, we establish that the asymptotic covariance structure of the bootstrapped sequence $\{Z_{T_m}^*\}_{m=1}^M$ is equivalent to that of the original sequence $\{Z_{T_m}\}_{m=1}^M$. Consequently, conditional on the observed data, the empirical distribution of the bootstrap samples provides a valid approximation to the joint distribution of the original statistics.

Based on these bootstrap samples, the procedure for determining the boundaries $\{\widehat{c}_m\}_{m=1}^M$ is detailed in Algorithm 1. By aligning the empirical crossing probabilities with the α-spending schedule, we guarantee valid asymptotic inference. As supported by Theorems 2 and 3 in Shi et al. (2023), under mild regularity conditions, our procedure strictly controls the cumulative Type-I error such that $\Pr\left(\bigcup_{j=1}^m \{Z_{T_j} > \widehat{c}_j\}\right) = \alpha(T_m) + o(1)$ under H_0. Furthermore, the test is consistent, with the rejection probability approaching 1 when the true effect size satisfies $\delta \gg T^{-1/2}$.

4 Theoretical Results

This section establishes the theoretical foundations of our proposed methodology. The primary objective is to derive the asymptotic normality of the test statistics and to theoretically justify the validity of the sequential monitoring procedure. This justification is achieved by demonstrating the asymptotic equivalence between the multiplier bootstrap samples and the original samples. Before proceeding to the main convergence results, we first impose the following regularity condition on the mathematical properties of the risk measure κ.

Assumption 1. *We assume that the functional $\kappa(\cdot)$ admits a second-order Bahadur-type expansion at the population distribution F. Specifically, for the empirical distribution F_n based on n samples, there exists a measurable influence function $\phi(X)$ with $\mathbb{E}[\phi(X)] = 0$ and $\mathbb{E}[\phi^2(X)] < \sigma^2 < \infty$, such that:*

$$\kappa(F_n) = \kappa(F) + \frac{1}{n}\sum_{i=1}^n \phi(X_i) + R_n, \tag{6}$$

where the remainder term satisfies $R_n = o_p(n^{-1/2})$ as $n \to \infty$. This condition implies that the functional is Hadamard-differentiable, ensuring that the non-linear quantile-based estimator can be asymptotically linearized without loss of first-order statistical information.

Having established the asymptotic linear representation of κ via its influence function, it is necessary to impose structural conditions on the class of these influence functions. Such conditions are essential to control the complexity of the function class, thereby ensuring that the resulting estimators satisfy the requirements for asymptotic normality.

Assumption 2. *Let $\mathcal{F}_\phi = \{\phi(x;\theta) : \theta \in \Theta\}$ be the class of influence functions given by Equation 6 associated with the family of functionals under consideration, where Θ is a compact parameter space. We assume that $\mathcal{F}_\phi$ is a P-Donsker class.*

With the aforementioned functional and structural assumptions in place, we are now positioned to formally establish the asymptotic normality of the proposed test statistics across the sequential monitoring stages.

Theorem 1. *Under conditions (C1)–(C5) in Appendix B.1 and Assumptions 1–2, let $\{Z_{T_m}\}_{1\le m\le M}$ be defined as in Equation 4. Then:*

- *the statistics $\{Z_{T_m}\}_{1\le m\le M}$ are jointly asymptotically normal;*
- *their asymptotic means are non-positive under H_0;*
- *their covariance matrix can be consistently estimated by a matrix $\widehat{\Gamma}$, where the closed form of $\widehat{\Gamma}$ is defined by Equation 8.*

The theorem explicitly incorporates transition-model estimation. In Appendix B.2, we derive the uniform first-order expansion of $(\widehat{A}, \widehat{B})$ and propagate its influence through $\widehat{A}_K$, the Lyapunov solution $\widehat{P}$, the simulated return law, and $\widehat{\delta}_{T_m}$. Hence OLS error is part of the influence function and of $\widehat{\Gamma}$; it is not treated as a fixed plug-in quantity. Building upon the results of Theorem 1, we can further verify the theoretical connection between the original test statistics and their multiplier bootstrap counterparts utilized in our algorithm. This relationship is crucial for confirming the structural validity of the proposed sequential inference procedure.

Corollary 1. *Under the same condition as Theorem 1 , let $\{Z^*_{T_m}\}_{1\le m\le M}$ be the bootstrap sample defined by Equation 5. Then the covariance matrices of $\{Z_{T_m}\}_{1\le m\le M}$ and $\{Z^*_{T_m}\}_{1\le m\le M}$ are asymptotically equivalent. Consequently, the distribution of $\{Z_{T_m}\}_{1\le m\le M}$ can be consistently approximated by $\{Z^*_{T_m}\}_{1\le m\le M}$.*

In summary, this subsection provides the rigorous theoretical justification for our inferential framework. These results establish the asymptotic validity of the proposed dynamic sequential monitoring procedure and its control of statistical error rates.

5 Experiments

We evaluate our sequential monitoring framework on both synthetic and real datasets. The synthetic data generation procedure, the parameter details and introduction to OhioT1DM dataset are provided in Appendix C.1. We study four complementary questions: statistical performance comparison of exsisting methods, sensitivity to risk measure, robustness beyond exact LQ dynamics, and capability in applying real clinical data.

Performance of Statistical Power. We firstly consider the $\kappa =$ Mean case. To isolate the immediate glucose-to-insulin response, all entries of K in $u_t = Kx_t$ are set to zero except the coefficient for current blood glucose, denoted by ρ. Fixing the new policy at $\rho_1 = 0.1$, we define the policy discrepancy as $\Delta = \rho_1 - \rho_0$. Monte Carlo simulations verify that $\Delta < 0$ yields a strictly positive true DTE (the alternative hypothesis), with the signal weakening as $\Delta \to 0$. Expect for the SMART, *Bonferroni* retains the same DTE estimator, standardized statistic, and stopping logic, but uses the constant boundary, corresponding to a one-sided $\alpha/5$ test at each of five looks. *Fixed-horizon DTE* uses the same estimator but tests only once at $T = 500$ with boundary, so it avoids repeated-testing

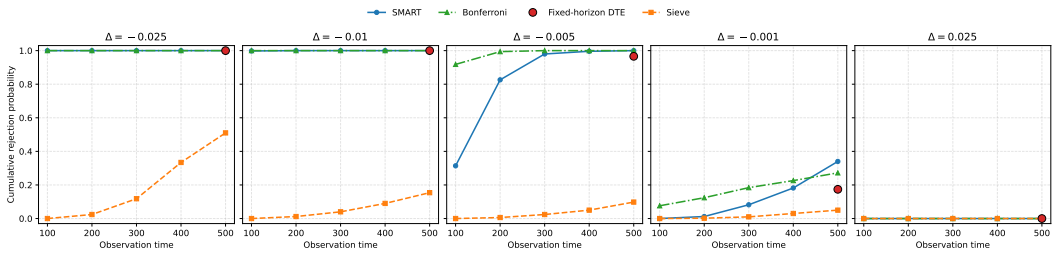


Figure 1: Cumulative rejection probabilities for SMART, Bonferroni, fixed-horizon DTE, and Sieve. Fixed-horizon DTE is evaluated only at $T = 500$ and is highlighted by large red points; Sieve is compared only for the shared expectation target. Each configuration uses 500 replications.

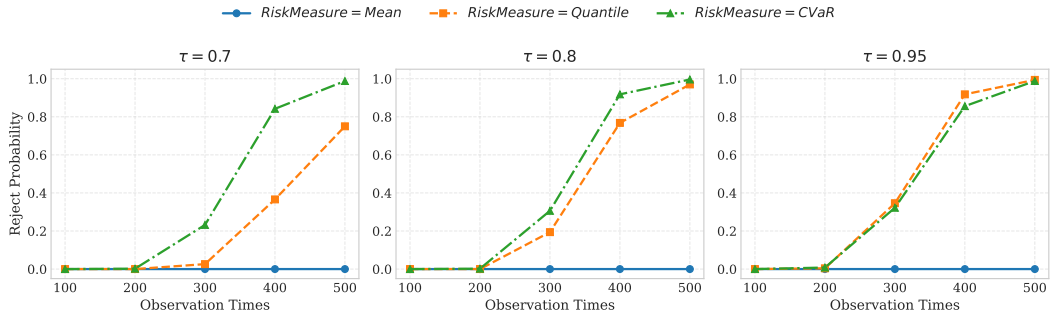


Figure 2: Empirical rejection probabilities for expectation, quantile, and CVaR DTEs in the synthetic experiment. Panels correspond to $\tau \in \{0.7, 0.8, 0.95\}$; each curve uses 500 replications.

multiplicity but cannot stop early. *Sieve* (Shi et al., 2023) is a sequential dynamic A/B test based on sieve approximation of mean Q-functions; it therefore suit for the $\kappa =$ Mean case.

Figure 1 uses the policy grid $\Delta \in \{-0.025, -0.01, -0.005, -0.001, 0.025\}$. At $\Delta = -0.005$, SMART and Bonferroni both attain final rejection probability 1.00 with expected stopping times 188.4 and 108.8, versus rejection 0.966 and stopping time 500 for fixed-horizon DTE. At the weakest alternative ($\Delta = -0.001$), their rejection probabilities are 0.340, 0.272, and 0.174, respectively. Hence the calibrated sequential procedures retain the same DTE target while providing an opportunity for materially earlier decisions.

Adaptivity to Tail Risks and Failure of ATE. We next highlight our method's adaptability when traditional ATE testing fails to capture critical tail risks. We configure two policies ($\rho_1 = 0.9, \rho_2 = 0.025$) such that their true DTE under standard expectation ($\kappa =$ Mean) is exactly zero, but strictly positive under Quantile and CVaR measures. We evaluate the statistical power across these risk measures for probability levels $\tau \in \{0.7, 0.8, 0.95\}$.

As shown in Figure 2, the Mean-based test completely fails to reject H_0, with its empirical rejection probability bounded at zero across all monitoring steps. Conversely, our framework utilizing Quantile and CVaR successfully detects the underlying structural tail differences. The rejection probabilities for both measures grow significantly with observation time, approaching 1.0 by the final node $T = 500$ for extreme levels ($\tau = 0.8$ and 0.95). Furthermore, CVaR consistently exhibits a steeper rejection trajectory than Quantile. These results verify our framework's capacity to identify severe tail risks even when ATE is entirely blind to policy discrepancies, enabling flexible sequential monitoring tailored to specific risk-aversion demands.

OhioT1DM Evaluation. Finally, we evaluate our framework on the OhioT1DM dataset. To verify correctness of the pooled LQ model, we conduct a diagnostics and sensitivity analysis. Full

diagnostics and sensitivity results are reported in Appendix C.2.2.
Justified by this audit, we conduct the proposed method on OhioT1DM. Setting the new policy to $\rho_1 = 0.5$, we test policy discrepancies $\Delta \in \{-0.05, 0.03, 0.05, 0.07\}$. Note that in this configuration, a positive Δ yields a strictly positive true DTE (the alternative hypothesis), with the effect size increasing monotonically with Δ.
Figure 3 supports the effectiveness of our algorithm on the OhioT1DM-based evaluation across Mean, Quantile, and CVaR measures. Under the null hypothesis ($\Delta = -0.05$), empirical rejection rates remain strictly zero, demonstrating rigorous Type-I error control. Conversely, when the alternative hypothesis holds ($\Delta > 0$), statistical power scales monotonically with signal strength. Under a strong signal ($\Delta = 0.07$), the rejection probability reaches 1.0 by $T = 500$ across all statistics, decreasing predictably as the signal attenuates ($\Delta \in \{0.05, 0.03\}$). Crucially, our framework detects complex tail risk improvements (Quantile and CVaR) with sensitivity matching the standard Mean statistic, proving its clinical utility in safeguarding against extreme glycemic fluctuations alongside average outcome improvements.

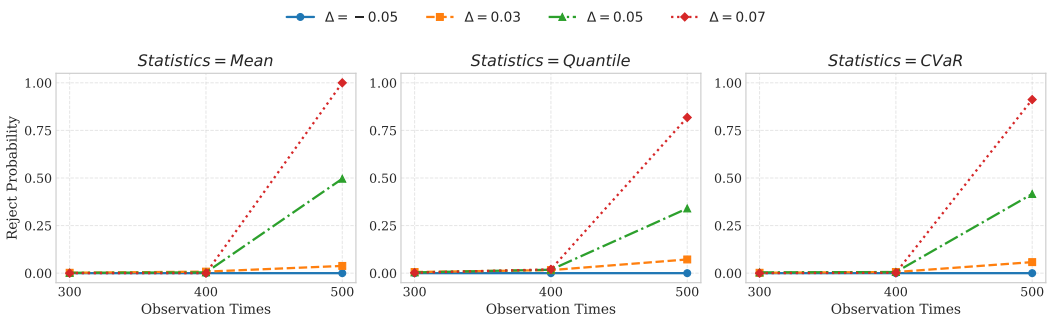


Figure 3: Empirical rejection probabilities evaluated under different risk measures across sequential observation times on the real-world dataset. The subplots from left to right correspond to the target statistics: Mean, Quantile, and CVaR, respectively. The distinct lines within each subplot illustrate the testing power under different Δ. The empirical probabilities are computed based on 500 replications.

Robustness and Design Sensitivity. We further stress-test SMART in seven non-medical energy-storage environments and vary bootstrap size, monitoring times, alpha spending, and the discount factor. Strong nonlinearity is the clearest misspecification failure mode, while alternative monitoring and spending schedules expose finite-sample power–validity tradeoffs. Full designs and results appear in Appendices C.2.1 and C.2.3.

6 Conclusion

In this work, we proposed a risk-sensitive sequential monitoring framework that enables safe evaluation and early stopping for A/B testing in dynamic environments. Power performance comparisons demonstrate the practical value of calibrated early stopping. The OhioT1DM results shows clinical utility in safeguarding against extreme glycemic fluctuations alongside average outcome improvements. While our method is designed to control the Type I error and performs well in the considered mHealth-inspired settings, its primary limitation lies in the reliance on fixed system matrices A and B within the LQ framework. Adaptive and iterative LQR are promising extensions, but nonlinear or time-varying SMART requires new theory for valid sequential distributional inference.

AI use statement

We used generative AI tools to improve the clarity and readability of the manuscript and to help identify relevant literature. AI tools also provided feedback on aspects of the research methodology, assisted in designing some robustness and parameter sensitivity experiments. The authors reviewed the AI-assisted work, verified the methods and experimental results, and take full responsibility for the final manuscript.

REFERENCES

Alberto Abadie. Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American statistical Association*, 97(457):284–292, 2002.

Patrick Bajari, Brian Burdick, Guido W Imbens, Lorenzo Masoero, James McQueen, Thomas Richardson, and Ido M Rosen. Multiple randomization designs. *arXiv preprint arXiv:2112.13495*, 2021.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pp. 449–458. PMLR, 2017.

Iavor Bojinov and Neil Shephard. Time series experiments and causal estimands: exact randomization tests and trading. *Journal of the American Statistical Association*, 114(528):1665–1682, 2019.

Iavor Bojinov, David Simchi-Levi, and Jinglong Zhao. Design and analysis of switchback experiments. *Management Science*, 69(7):3759–3777, 2023.

Undral Byambadalai, Tatsushi Oka, and Shota Yasui. Estimating distributional treatment effects in randomized experiments: Machine learning for variance reduction. In *International Conference on Machine Learning*, pp. 5082–5113. PMLR, 2024.

Brantly Callaway, Tong Li, and Tatsushi Oka. Quantile treatment effects in difference in differences models under dependence restrictions and with only two time periods. *Journal of Econometrics*, 206(2):395–413, 2018.

Shuze Chen, David Simchi-Levi, and Chonghuan Wang. Experimenting on markov decision processes with local treatments. *arXiv e-prints*, pp. arXiv–2407, 2024.

Xi Chen and Wen-Xin Zhou. Robust inference via multiplier bootstrap. *The Annals of Statistics*, 48(3):pp. 1665–1691, 2020. ISSN 00905364, 21688966. URL `https://www.jstor.org/stable/26931527`.

Victor Chernozhukov and Christian Hansen. An iv model of quantile treatment effects. *Econometrica*, 73(1):245–261, 2005.

Victor Chernozhukov, Iván Fernández-Val, and Blaise Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

Victor Chernozhukov, Ivan Fernandez-Val, Blaise Melly, and Kaspar Wüthrich. Generic inference on quantile and quantile effect functions for discrete outcomes. *Journal of the American Statistical Association*, 2020.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International Conference on Machine Learning*, pp. 1096–1105. PMLR, 2018a.

Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018b.

Kjell Doksum. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The Annals of Statistics*, pp. 267–277, 1974.

Vivek Farias, Andrew Li, Tianyi Peng, and Andrew Zheng. Markovian interference in experiments. *Advances in Neural Information Processing Systems*, 35:535–549, 2022.

Sergio Firpo. Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75(1): 259–276, 2007.

Qiyang Ge, Xuelin Huang, Shenying Fang, Shicheng Guo, Yuanyuan Liu, Wei Lin, and Momiao Xiong. Conditional generative adversarial networks for individualized treatment effect estimation and treatment selection. *Frontiers in Genetics*, 11:585804, 2020.

Peter W Glynn, Ramesh Johari, and Mohammad Rasouli. Adaptive experimental design with temporal interference: A maximum likelihood approach. *Advances in Neural Information Processing Systems*, 33:15054–15064, 2020.

Ben Hambly, Renyuan Xu, and Huining Yang. Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *SIAM Journal on Control and Optimization*, 59(5):3359–3391, 2021.

Yuchen Hu, Shuangning Li, and Stefan Wager. Average direct and indirect causal effects under interference. *Biometrika*, 109(4):1165–1172, 2022.

Guido W Imbens and Donald B Rubin. Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies*, 64(4):555–574, 1997.

Ramesh Johari, Leo Pekelis, and David J Walsh. Always valid inference: Bringing sequential analysis to a/b testing. *arXiv preprint arXiv:1512.04922*, 2015.

Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. 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*, pp. 1517–1525, 2017.

Nathan Kallus and Miruna Oprescu. Robust and agnostic learning of conditional distributional treatment effects. In *International Conference on Artificial Intelligence and Statistics*, pp. 6037–6060. PMLR, 2023.

Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond. *Journal of Machine Learning Research*, 25 (16):1–59, 2024.

Eugene Kharitonov, Aleksandr Vorobev, Craig Macdonald, Pavel Serdyukov, and Iadh Ounis. Sequential testing for early stopping of online experiments. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 473–482, 2015.

Roger Koenker. Quantile regression [m]. *Econometric Society Monographs, Cambridge University Press, Cambridge*, 2005.

Nicholas Larsen, Jonathan Stallrich, Srijan Sengupta, Alex Deng, Ron Kohavi, and Nathaniel T Stevens. Statistical challenges in online controlled experiments: A review of a/b testing methodology. *The American Statistician*, 78(2):135–149, 2024.

Erich Leo Lehmann and Howard JM D'Abrera. *Nonparametrics: statistical methods based on ranks*, volume 464. Springer New York, 2006.

Michael P Leung. Rate-optimal cluster-randomized designs for spatial interference. *The Annals of Statistics*, 50(5):3064–3087, 2022.

Ting Li, Chengchun Shi, Jianing Wang, Fan Zhou, et al. Optimal treatment allocation for efficient policy evaluation in sequential decision making. *Advances in Neural Information Processing Systems*, 36:48890–48905, 2023.

Ting Li, Chengchun Shi, Zhaohua Lu, Yi Li, and Hongtu Zhu. Evaluating dynamic conditional quantile treatment effects with applications in ridesharing. *Journal of the American Statistical Association*, 119(547):1736–1750, 2024. doi: 10.1080/01621459.2024.2314316.

Milena Soriano Marcolino, João Antonio Queiroz Oliveira, Marcelo D'Agostino, Antonio Luiz Ribeiro, Maria Beatriz Moreira Alkmim, and David Novillo-Ortiz. The impact of mhealth interventions: systematic review of systematic reviews. *JMIR mHealth and uHealth*, 6(1):e23, 2018.

Jeffrey Näf and Herbert Susmann. Causal-drf: Conditional kernel treatment effect estimation using distributional random forest. *arXiv preprint arXiv:2411.08778*, 2024.

Tu Ni and Iavor Bojinov. Enhancing efficiency and robustness for switchback experiments: A practical model-assisted framework. *Available at SSRN 5229804*, 2025.

Tatsushi Oka, Shota Yasui, Yuta Hayakawa, and Undral Byambadalai. Regression adjustment for estimating distributional treatment effects in randomized controlled trials. *Econometric Reviews*, 45(1):2–17, 2026.

Bo Pan, Jianya Lu, Yafei Wang, Hao Li, Bei Jiang, and Linglong Kong. Toward optimal statistical inference in noisy linear quadratic reinforcement learning over a finite horizon. *arXiv preprint arXiv:2508.08436*, 2025.

Junhyung Park, Uri Shalit, Bernhard Schölkopf, and Krikamol Muandet. Conditional distributional treatment effect with kernel conditional mean embeddings and u-statistic regression. In *International Conference on Machine Learning*, pp. 8401–8412. PMLR, 2021.

Temiloluwa Prioleau, Abigail Bartolome, Richard Comi, and Catherine Stanger. Diatrend: A dataset from advanced diabetes technology to enable development of novel analytic solutions. *Scientific Data*, 10(1):556, 2023.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Federico Quin, Danny Weyns, Matthias Galster, and Camila Costa Silva. A/b testing: A systematic literature review. *Journal of Systems and Software*, 211:112011, 2024.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7 (9-12):1393–1512, 1986.

Chengchun Shi, Runzhe Wan, Rui Song, Wenbin Lu, and Ling Leng. Does the markov decision process fit the data: Testing for the markov property in sequential decision making. In *International Conference on Machine Learning*, pp. 8807–8817. PMLR, 2020.

Chengchun Shi, Xiaoyu Wang, Shikai Luo, Hongtu Zhu, Jieping Ye, and Rui Song. Dynamic causal effects evaluation in a/b testing with a reinforcement learning framework. *Journal of the American Statistical Association*, 118(543):2059–2071, 2023.

Michael E Sobel and Martin A Lindquist. Causal inference for fmri time series data with systematic errors of measurement in a balanced on/off study of social evaluative threat. *Journal of the American Statistical Association*, 109(507):967–976, 2014.

Ke Sun, Linglong Kong, Hongtu Zhu, and Chengchun Shi. Arma-design: Optimal treatment allocation strategies for a/b testing in partially observable time series experiments. *arXiv preprint arXiv:2408.05342*, 2024.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Leonardo Felipe Toso, Donglin Zhan, James Anderson, and Han Wang. Meta-learning linear quadratic regulators: a policy gradient maml approach for model-free lqr. In *6th Annual Learning for Dynamics & Control Conference*, pp. 902–915. PMLR, 2024.

Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 329–337, 2013.

Davide Viviano, Lihua Lei, Guido Imbens, Brian Karrer, Okke Schrijvers, and Liang Shi. Causal clustering: design of cluster experiments under network interference. *arXiv preprint arXiv:2310.14983*, 2023.

Feicheng Wang and Lucas Janson. Exact asymptotics for linear quadratic adaptive control. *Journal of Machine Learning Research*, 22(265):1–112, 2021.

Zifan Wang, Yulong Gao, Siyi Wang, Michael M Zavlanos, Alessandro Abate, and Karl Henrik Johansson. Policy evaluation in distributional lqr. In *Learning for Dynamics and Control Conference*, pp. 1245–1256. PMLR, 2023.

Ian Waudby-Smith, Lili Wu, Aaditya Ramdas, Nikos Karampatziakis, and Paul Mineiro. Anytime-valid off-policy inference for contextual bandits. *ACM/IMS Journal of Data Science*, 1(3):1–42, 2024.

Qianglin Wen, Chengchun Shi, YingYang, Niansheng Tang, and Hongtu Zhu. Unraveling the interplay between carryover effects and reward autocorrelations in switchback experiments. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=ZwcMZ443BF.

Chien-Fu Jeff Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295, 1986.

Ruoxuan Xiong, Alex Chin, and Sean J Taylor. Data-driven switchback experiments: Theoretical tradeoffs and empirical bayes designs. *arXiv preprint arXiv:2406.06768*, 2024.

Ruofan Xu, Jiti Gao, Tatsushi Oka, and Yoon-Jae Whang. Quantile random-coefficient regression with interactive fixed effects: Heterogeneous group-level policy evaluation. *Econometric Reviews*, 44(5):630–648, 2025.

Fanny Yang, Aaditya Ramdas, Kevin G Jamieson, and Martin J Wainwright. A framework for multi-a (rmed)/b (andit) testing with online fdr control. *Advances in Neural Information Processing Systems*, 30, 2017.

Tianhui Zhou, William E Carson IV, and David Carlson. Estimating potential outcome distributions with collaborating causal networks. *Transactions on Machine Learning Research*, 2022.

A Additional Technical Materials

A.1 Algorithm

Algorithm 1 Sequential A/B Testing for Distributional Treatment Effects

Require: Data $\{(x_t, u_t, x_{t+1})\}_{t=1}^{T+1}$, policies K_0, K_1, risk measure κ, bootstrap size Q, α spending function $\alpha(\cdot)$, projection operator Proj defined in Appendix A.2, time points $T_1, T_2, \cdots, T_M$.

Set $T_0 = 0$, $\mathcal{I} = \{1, 2, \cdots, Q\}$. Initialize $\widehat{\Sigma}$, $\widehat{\eta}$ and $\widehat{\eta}^{(q)}$ as zero matrix.
for $m = 1, \ldots, M$ **do**
  Estimate $(\widehat{A}, \widehat{B})$ via linear regression using data $\{(x_t, u_t, x_{t+1})\}_{t=1}^{T_m}$ and let $z_t = (x_t^\top, u_t)^\top$.

$$\widehat{\Sigma} = \frac{T_{m-1}}{T_m}\widehat{\Sigma} + \frac{1}{T_m}\sum_{t=T_{m-1}}^{T_m} z_t z_t^\top;$$

$$\widehat{\eta} = \frac{T_{m-1}}{T_m}\widehat{\eta} + \frac{1}{T_m}\sum_{t=T_{m-1}}^{T_m} z_t x_{t+1}^\top.$$

  Set $(\widehat{A}, \widehat{B}) = \widehat{\Sigma}^{-1}\widehat{\eta}$ and conduct projection $\widehat{A} = \mathrm{Proj}(\widehat{A}, \widehat{B}, K)$.
  Compute residuals $\widehat{w}_t = x_{t+1} - \widehat{A}x_t - \widehat{B}u_t$ and $\widehat{\delta}$ by Equation 1.
  for $q = 1, \ldots, Q$ **do**
    Re-estimate $(\widehat{A}^{(q)}, \widehat{B}^{(q)})$ via multiplier bootstrap

$$\widehat{\eta}^{(q)} = \frac{T_{m-1}}{T_m}\widehat{\eta}^{(q)} + \frac{1}{T_m}\sum_{t=T_{m-1}}^{T_m} \zeta_t^{(q)} z_t x_{t+1}^\top, \quad \zeta_t^{(q)} \sim N(0,1) \quad \text{for} \quad T_{m-1} \le t \le T_m.$$

    Set $(\widehat{A}^{(q)}, \widehat{B}^{(q)}) = \widehat{\Sigma}^{-1}\widehat{\eta}^{(q)}$ and conduct projection $\widehat{A}^{(q)} = \mathrm{Proj}(\widehat{A}^{(q)}, \widehat{B}^{(q)}, K)$.
    Compute residuals $\widehat{w}_t^{(q)} = x_{t+1} - \widehat{A}^{(q)}x_t - \widehat{B}^{(q)}u_t$ and $\delta^{(q)}$.
  end for
  Compute $\widehat{\mathrm{Var}}(\widehat{\delta})$ by Equation 3.
  Set

$$\widehat{Z}^{(q)} = \frac{\delta^{(q)}}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\delta})}}. \tag{7}$$

  Set c_m to the upper $\{\alpha(t) - |\mathcal{I}^c|/Q\} / (1 - |\mathcal{I}^c|/Q)$-th quantile of $\{\widehat{Z}^{(q)}\}_{q\in\mathcal{I}}$.
  Update $\mathcal{I} \leftarrow \{q \in \mathcal{I}, \widehat{Z}^{(q)} \le c_m\}$.
  Reject the null if $\widehat{\delta}/\sqrt{\widehat{\mathrm{Var}}(\widehat{\delta})} > c_m$.
end for

A.2 Stationarity Guarantee

In the identification of the discrete-time linear dynamical system $x_{t+1} = (A + BK)x_t + w_t$, unconstrained estimation of $A_K = A + BK$ does not guarantee system stationarity. For example, as evidenced by the spectral analysis of estimating $\widehat{A}_K$ under small sample size, certain eigenvalues $\lambda_i(\widehat{A}_K)$ exhibit a spectral radius larger than one, leading to divergent trajectories. To improve robustness and enforce stationarity, we employ a spectral projection method onto the unit disk.

Given that A_K is non-symmetric, we consider its eigendecomposition $A_K = V\Lambda V^{-1}$, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ contains potentially complex-valued eigenvalues. For each eigenvalue λ_i, we define a contraction mapping to ensure it resides within the stable manifold $\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$. To preserve the underlying oscillatory frequency and the real-valued nature of the reconstructed matrix, we apply a modulus-only shrinkage that maintains the phase $\angle\lambda_i$:

$$\tilde{\lambda}_i = \begin{cases} (1-\epsilon)\frac{\lambda_i}{|\lambda_i|}, & \text{if } |\lambda_i| \geq 1 \\ \lambda_i, & \text{if } |\lambda_i| < 1 \end{cases},$$

where $\epsilon > 0$ is a small stability margin (e.g., $\epsilon = 10^{-3}$). This transformation ensures that complex conjugate pairs $\lambda_i, \bar{\lambda}_i$ remain conjugate, thereby guaranteeing that the stabilized matrix $\tilde{A}_K = V\tilde{\Lambda}V^{-1}$ remains in the real domain $\mathbb{R}^{n\times n}$.

To incorporate this stability constraint while preserving the fixed policy K and the input matrix B, we attribute the spectral discrepancy to the uncertainty in the autonomous dynamics. Specifically, we define the projection operator $\mathrm{Proj}(\cdot,\cdot,\cdot)$ as $\tilde{A} = \mathrm{Proj}(A, B, K) = \tilde{A}_K - BK$, which effectively projects the closed-loop dynamics into the stable parameter space while preserving the structural coupling between the control input and the state evolution. The resulting system is guaranteed to be stable by construction, as $\max_i |\lambda_i(\tilde{A} + BK)| \leq 1 - \epsilon$.

Rationale for the Stationary Projection. The theoretical and practical justification for employing this projection operator is grounded in three key considerations. First, our entire framework operates under the fundamental assumption that the underlying true dynamic system is inherently stable; the analysis of intrinsically unstable systems falls outside the scope of this paper. Second, in the finite-sample regime, the OLS estimates $\widehat{A}$ and $\widehat{B}$ can exhibit substantial estimation variance. This statistical error may artificially push the eigenvalues of the estimated closed-loop matrix $\widehat{A}_K$ outside the unit circle, despite the true system being strictly stable. In such scenarios, the projection is a practically indispensable intervention to prevent mathematically divergent simulated trajectories and ensure the viability of the algorithm. Finally, in the large-sample asymptotic regime, the consistency of the OLS estimators guarantees that $\widehat{A}$ and $\widehat{B}$ will tightly converge to their true, stable values. Consequently, as the sample size T grows, the estimated eigenvalues will naturally fall strictly within the stable manifold $\mathbb{D}$, so that the projection operator asymptotically reduces to the identity map. Therefore, applying this stabilization step is highly effective and necessary for small-sample robustness, while remaining theoretically innocuous to the large-sample asymptotic properties of the system.

B Proofs

B.1 Additional Conditions

(C1) Stability of the Closed-Loop System $A + BK$: The eigenvalues of the closed-loop matrix $A_K = A + BK$ must be strictly bounded within the unit circle, i.e., $\max_i |\lambda_i(A_K)| \leq 1 - \epsilon$ for some stability margin $\epsilon > 0$. Furthermore, to ensure the absolute convergence of the infinite-horizon discounted return, the spectral radius must satisfy $\rho(\sqrt{\gamma}A_K) < 1$.

(C2) Geometric Ergodicity and Moment Conditions: The underlying data generating process $\{(x_t, u_t, x_{t+1})\}$ must form a geometrically ergodic Markov chain. Additionally, there must exist a constant $\delta_0 > 0$ such that the $(2+\delta_0)$-th moments are bounded: $\mathbb{E}[\|z_t\|^{2+\delta_0}] < \infty$ and $\mathbb{E}[\|\epsilon_{t+1}\|^{2+\delta_0}] < \infty$, where $z_t = (x_t^\top, u_t^\top)^\top$ and ϵ_{t+1} represents the system innovation noise.

(C3) Non-degeneracy of the OLS Regressor z_t: The population covariance matrix of the augmented state-action regressors, defined as $\Sigma := \mathbb{E}[z_t z_t^\top]$, must be strictly positive definite ($\Sigma \succ 0$). This ensures that the empirical covariance matrix $\widehat{\Sigma}_m$ is invertible with probability tending to one as $T \to \infty$.

(C4) Validity of the Bootstrap Multipliers: The random multipliers $\zeta_t^{(q)}$ used in the multiplier bootstrap procedure must be drawn independently of the observed data trajectory. They must satisfy the moment matching conditions $\mathbb{E}[\zeta_t^{(q)}] = 0$ and $\mathrm{Var}(\zeta_t^{(q)}) = 1$ (e.g., standard normal multipliers $\zeta_t^{(q)} \sim \mathcal{N}(0,1)$).

(C5) Growth Condition for the Bootstrap Size Q: To guarantee that the Monte Carlo simulation error of the empirical return distribution is asymptotically negligible compared to the statistical estimation

error, the bootstrap sample size Q_T must grow to infinity at a rate satisfying $Q_T^{-1/2} = o(T^{-1/2})$ as $T \to \infty$.

Proof of Theorem 1 and Corollary 1. We organize the proof into seven steps. Throughout, let

$$\theta_0 := (A, B), \qquad \widehat{\theta}_m := (\widehat{A}_{T_m}, \widehat{B}_{T_m}),$$

and let

$$\varepsilon_{t+1} := x_{t+1} - Ax_t - Bu_t.$$

Let $\mathcal{F}_t := \sigma\{(x_s, u_s, \varepsilon_{s+1}) : 1 \le s \le t\}$. By the conditional mean independence assumption,

$$\mathbb{E}(\varepsilon_{t+1} \mid \mathcal{F}_t) = 0.$$

We also assume the usual moment conditions implied by geometric ergodicity and quadratic rewards: there exists $\delta_0 > 0$ such that

$$\mathbb{E}\|z_t\|^{2+\delta_0} < \infty, \qquad \mathbb{E}\|\varepsilon_{t+1}\|^{2+\delta_0} < \infty.$$

Let

$$\Sigma := \mathbb{E}(z_t z_t^\top),$$

and assume Σ is positive definite. We further assume the monitoring fractions satisfy

$$T_m/T \to \rho_m \in (0, 1], \qquad 1 \le m \le M,$$

with fixed M. Finally, to justify the algorithmic construction of $\widehat{F}_x^K$ from Q simulated bootstrap returns, we impose the standard simulation-growth condition

$$Q = Q_T \to \infty, \qquad Q_T^{-1/2} = o(T^{-1/2}),$$

so that Monte Carlo error in the empirical return law is asymptotically negligible relative to statistical error. We also interpret the algorithmic bootstrap studentization as

$$Z^*_{T_m} = \frac{\delta^{(q)}_{T_m}}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\delta}_{T_m})}},$$

i.e. with the same square-root denominator as Z_{T_m}.

B.2 Online least squares: consistency and uniform first-order expansion

For stage m, unrolling the online recursion gives

$$\widehat{\Sigma}_m = \frac{1}{T_m}\sum_{t=1}^{T_m} z_t z_t^\top, \qquad \widehat{\eta}_m = \frac{1}{T_m}\sum_{t=1}^{T_m} z_t x_{t+1}^\top.$$

Hence the unprojected estimator is

$$\widetilde{\theta}_m := \widehat{\Sigma}_m^{-1}\widehat{\eta}_m.$$

Because

$$x_{t+1}^\top = z_t^\top \theta_0^\top + \varepsilon_{t+1}^\top,$$

we have

$$\widehat{\eta}_m = \frac{1}{T_m}\sum_{t=1}^{T_m} z_t z_t^\top \theta_0^\top + \frac{1}{T_m}\sum_{t=1}^{T_m} z_t \varepsilon_{t+1}^\top = \widehat{\Sigma}_m \theta_0^\top + \frac{1}{T_m}\sum_{t=1}^{T_m} z_t \varepsilon_{t+1}^\top.$$

Therefore

$$\widetilde{\theta}_m - \theta_0 = \widehat{\Sigma}_m^{-1}\Big(\frac{1}{T_m}\sum_{t=1}^{T_m} z_t \varepsilon_{t+1}^\top\Big).$$

By geometric ergodicity, the law of large numbers for additive functionals of Markov chains yields

$$\widehat{\Sigma}_m \to_p \Sigma$$

uniformly over the finite set $m = 1, \dots, M$. Since Σ is nonsingular, with probability tending to one all $\widehat{\Sigma}_m$ are invertible and

$$\widehat{\Sigma}_m^{-1} = \Sigma^{-1} + o_p(1)$$

uniformly in m.

Next, the sequence $\{z_t \varepsilon_{t+1}^\top, \mathcal{F}_t\}$ is a square-integrable martingale difference array because

$$\mathbb{E}(z_t \varepsilon_{t+1}^\top \mid \mathcal{F}_t) = z_t \mathbb{E}(\varepsilon_{t+1}^\top \mid \mathcal{F}_t) = 0.$$

Hence, for each fixed m,

$$\frac{1}{\sqrt{T_m}} \sum_{t=1}^{T_m} \operatorname{vec}(z_t \varepsilon_{t+1}^\top) = O_p(1),$$

and the finite-dimensional martingale CLT applies jointly across the nested partial sums $T_1, \dots, T_M$.

Combining these facts,

$$\sqrt{T_m} \operatorname{vec}(\widetilde{\theta}_m - \theta_0) = (\mathbf{I} \otimes \Sigma^{-1}) \frac{1}{\sqrt{T_m}} \sum_{t=1}^{T_m} \operatorname{vec}(z_t \varepsilon_{t+1}^\top) + o_p(1),$$

uniformly over $m = 1, \dots, M$.

Now consider the projection step. Let Π_K denote the projection map used to enforce the structural constraint attached to policy K. Assume, as stated in the problem, that the true parameter satisfies the constraint and that Π_K is Hadamard differentiable at θ_0 with derivative equal to the identity on the relevant tangent space. Then

$$\widehat{\theta}_m - \widetilde{\theta}_m = o_p(T_m^{-1/2}),$$

uniformly in m. Consequently,

$$\sqrt{T_m} \operatorname{vec}(\widehat{\theta}_m - \theta_0) = (\mathbf{I} \otimes \Sigma^{-1}) \frac{1}{\sqrt{T_m}} \sum_{t=1}^{T_m} \operatorname{vec}(z_t \varepsilon_{t+1}^\top) + o_p(1)$$

uniformly in m. This proves consistency and the required first-order linearization of $(\widehat{A}, \widehat{B})$.

B.3 Propagation through $\widehat{A}_K$, $\widehat{P}$, and $\widehat{G}^K(x)$

Fix $K \in \{K_0, K_1\}$. Write

$$A_K := A + BK, \qquad \widehat{A}_{K,m} := \widehat{A}_m + \widehat{B}_m K.$$

From Step 1,

$$\sqrt{T_m} \operatorname{vec}(\widehat{A}_{K,m} - A_K) = O_p(1),$$

uniformly in m.

By assumption, the discounted return expansion is well-defined; equivalently, for each fixed K,

$$\rho(\sqrt{\gamma}\, A_K) < 1,$$

where $\rho(\cdot)$ denotes the spectral radius. By continuity of the spectral radius and consistency of $\widehat{A}_{K,m}$, with probability tending to one,

$$\rho(\sqrt{\gamma}\, \widehat{A}_{K,m}) < 1$$

for every m. Thus the discrete Lyapunov equation

$$P_K = \mathbf{Q} + K^\top R K + \gamma A_K^\top P_K A_K$$

has a unique solution, and similarly

$$\widehat{P}_{K,m} = \mathbf{Q} + K^\top R K + \gamma \widehat{A}_{K,m}^\top \widehat{P}_{K,m} \widehat{A}_{K,m}$$

has a unique solution.

Define the Lyapunov operator

$$\mathcal{L}_K(P,\Theta) := P - \gamma A_K(\Theta)^\top P A_K(\Theta) - \{\mathbf{Q} + K^\top R K\},$$

where $A_K(\Theta) = A + BK$ when $\Theta = (A, B)$. Since $\rho(\sqrt{\gamma}A_K) < 1$, the derivative of $\mathcal{L}_K$ with respect to P is invertible. Therefore, by the implicit function theorem, the map

$$\Theta \mapsto P_K(\Theta)$$

is continuously Fréchet differentiable in a neighborhood of θ_0. Hence

$$\sqrt{T_m}\,\mathrm{vec}(\widehat{P}_{K,m} - P_K) = D_\Theta P_K(\theta_0)\Big[\sqrt{T_m}\,\mathrm{vec}(\widehat{\theta}_m - \theta_0)\Big] + o_p(1),$$

uniformly in m.

Next consider the residuals

$$\widehat{w}_{t,m} := x_{t+1} - \widehat{A}_m x_t - \widehat{B}_m u_t = \varepsilon_{t+1} - (\widehat{\theta}_m - \theta_0) z_t.$$

Hence, uniformly in m,

$$\widehat{w}_{t,m} - \varepsilon_{t+1} = -(\widehat{\theta}_m - \theta_0) z_t = O_p(T_m^{-1/2}) \cdot \|z_t\|.$$

Because $\mathbb{E}\|z_t\|^{2+\delta_0} < \infty$, this perturbation is square-integrable.

Now define $G^K(x)$ as the population counterpart of $\widehat{G}^K(x)$, with P_K, A_K, and the innovation sequence $\{\varepsilon_t\}$ in place of $\widehat{P}, \widehat{A}_K, \{\widehat{w}_t\}$. Every term in the discounted series for $\widehat{G}^K(x)$ is a finite-degree polynomial in $(\widehat{P}_{K,m}, \widehat{A}_{K,m}, \widehat{w}_{t,m})$, multiplied by powers of γ, and the series is absolutely summable because $\rho(\sqrt{\gamma}A_K) < 1$. Therefore the map

$$(\Theta, \nu) \mapsto G^K(x; \Theta, \nu)$$

from the parameter Θ and innovation law ν to the return is Fréchet differentiable in a neighborhood of (θ_0, ν_0), where ν_0 is the law of ε_t. Applying the chain rule to the expansions above yields, for each fixed x,

$$\sqrt{T_m}\big(\widehat{G}^K_m(x) - G^K(x)\big) = \frac{1}{\sqrt{T_m}} \sum_{t=1}^{T_m} \psi_{G,K}(x; \mathcal{Y}_t) + o_p(1)$$

uniformly in m, where $\mathcal{Y}_t := (x_t, u_t, x_{t+1})$, and $\psi_{G,K}(x; \mathcal{Y}_t)$ is a square-integrable influence function depending linearly on the score term $\mathrm{vec}(z_t \varepsilon_{t+1}^\top)$ from Step 1 and on the first-order perturbation of the residual empirical law. Because the return map is smooth and the discounted coefficients are summable, $\psi_{G,K}(x; \cdot)$ admits an integrable envelope in x with respect to $\mathbb{G}$.

B.4 Bahadur expansion of κ for the empirical return laws

For each $K \in \{K_0, K_1\}$ and x, let F_x^K denote the population law of $G^K(x)$. Let $\widetilde{F}^K_{x,m}$ denote the "ideal" conditional law generated by $(\widehat{\theta}_m, \widehat{P}_{K,m}, \widehat{w}_{t,m})$, before replacing it by the empirical distribution built from the finite Monte Carlo sample $\{\widehat{G}^{K,(q)}(x)\}_{b=1}^{B_T}$. By construction of $\widehat{F}_x^K$,

$$\widehat{F}_x^K - \widetilde{F}^K_{x,m}$$

is purely Monte Carlo error conditional on the data. Since $\widehat{F}_x^K$ is an empirical distribution based on Q_T draws,

$$\sup_x d(\widehat{F}_x^K, \widetilde{F}^K_{x,m}) = O_p(Q_T^{-1/2}),$$

for any metric d dominating bounded Lipschitz functionals. Our condition $Q_T^{-1/2} = o(T^{-1/2})$ therefore implies

$$\sup_x \big|\kappa(\widehat{F}_x^K) - \kappa(\widetilde{F}^K_{x,m})\big| = o_p(T_m^{-1/2}),$$

provided κ is locally Lipschitz along the Bahadur expansion, which is exactly what Assumption 1 supplies.

By Assumption 1, for each x and each K,

$$\kappa(\widetilde{F}^K_{x,m}) = \kappa(F^K_x) + \int \phi(g; x, K)\, d\big(\widetilde{F}^K_{x,m} - F^K_x\big)(g) + r_{m,K}(x),$$

with

$$\sup_x |r_{m,K}(x)| = o_p(T_m^{-1/2}).$$

Assumption 2 implies that the class

$$\mathcal{F}_\phi = \{\phi(\cdot; x, K, \vartheta) : (x, K, \vartheta) \in \mathrm{supp}(\mathbb{G}) \times \{K_0, K_1\} \times \Theta\}$$

is P-Donsker; combined with the consistency of $\widehat{\theta}_m$ and the asymptotic linearity of $\widehat{G}^K_m(x)$, this yields stochastic equicontinuity and allows us to replace the perturbation of $\widetilde{F}^K_{x,m}$ by its first-order influence representation. Therefore,

$$\sqrt{T_m}\big(\kappa(\widehat{F}^K_x) - \kappa(F^K_x)\big) = \frac{1}{\sqrt{T_m}} \sum_{t=1}^{T_m} \psi_{\kappa,K}(x; \mathcal{Y}_t) + o_p(1)$$

uniformly in m, where $\psi_{\kappa,K}$ is the influence function obtained by composing the return influence $\psi_{G,K}$ with the derivative of κ.

Hence the conditional distributional treatment effect satisfies

$$\widehat{\mathrm{CDTE}}_m(x, \kappa) - \mathrm{CDTE}(x, \kappa) = \frac{1}{T_m} \sum_{t=1}^{T_m} \Delta\psi_\kappa(x; \mathcal{Y}_t) + o_p(T_m^{-1/2}),$$

uniformly in m, where

$$\Delta\psi_\kappa(x; \mathcal{Y}_t) := \psi_{\kappa,K_1}(x; \mathcal{Y}_t) - \psi_{\kappa,K_0}(x; \mathcal{Y}_t).$$

B.5 Integration over $\mathbb{G}$ and asymptotic linearization of $\widehat{\delta}_{T_m}$

By definition,

$$\widehat{\delta}_{T_m} = \int \widehat{\mathrm{CDTE}}_m(x, \kappa)\, \mathbb{G}(dx), \qquad \delta = \int \mathrm{CDTE}(x, \kappa)\, \mathbb{G}(dx).$$

Because $\Delta\psi_\kappa(x; \mathcal{Y}_t)$ has an integrable envelope in x and the remainder is uniform in x, Fubini's theorem yields

$$\sqrt{T_m}\,(\widehat{\delta}_{T_m} - \delta) = \frac{1}{\sqrt{T_m}} \sum_{t=1}^{T_m} \psi_t + o_p(1),$$

where

$$\psi_t := \int \Delta\psi_\kappa(x; \mathcal{Y}_t)\, \mathbb{G}(dx).$$

Equivalently,

$$\widehat{\delta}_{T_m} - \delta = \frac{1}{T_m} \sum_{t=1}^{T_m} \psi_t + o_p(T_m^{-1/2})$$

uniformly over $m = 1, \ldots, M$.

The centering of ψ_t follows from the definition of influence functions:

$$\mathbb{E}(\psi_t) = 0.$$

Moreover, since ψ_t is obtained by composing square-integrable influence terms with smooth maps and integrating over $\mathbb{G}$, we have $\mathbb{E}|\psi_t|^{2+\delta} < \infty$ for some $\delta > 0$.

Under the null $H_0 : \delta \le 0$, the deterministic component of $\widehat{\delta}_{T_m}$ is nonpositive. Thus

$$\sqrt{T_m}\,\widehat{\delta}_{T_m} = \sqrt{T_m}\,\delta + \frac{1}{\sqrt{T_m}} \sum_{t=1}^{T_m} \psi_t + o_p(1).$$

Hence the drift term $\sqrt{T_m}\delta$ is componentwise nonpositive. On the boundary null $\delta = 0$, the mean vector is 0; under local nulls $\delta = h/\sqrt{T}$ with $h \le 0$, the limiting mean vector is finite and componentwise nonpositive; for fixed strict null $\delta < 0$, the same representation implies a negative diverging drift, which is conservative for upper-tail rejection.

B.6 JOINT CLT FOR $(\widehat{\delta}_{T_1}, \ldots, \widehat{\delta}_{T_M})$ AND CONSTRUCTION OF $\widehat{\Gamma}$

Set

$$S_T(r) := \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} \psi_t, \qquad r \in [0,1].$$

Because $\{\mathcal{Y}_t\}$ is geometrically ergodic and ψ_t is a square-integrable additive functional of this chain, the functional central limit theorem for geometrically ergodic Markov chains applies. Under the conditional mean independence condition, the leading influence process is a square-integrable martingale difference sequence, so

$$S_T(\cdot) \Rightarrow \sigma_\psi W(\cdot) \quad \text{in } D[0,1],$$

where W is standard Brownian motion and

$$\sigma_\psi^2 = \lim_{T\to\infty} \mathrm{Var}\Big(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \psi_t\Big) > 0.$$

In the martingale-difference case,

$$\sigma_\psi^2 = \mathbb{E}(\psi_t^2),$$

while under the more general mixing interpretation one may equivalently write

$$\sigma_\psi^2 = \sum_{k\in\mathbb{Z}} \mathrm{Cov}(\psi_0, \psi_k).$$

By the asymptotic linear representation from Step 4 and the continuous mapping theorem,

$$\Big(\sqrt{T_1}(\widehat{\delta}_{T_1} - \delta), \ldots, \sqrt{T_M}(\widehat{\delta}_{T_M} - \delta)\Big) \Rightarrow N(0, \Omega),$$

where for $i \le j$,

$$\Omega_{ij} = \mathrm{Cov}\Big(\frac{S_T(\rho_i)}{\sqrt{\rho_i}}, \frac{S_T(\rho_j)}{\sqrt{\rho_j}}\Big) \to \sigma_\psi^2 \sqrt{\frac{\rho_i}{\rho_j}},$$

and symmetry determines the remaining entries. Thus

$$\Omega_{ij} = \sigma_\psi^2 \sqrt{\frac{\rho_i \wedge \rho_j}{\rho_i \vee \rho_j}} = \frac{\sigma_\psi^2 \, \min(\rho_i, \rho_j)}{\sqrt{\rho_i \rho_j}}.$$

To estimate Ω, let $\widehat{\psi}_t$ be the plug-in version of ψ_t, obtained by replacing all population quantities by their estimators. By Steps 1–4 and stochastic equicontinuity,

$$\frac{1}{T} \sum_{t=1}^{T} (\widehat{\psi}_t - \psi_t)^2 \to_p 0.$$

Therefore any consistent estimator of σ_ψ^2 built from $\{\widehat{\psi}_t\}$ is valid. In the martingale-difference case,

$$\widehat{\sigma}_\psi^2 := \frac{1}{T} \sum_{t=1}^{T} \widehat{\psi}_t^2 \to_p \sigma_\psi^2.$$

More generally, one may use a HAC estimator. Define

$$\widehat{\Omega}_{ij} := \widehat{\sigma}_\psi^2 \sqrt{\frac{\rho_i \wedge \rho_j}{\rho_i \vee \rho_j}}$$

(or the corresponding HAC version). Then

$$\widehat{\Omega} \to_p \Omega.$$

This yields a consistent nondegenerate estimator for the covariance matrix of the unstudentized sequential vector.

B.7 STUDENTIZATION AND JOINT ASYMPTOTIC NORMALITY OF $(Z_{T_1}, \dots, Z_{T_M})$

We now prove consistency of the variance estimator actually used in the algorithm. For replicate q, define the weighted least-squares criterion

$$L_T^{(q)}(A, B) = \frac{1}{T} \sum_{t=1}^{T} (1 + \xi_t^{(q)}) \|x_{t+1} - Ax_t - Bu_t\|^2,$$

where $\mathbb{E}\xi_t^{(q)} = 0$, $\mathbb{E}(\xi_t^{(q)})^2 = 1$, and the multipliers are independent of the data. The weighted score equation gives

$$0 = \frac{1}{T} \sum_{t=1}^{T} (1 + \xi_t^{(q)}) z_t \big(x_{t+1} - \widehat{A}^{(q)} x_t - \widehat{B}^{(q)} u_t\big)^\top.$$

Linearizing this equation around $\widehat{\theta}_T$ and using the same matrix expansion as in Step 1 yields, conditionally on the data,

$$\sqrt{T}\,\mathrm{vec}(\widehat{\theta}_T^{(q)} - \widehat{\theta}_T) = (\mathbf{I} \otimes \widehat{\Sigma}_T^{-1}) \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \xi_t^{(q)}\, \mathrm{vec}(z_t \widehat{\varepsilon}_{t+1}^\top) + o_{p^*}(1),$$

where p^* denotes conditional probability given the data and $\widehat{\varepsilon}_{t+1} = x_{t+1} - \widehat{A}x_t - \widehat{B}u_t$.

Propagating this expansion through the same differentiable maps as in Steps 2–4 gives the conditional bootstrap linearization

$$\sqrt{T}\big(\widehat{\delta}_T^{(q)} - \widehat{\delta}_T\big) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \xi_t^{(q)} \widehat{\psi}_t + o_{p^*}(1)$$

uniformly in q, with $\widehat{\psi}_t$ the same plug-in influence proxy as above.

Conditional on the data,

$$\mathrm{Var}^*\Big(\sqrt{T}(\widehat{\delta}_T^{(q)} - \widehat{\delta}_T)\Big) = \frac{1}{T} \sum_{t=1}^{T} \widehat{\psi}_t^2 + o_p(1) \to_p \sigma_\psi^2.$$

As $Q \to \infty$, the empirical variance across bootstrap replicates consistently estimates the conditional variance, hence

$$T\,\widehat{\mathrm{Var}}(\widehat{\delta}_T) \to_p \sigma_\psi^2, \qquad T_m\,\widehat{\mathrm{Var}}(\widehat{\delta}_{T_m}) \to_p \sigma_\psi^2$$

for each m. In particular, the variance estimators are nondegenerate.

Therefore,

$$Z_{T_m} = \frac{\widehat{\delta}_{T_m}}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\delta}_{T_m})}} = \frac{\sqrt{T_m}\delta}{\sigma_\psi} + \frac{1}{\sigma_\psi \sqrt{T_m}} \sum_{t=1}^{T_m} \psi_t + o_p(1).$$

Hence, under the boundary null $\delta = 0$,

$$(Z_{T_1}, \dots, Z_{T_M}) \Rightarrow N(0, \Gamma),$$

where

$$\Gamma_{ij} = \sqrt{\frac{\rho_i \wedge \rho_j}{\rho_i \vee \rho_j}} = \frac{\min(\rho_i, \rho_j)}{\sqrt{\rho_i \rho_j}}. \tag{8}$$

Under local nulls $\delta = h/\sqrt{T}$ with $h \le 0$,

$$(Z_{T_1}, \dots, Z_{T_M}) \Rightarrow N(\mu, \Gamma), \qquad \mu_m = \frac{\sqrt{\rho_m}\, h}{\sigma_\psi} \le 0.$$

For fixed strict null $\delta < 0$, the same representation yields componentwise drift to $-\infty$, again with nonpositive asymptotic mean behavior. Thus the theorem's mean claim is established in the precise sense relevant for one-sided testing.

A consistent estimator $\widehat{\Gamma}$ is obtained by replacing σ_ψ^2 and the monitoring fractions by their consistent estimators in the covariance formula above; equivalently, because the studentized covariance depends only on $\rho_1, \dots, \rho_M$ in the martingale-difference case, one may take

$$\widehat{\Gamma}_{ij} := \sqrt{\frac{(T_i/T) \wedge (T_j/T)}{(T_i/T) \vee (T_j/T)}} = \frac{\min(T_i, T_j)}{\sqrt{T_i T_j}},$$

or its plug-in HAC analogue in the more general mixing setup. In either formulation,

$$\widehat{\Gamma} \to_p \Gamma.$$

This proves Theorem 1.

B.8 Multiplier bootstrap validity and stagewise boundary calibration

We now prove Corollary 1. Fix $m \le M$. The online bootstrap recursion satisfies

$$\widehat{\eta}_m^{(q)} = \frac{1}{T_m} \sum_{t=1}^{T_m} \zeta_t^{(q)} z_t x_{t+1}^\top,$$

because the recursion unrolls exactly as in Step 1. Therefore

$$\widehat{\theta}_m^{(q)} = \widehat{\Sigma}_m^{-1} \widehat{\eta}_m^{(q)}.$$

Since $\widehat{\Sigma}_m \to_p \Sigma$ and the multipliers are independent $N(0,1)$, conditionally on the data,

$$\sqrt{T_m}\, \mathrm{vec}(\widehat{\theta}_m^{(q)}) = (\mathbf{I} \otimes \widehat{\Sigma}_m^{-1}) \frac{1}{\sqrt{T_m}} \sum_{t=1}^{T_m} \zeta_t^{(q)}\, \mathrm{vec}(z_t x_{t+1}^\top) + o_{p^*}(1).$$

After recentering around the original estimator and applying the same projection-delta argument as in Step 1,

$$\sqrt{T_m}\, \mathrm{vec}\big(\widehat{\theta}_m^{(q)} - \widehat{\theta}_m\big) = (\mathbf{I} \otimes \Sigma^{-1}) \frac{1}{\sqrt{T_m}} \sum_{t=1}^{T_m} \zeta_t^{(q)}\, \mathrm{vec}(z_t \widehat{\varepsilon}_{t+1}^\top) + o_{p^*}(1).$$

Propagating this perturbation through the closed-loop map, the Lyapunov map, the residual map, the return map, and finally the risk functional κ, exactly as in Steps 2–4, gives

$$\sqrt{T_m}\, \delta_{T_m}^{(q)} = \frac{1}{\sqrt{T_m}} \sum_{t=1}^{T_m} \zeta_t^{(q)} \widehat{\psi}_t + o_{p^*}(1)$$

conditionally on the data, uniformly in m. Because the same influence proxy $\widehat{\psi}_t$ appears here as in the original statistic, the bootstrap reproduces the same first-order covariance structure.

Indeed, for $i \le j$,

$$\mathrm{Cov}^*\left(\frac{1}{\sqrt{T_i}} \sum_{t=1}^{T_i} \zeta_t \widehat{\psi}_t,\ \frac{1}{\sqrt{T_j}} \sum_{t=1}^{T_j} \zeta_t \widehat{\psi}_t \right) = \frac{1}{\sqrt{T_i T_j}} \sum_{t=1}^{T_i} \widehat{\psi}_t^2 \to_p \sigma_\psi^2 \sqrt{\frac{\rho_i}{\rho_j}},$$

which is exactly the covariance of the Gaussian limit of the original unstudentized vector. After division by the common studentizing factor $\sqrt{\widehat{\mathrm{Var}}(\widehat{\delta}_{T_m})}$, we obtain

$$\mathrm{Cov}^*(Z_{T_1}^*, \dots, Z_{T_M}^*) \to_p \Gamma.$$

Thus the bootstrap and original sequential vectors are asymptotically covariance equivalent.

Now Assumption 2 implies that the influence-function class is P-Donsker. Hence the conditional multiplier central limit theorem applies to the process

$$m \mapsto \frac{1}{\sqrt{T_m}} \sum_{t=1}^{T_m} \zeta_t^{(q)} \widehat{\psi}_t.$$

Therefore, conditionally on the data,

$$(Z^*_{T_1}, \dots, Z^*_{T_M}) \Rightarrow^* N(0, \Gamma)$$

in probability, while under the boundary null,

$$(Z_{T_1}, \dots, Z_{T_M}) \Rightarrow N(0, \Gamma).$$

Under local nulls, the bootstrap matches the centered Gaussian fluctuation part, which is sufficient for boundary calibration. Consequently,

$$\sup_{h \in \mathrm{BL}_1(\mathbb{R}^M)} \left| \mathbb{E}^* h(Z^*_{T_1}, \dots, Z^*_{T_M}) - \mathbb{E} h(Z_{T_1} - \mu_1, \dots, Z_{T_M} - \mu_M) \right| \to_p 0,$$

and in particular, on the boundary null $\mu = 0$,

$$\sup_{h \in \mathrm{BL}_1(\mathbb{R}^M)} \left| \mathbb{E}^* h(Z^*_{T_1}, \dots, Z^*_{T_M}) - \mathbb{E} h(Z_{T_1}, \dots, Z_{T_M}) \right| \to_p 0.$$

Finally, because the map $z \mapsto \max_{1 \le j \le m}(z_j - b_j)$ is continuous except on a Gaussian null set, the continuous mapping theorem transfers the bootstrap approximation to crossing events. Therefore the bootstrap quantiles computed from the active set $\mathcal{I}$ satisfy the recursive spending identity asymptotically:

$$\Pr\left\{ Z_{T_m} > \widehat{c}_m \,\middle|\, \max_{1 \le j < m} (Z_{T_j} - \widehat{c}_j) \le 0 \right\} = \frac{\alpha(T_m) - \alpha(T_{m-1})}{1 - \alpha(T_{m-1})} + o(1),$$

which implies

$$\Pr\left(\bigcup_{j=1}^{m} \{ Z_{T_j} > \widehat{c}_j \} \right) = \alpha(T_m) + o(1), \qquad 1 \le m \le M.$$

Hence the stagewise bootstrap boundaries provide asymptotically valid α-spending control for the sequential A/B testing procedure.

This proves Corollary 1 and completes the justification of Algorithm 1. □

C Additional Materials for Experiments

C.1 OhioT1DM Dataset and Sythetic Data Parameters

In this section, we evaluate the empirical performance of our proposed sequential monitoring framework. The experiments are conducted on both synthetic and real datasets. The data generation process for the synthetic experiments closely follows the simulation environment designed in Section 5.2.2 of Shi et al. (2020).

For the real-world empirical component of our study, we utilize the **OhioT1DM dataset** (Marcolino et al., 2018), a widely recognized benchmark resource containing continuous measurements for six patients with type 1 diabetes over an eight-week period. To formulate our Markov Decision Process (MDP) and maintain alignment with other contemporary datasets (e.g., the DiaTrend dataset (Prioleau et al., 2023)), we define the following features:

- **State ($\tilde{x}_t$):** Originally, the dataset tracks three components: average blood glucose levels (G_t), carbohydrate estimates for meals (C_t), and exercise intensity (Ex_t). Because exercise intensity is not consistently available across general mobile health devices, we exclude it from our representation. Thus, our base state is a 2-dimensional vector $\tilde{x}_t = (G_t, C_t)^\top$.
- **Action (u_t):** The original dataset categorizes insulin doses into five discrete levels, i.e., $\{0, 1, 2, 3, 4\}$, ranging from no dose to high doses. Because our methodology is designed to evaluate continuous interventions via a linear policy $u_t = Kx_t$, we treat the action space as a continuous variable ranging from 0 to 30. Specifically, we generate continuous actions by uniformly sampling within the corresponding continuous intervals based on the original discrete levels.

- **Reward (r_t):** The objective is aligned with the Index of Glycemic Control, which functions as a quadratic function of the average blood glucose levels during the time interval $[t, t+1)$. This quadratic structure naturally fits within our Linear Quadratic (LQ) framework.

To address concerns regarding the Markovian assumption in sequential decision-making, we rely on the testing results from Shi et al. (2020), which suggest that physiological data of this nature typically satisfy a 4-th order Markov property. Consequently, we reconstruct the state variable by concatenating past measurements and actions to meet this assumption. The augmented state vector incorporates an intercept term and is defined as:

$$x_t = (1, \tilde{x}_t^\top, u_{t-1}, \tilde{x}_{t-1}^\top, u_{t-2}, \tilde{x}_{t-2}^\top, u_{t-3}, \tilde{x}_{t-3}^\top)^\top.$$

Given that $\tilde{x}_t \in \mathbb{R}^2$ and $u_t \in \mathbb{R}$, this concatenation yields a 12-dimensional augmented state vector $x_t \in \mathbb{R}^{12}$.

Based on the linear dynamic assumption, the evolution of the augmented state can be explicitly decomposed into the following element-wise transitions. The first element remains a constant 1 (i.e., $x_{t+1}^{(1)} = 1$), while the remaining dimensions evolve as:

$$\begin{cases} x_{t+1}^{(2)} = \boldsymbol{l}^\top x_t + b u_t + \epsilon_t^{(2)}; \\ x_{t+1}^{(3)} = \mu + \epsilon_t^{(3)}; \\ x_{t+1}^{(4)} = u_t + \epsilon_t^{(4)}; \\ x_{t+1}^{(k)} = x_t^{(k-3)} + \epsilon_t^{(k)}, \quad \text{for } k = 5, \ldots, 12, \end{cases}$$

where $b = -0.070$, $\mu = 10.185$, and

$$\begin{aligned} \boldsymbol{l} = (&39.030, 1.171, -0.119, -0.056, -0.481, 0.049, \\ &-0.338, 0.106, 0.012, -0.055, -0.008, -0.009)^\top. \end{aligned}$$

This deterministic shifting of historical variables and linear dependence can be compactly expressed in the standard linear system form:

$$x_{t+1} = A x_t + B u_t + \epsilon_t,$$

where the transition matrices $A \in \mathbb{R}^{12\times 12}$ and $B \in \mathbb{R}^{12\times 1}$ are explicitly constructed as:

$$A = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ & & & & \boldsymbol{l}^\top & & & \\ \mu & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 & 0 & 0 \end{pmatrix}, \quad \text{and} \quad B = \begin{pmatrix} 0 \\ b \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Overall, this rigorous data preprocessing and environmental formulation provide a robust, physiologically informed foundation for evaluating our sequential monitoring approach.

C.2 Supplementary Empirical Audits

This appendix reports the compact numerical summaries underlying the revised empirical discussion. The statistical power comparison uses 500 replications; the cross-domain, OhioT1DM audit, and sequential-sensitivity cells use 100 outer Monte Carlo replications. With 100 replications, rejection-rate uncertainty is material; for example, the default sequential-design null estimate 0.09 has Wilson 95% interval $[0.048, 0.162]$.

Matched-Estimand Baselines: final rejection probability / expected stopping time.

Δ	SMART	Bonferroni	Fixed-horizon DTE	Sieve
-0.025	$1.000/100.0$	$1.000/100.0$	$1.000/500.0$	$0.510/452.4$
-0.010	$1.000/100.2$	$1.000/100.0$	$1.000/500.0$	$0.154/485.8$
-0.005	$1.000/188.4$	$1.000/108.8$	$0.966/500.0$	$0.098/492.0$
-0.001	$0.340/472.4$	$0.272/439.0$	$0.174/500.0$	$0.050/495.8$
0.025	$0.000/500.0$	$0.000/500.0$	$0.000/500.0$	$0.000/500.0$

The original grid contains an interior null but not the boundary case $\delta = 0$, so it is not used as a precise size study. SMART and the new DTE procedures share the fixed synthetic dataset; Sieve uses the randomized-design stream required by its implementation and is not a paired same-data comparison.

C.2.1 Cross-Domain LQ Misspecification Benchmark

We introduce a non-medical energy-storage system whose state records storage level and demand and whose action is charging/discharging intensity. The correctly specified benchmark is linear with quadratic energy-imbalance and control costs. We then perturb one component at a time: nonlinear transitions, sinusoidally time-varying dynamics, or state-dependent finite-moment heavy-tailed noise, each at mild and strong intensity, producing seven environments in total. The policies and SMART algorithm are unchanged. Across expectation, quantile, and CVaR targets, empirical CI coverage is 0.85–1.00. Time variation is comparatively benign (maximum absolute DTE bias 0.11 and 0.07 at mild and strong levels), while state-dependent tail noise increases maximum bias to 0.32 and 0.43. Strong nonlinearity is the clearest failure mode, with maximum absolute bias 2.15. Thus, the experiment identifies a practical robustness range but does not extend the theory beyond LQ systems. The table reports ranges across the three estimands.

Environment	Null rejection	Power	Finite-CI coverage	Maximum |bias|
Correct LQ	0.10–0.11	0.17–0.23	0.91–0.96	0.14
Nonlinear, mild	0.05–0.09	0.08–0.15	0.85–0.98	0.20
Nonlinear, strong	0.06–0.13	0.18–0.28	0.88–0.98	2.15
Time-varying, mild	0.10–0.11	0.18–0.20	0.90–0.96	0.11
Time-varying, strong	0.08–0.13	0.16–0.20	0.89–0.97	0.07
Tail noise, mild	0.06–0.08	0.15–0.21	0.95–1.00	0.32
Tail noise, strong	0.05–0.11	0.17–0.18	0.93–1.00	0.43

C.2.2 OhioT1DM Model, Preprocessing, and Patient Audit

The legacy code pools transitions across all six patients but preserves patient boundaries. For uniform-within-bin mapping, held-out linear one-step RMSE is 9.18–15.03, compared with 10.62–19.60 for a random forest. Residual lag-one autocorrelation for glucose is 0.01; squared-residual correlations with state norm and absolute action are 0.081 and 0.042. The regressor covariance condition number is 1.65×10^6, and fitted closed-loop spectral radii are 0.76 and 0.73. Uniform and midpoint mappings yield null rejection ranges of 0.05–0.09 and 0.03–0.09, respectively; their power ranges are 0.23–0.42 and 0.48–0.69. Across six leave-one-patient-out folds, median null rejection is 0.05 (mean), 0.07 (quantile), and 0.065 (CVaR), while power ranges from 0.12–0.81, 0.16–0.72, and 0.11–0.61. With one patient, parameter variability increases sharply and CVaR null rejection reaches 0.15, confirming the finite-sample limitation. Taken together, these diagnostics support the pooled LQ system as an interpretable working approximation.

C.2.3 Sequential-Design Sensitivity

Using one null and one moderate alternative, we vary one design component at a time around the default $Q = 100$, $T_m = (100, 200, 300, 400, 500)$, early-conservative spending, and $\gamma = 0.98$. Here Q is the number of multiplier-bootstrap draws used to estimate standardized boundaries c_m. The

sparse schedule uses only $(250, 500)$, whereas *front-loaded* monitoring uses $(50, 100, 200, 350, 500)$ and therefore places more opportunities early without changing the horizon. *Linear spending* allocates cumulative Type-I error in proportion to information time; the *Pocock-like* schedule uses approximately equal boundaries across looks; the default early-conservative schedule spends little error initially and relaxes later.

Configuration	Null rejection	Power	Alternative $E[T]$
Default	0.09	0.32	474.0
$Q = 50$ / $Q = 200$	$0.07/0.09$	$0.34/0.36$	$475.0/470.0$
Sparse / front-loaded	$0.07/0.05$	$0.21/0.15$	$485.0/492.5$
Linear / Pocock-like	$0.19/0.20$	$0.51/0.57$	$412.0/394.0$
$\gamma = 0.95$ / 0.99	$0.08/0.09$	$0.21/0.46$	$480.0/448.0$

Changing Q from 50 to 200 changes power only from 0.34 to 0.36 around the default 0.32. Sparse and front-loaded monitoring reduce power to 0.21 and 0.15. Linear and Pocock-like schedules stop earlier, but their empirical null rejection rises to 0.19 and 0.20, so those apparent efficiency gains are not reliable in this finite-sample regime. Changing γ from 0.95 to 0.99 changes power from 0.21 to 0.46, reflecting a genuine change in the long-run return target rather than a boundary-tuning parameter. For a fixed spending rule, c_m is deterministic rather than user-tuned; seed variability enters through $c_m \widehat{\mathrm{SE}}_m$. The overall α is a prespecified experiment-wide error budget. Linear and Pocock-like schedules are reported, not selected post hoc, despite their inflated null rejection.