SAFE POLICY SELECTION FOR RESOURCE-CONSTRAINED TREATMENT ALLOCATION WITH BATCH-LEVEL UNDERPERFORMANCE CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Resource-constrained treatment allocation assigns a limited intervention jointly across each deployment batch. Once a baseline policy is in use, updated data and models often yield several candidate policies, making deployment a policy-selection problem. Mean gain alone is insufficient when a candidate can improve average outcomes yet incur a material loss on some deployment batches. In this paper, we study selection from a fixed candidate set under a constraint on the probability of batch-level underperformance relative to the baseline. This event is counterfactual and must be inferred from finite randomized calibration logs. We construct conditionally valid one-sided lower bounds for paired batch gains, convert them into conservative loss flags, and calibrate their frequencies with simultaneous exact binomial upper bounds. Selection then maximizes estimated mean gain among the certified candidates and the baseline. We prove finite-sample control of the selected policy's underperformance risk and characterize certification probability and the mean gain retained by finite calibration. Experiments with learned policies and controlled repeated-log and repeated-calibration designs illustrate the two sources of uncertainty and the resulting finite-sample behavior.

1 INTRODUCTION

Causal policy learning uses heterogeneous treatment responses to learn treatment rules with high expected welfare (Kitagawa & Tetenov, 2018; Athey & Wager, 2021). In many applications, however, shared budgets, quotas, or capacity limits couple the assignments within each deployment batch, so a policy must map the pretreatment information of the complete batch to a feasible joint allocation (Ai et al., 2022; Zhou et al., 2024). Once a baseline policy has been deployed, updated data and models often produce several candidate policies. The operational question then changes from learning a policy to choosing the next deployed policy from the baseline and a fixed candidate set, a setting closely related to off-policy selection from logged data (Kuzborskij et al., 2021; Thomas et al., 2015).

Expected policy value is central to this decision, but it does not describe how a candidate behaves across repeated deployment batches. A candidate can improve mean potential outcomes while still producing a material loss relative to the baseline on some deployment batches. Safe policy improvement addresses uncertainty in expected baseline-relative performance (Laroche et al., 2019; Ben-Michael et al., 2025), while other causal formulations study aggregate success probabilities, conditional utility, or individual counterfactual harm (Betlei et al., 2024; Jia et al., 2026; Jin & Egami, 2026). These criteria answer different deployment questions. For resource-constrained treatment allocation, we focus on the probability that a candidate policy's true average causal gain falls below a prespecified loss tolerance on a deployment batch. This batch-level underperformance risk is defined over repeated pretreatment batches and therefore measures variation in causal performance across deployments, excluding outcome noise realized after treatment.

Controlling this risk creates a statistical problem that is not present when the calibration loss is directly observed. The batch-level underperformance event is counterfactual because the candidate and baseline allocations cannot both be deployed on the same units. Randomized treatment logs identify their causal contrast under randomization and positivity, but a finite log still leaves uncer-

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.

tainty about whether the true batch gain crosses the loss threshold. A second source of uncertainty appears across calibration batches, where the frequency of this latent event must be bounded from finitely many observations. Risk-calibration frameworks provide finite-sample guarantees through valid risk bounds or tests (Bates et al., 2021; Angelopoulos et al., 2025). Applying them here requires constructing such evidence for a latent causal event from randomized outcomes. This motivates the central question of this paper: *Can finite randomized calibration logs support policy selection with a finite-sample guarantee on the selected policy's batch-level underperformance risk?*

To address this, we develop a finite-sample certification procedure that uses randomized calibration logs to evaluate each candidate against the deployed baseline. Within each batch, a one-sided causal lower bound yields a conservative indicator of material underperformance. Across batches, simultaneous calibration converts these indicators into risk certificates, after which selection favors the certified policy with the highest estimated mean gain. Our theory establishes finite-sample control after selection and characterizes how randomized evidence and calibration sample size determine certification probability and retained mean gain. Experimental results show that the procedure can preserve most of the mean gain achieved by mean-based selection while controlling batch-level underperformance risk.

Our main **contributions** are summarized as follows:

(i) We formulate policy selection for resource-constrained treatment allocation under a batch-level underperformance-risk constraint, with a fixed deployed baseline and a fixed set of candidate policies.

(ii) We combine conditionally valid one-sided causal evaluation within batches with simultaneous risk calibration across batches, and prove finite-sample control of the selected policy's batch-level underperformance risk. We also characterize certification probability and retained mean gain.

(iii) We evaluate the complete procedure with learned treatment policies and controlled experiments that separately examine within-batch causal evaluation and across-batch calibration uncertainty.

2 Related Work

Policy learning, constrained allocation, and baseline-relative safety. Empirical welfare maximization and doubly robust policy learning optimize expected population welfare over a policy class (Kitagawa & Tetenov, 2018; Athey & Wager, 2021). Resource-constrained methods extend this objective to joint assignments under shared budgets or capacity limits. LBCF combines heterogeneous treatment effect estimation with budget-constrained allocation (Ai et al., 2022), and Decision Focused Causal Learning trains causal models using the downstream allocation objective (Zhou et al., 2024). Off-policy policy selection instead begins with a fixed candidate set and uses logged data to compare expected policy value (Kuzborskij et al., 2021; Thomas et al., 2015). Safe policy improvement adds baseline-relative guarantees for expected performance (Laroche et al., 2019; Ben-Michael et al., 2025). Related causal work studies aggregate success probabilities, conditional utility relative to a baseline, and individual counterfactual harm (Betlei et al., 2024; Jia et al., 2026; Jin & Egami, 2026). Our setting combines a fixed candidate set with resource-constrained batch allocation and controls the probability of a material loss in true average causal gain across deployment batches.

Risk calibration for a counterfactual batch-level loss. Risk-controlling prediction sets use held-out data and one-sided confidence bounds to calibrate a predictor to a prescribed population risk (Bates et al., 2021). Learn then Test supports selection among multiple candidate configurations by combining valid risk tests with simultaneous error control (Angelopoulos et al., 2025). Applying this framework to batch-level causal loss requires valid tests for an event that is not directly observed. We first use randomized treatment logs to construct a conditionally valid lower bound for each candidate's batch gain and convert this bound into a conservative loss flag. We then calibrate the flag frequency across batches to obtain a finite-sample certificate for the latent batch-level underperformance risk.

Taken together, these literatures provide the ingredients for learning resource-constrained policies, comparing fixed policies with a deployed baseline, and calibrating population-level risk. They do

not directly yield a selection guarantee when the safety event itself is a counterfactual property of an entire deployment batch. This motivates evaluating the batch-level causal loss from randomized logs before calibrating its frequency across deployment batches, which is the problem studied here.

3 Problem Setup

We study policy selection among a deployed baseline and a fixed set of candidate policies for resource-constrained treatment allocation. Let $\mathcal{T}$ contain all data, fitted nuisance functions, candidate-construction choices, and randomization used before calibration. The baseline and candidate policies are fixed conditional on $\mathcal{T}$, while calibration batches, their randomized logs, and future deployment batches remain random. All probabilities and expectations below condition on $\mathcal{T}$ unless another conditioning set is displayed explicitly.

Batches, policies, and feasible allocations. Let $H \in \mathcal{H}$ denote the pretreatment information available for one deployment batch. It contains the batch size $N = N(H)$, covariates $X_1, \ldots, X_N$, and the resource constraints used at deployment. Treatments belong to the finite set $\mathcal{A} = \{0, \ldots, J\}$. Unit i has potential outcome $Y_i(a) \in [0, 1]$ under treatment a, with conditional mean $\mu_i(a, H) = \mathbb{E}[Y_i(a) \mid H, \mathcal{T}]$ (Rubin, 1974). A feasible allocation $z = (z_1, \ldots, z_N)$ lies in a known set $\mathcal{Z}(H) \subseteq \mathcal{A}^N$. For example, with treatment cost $b_i(a, H)$, total budget $B(H)$, and treatment-count bounds $l_a(H), u_a(H)$,

$$\sum_{i=1}^{N} b_i(z_i, H) \le B(H), \qquad l_a(H) \le \sum_{i=1}^{N} \mathbf{1}\{z_i = a\} \le u_a(H). \tag{1}$$

These restrictions couple assignments within a batch. A resource-constrained treatment policy π maps each H to an allocation satisfying $\pi(H) \in \mathcal{Z}(H)$. Conditional on $\mathcal{T}$, the baseline π_0 and candidates $\pi_1, \ldots, \pi_K$ are fixed. Write $a_i^k(H)$ for the treatment assigned to unit i by $\pi_k(H)$. Each policy may use the complete pretreatment batch information. Evaluation uses randomized treatments and outcomes generated after these policy outputs are fixed.

Causal gain and future-batch underperformance risk. For a future deployment batch H, candidate k has average causal gain relative to the baseline

$$\Delta_H(k) = \frac{1}{N(H)} \sum_{i=1}^{N(H)} \{\mu_i(a_i^k(H), H) - \mu_i(a_i^0(H), H)\}. \tag{2}$$

For a prespecified loss tolerance $\epsilon \ge 0$, define its mean gain and future-batch underperformance risk by

$$U(k) = \mathbb{E}[\Delta_H(k) \mid \mathcal{T}], \qquad R(k) = \mathbb{P}\{\Delta_H(k) < -\epsilon \mid \mathcal{T}\}. \tag{3}$$

Batches receive equal weight in both targets, and $\Delta_H(0) = U(0) = R(0) = 0$. For a risk limit $\delta \in (0, 1)$, the deployment goal is to select a policy with high mean gain while controlling

$$\mathbb{P}_{\text{cal}}\{R(\widehat{k}) > \delta \mid \mathcal{T}\} \le \alpha. \tag{4}$$

The selected policy is $\widehat{\pi} = \pi_{\widehat{k}}$. The tolerance ϵ defines a material batch-level loss. The risk limit δ bounds how often that loss may occur over future batches. The error level α controls the probability that calibration selects a policy whose true risk exceeds δ. Because the loss event uses conditional expected outcomes, $R(k)$ measures variation in causal performance across deployment batches and excludes outcome noise realized after treatment.

Randomized calibration logs. For calibration batch c, observe H_c and $\{(A_{ci}, Y_{ci})\}_{i=1}^{N_c}$, where $Y_{ci} = Y_{ci}(A_{ci})$. Randomized logs identify the causal contrast between each fixed candidate allocation and the baseline.

Assumption 1 (Treatment logs)**.** *Consistency and no interference hold. Conditional on $(H_c, \mathcal{T})$, the pairs (A_{ci}, Y_{ci}) are independent across i. Each A_{ci} is independent of the potential outcomes conditional on $(H_c, \mathcal{T})$, with known assignment probabilities $e_{ci}(a) = \mathbb{P}(A_{ci} = a \mid H_c, \mathcal{T})$. For every treatment used by the baseline or candidates, $e_{ci}(a) \ge \underline{e} > 0$.*

Randomization identifies the conditional treatment means and positivity supplies observations for every evaluated action (Rubin, 1974; Rosenbaum & Rubin, 1983). Known assignment probabilities also give unbiased augmented scores for fixed candidates when the outcome predictions are misspecified (Bang & Robins, 2005; Dudík et al., 2011). Conditional independence across units is used to construct the finite-sample lower bound in Appendix B.

Assumption 2 (Calibration sampling)**.** *Conditional on $\mathcal{T}$, the C calibration batches, including their randomized logs, are independent and identically distributed. A future batch and its hypothetical evaluation log have the same law and are independent of the calibration data.*

Batch sizes, covariates, and resource constraints may vary across draws from this common law. All candidates may share the same calibration logs, and dependence across candidates is allowed. The target $R(k)$ depends on the latent causal gain $\Delta_H(k)$, which is unavailable from a finite randomized log without estimation. Certification must therefore address uncertainty within each batch before estimating the frequency of harmful batches.

4 Risk Certification and Policy Selection

The selection procedure addresses the two sources of uncertainty in sequence. First, it evaluates each candidate policy against the baseline within a calibration batch and constructs a one-sided causal lower bound. Second, it calibrates the frequency of conservative loss flags across batches and selects by estimated mean gain among the certified candidates and the baseline. Complete proofs are given in Appendix A.

4.1 Paired causal evaluation

Let $m_{ci}(a) \in [0,1]$ be an outcome prediction trained using $\mathcal{T}$. For each evaluated action, define the augmented inverse propensity weighted (AIPW) score (Bang & Robins, 2005; Dudík et al., 2011)

$$\widehat{\psi}_{ci}(a) = m_{ci}(a) + \frac{\mathbf{1}\{A_{ci} = a\}}{e_{ci}(a)}\{Y_{ci} - m_{ci}(a)\}. \tag{5}$$

For candidate k, the paired contribution and estimated batch gain are

$$D_{ci}(k) = \widehat{\psi}_{ci}(a^k_{ci}) - \widehat{\psi}_{ci}(a^0_{ci}), \qquad \widehat{\Delta}_c(k) = \frac{1}{N_c}\sum_{i=1}^{N_c} D_{ci}(k). \tag{6}$$

The two allocations are evaluated on the same randomized observation. Hence $D_{ci}(k) = 0$ whenever candidate and baseline assign the same treatment, and only their disagreements contribute sampling noise. The normalization by the full batch size matches the target $\Delta_H(k)$ in (2).

Proposition 1 (Conditional unbiasedness)**.** *Under Assumption 1, $\mathbb{E}[\widehat{\Delta}_c(k) \mid H_c, \mathcal{T}] = \Delta_{H_c}(k)$ for every fixed candidate and every training-only choice of $m_{ci}(a) \in [0,1]$.*

Proof is given in Appendix A.1. With known randomized assignment probabilities, the paired AIPW gain is conditionally unbiased for every fixed candidate even when the outcome predictions are misspecified.

A point estimate of $\Delta_{H_c}(k)$ is insufficient for calibrating the event $\Delta_{H_c}(k) < -\epsilon$, because finite-log error can move a harmful batch across the threshold. Let $L_c(k) = \widehat{\Delta}_c(k) - r_c(k)$ be a one-sided lower bound with nonnegative radius $r_c(k)$, and set $L_c(0) = 0$. We require coverage conditional on the deployment batch.

Assumption 3 (Conditional lower bound)**.** *Conditional on $\mathcal{T}$, the rule computing $L_c(k)$ from H_c and its own randomized log is fixed across batches for each k. For a fixed $\eta \in (0,1)$ and every $k \in \{1, \dots, K\}$,*

$$\mathbb{P}\{L_c(k) \le \Delta_{H_c}(k) \mid H_c, \mathcal{T}\} \ge 1 - \eta \quad \textit{for almost every } H_c. \tag{7}$$

For a finite-action construction, let $h_{ci}(k)$ be the range width of $D_{ci}(k)$ over logged actions of positive probability and outcome endpoints $0, 1$, conditional on $H_c, \mathcal{T}$. This width is determined by

the fixed predictions, logging design, and chosen actions, and is zero where the allocations agree. Conditional independence and Hoeffding's inequality (Hoeffding, 1963) give the radius

$$r_c^{\mathrm{H}}(k) = \frac{1}{N_c}\sqrt{\frac{\log(1/\eta)}{2}\sum_{i=1}^{N_c} h_{ci}(k)^2},$$

for which $\widehat{\Delta}_c(k) - r_c^{\mathrm{H}}(k)$ satisfies Assumption 3. Appendices B.1 and B.2 give the range calculation and proof. The experiments use the binary-treatment construction in (34) of Appendix B.3. It combines deterministic support bounds with observed squared residual terms to obtain candidate-specific radii. Proposition 8 establishes its conditional coverage with estimated variance terms.

4.2 From batch-level lower bounds to risk certification

A calibration batch is flagged when its lower bound falls below the tolerated-loss threshold. Define

$$F_c(k) = \mathbf{1}\{L_c(k) < -\epsilon\}, \qquad q(k) = \mathbb{P}\{F_c(k) = 1 \mid \mathcal{T}\}. \tag{8}$$

The flag probability $q(k)$ averages over both the deployment batch and the randomized evaluation log. The target $R(k)$ averages only the event defined by the true batch gain. Conditional coverage links these two probabilities.

Theorem 2 (Risk bound from conditional coverage)**.** *Under Assumption 3, every candidate satisfies*

$$(1-\eta)R(k) \le q(k), \qquad R(k) \le \min\left\{1, \frac{q(k)}{1-\eta}\right\}. \tag{9}$$

For $0 \le \epsilon < 1$, the coefficient $1-\eta$ is sharp among lower bounds satisfying (7)*.*

Proof is given in Appendix A.2. The result follows from the harmful batches themselves. Whenever $\Delta_H(k) < -\epsilon$, coverage of the lower bound implies $L_H(k) < -\epsilon$, so a harmful batch is flagged with conditional probability at least $1-\eta$. Averaging this statement over future batches gives (9). Conditional coverage in Assumption 3 is essential for this multiplicative relation.

For candidate k, let $S_k = \sum_{c=1}^{C} F_c(k)$ be the number of calibration flags. Assumption 2 gives $S_k \sim \mathrm{Binomial}(C, q(k))$ conditional on $\mathcal{T}$. Write $Q_p(a,b)$ for the p quantile of a beta distribution with shape parameters a, b. The one-sided Clopper–Pearson upper bound at level $\beta = \alpha/K$ is (Clopper & Pearson, 1934)

$$u_{\mathrm{CP}}(s, C, \beta) = \begin{cases} Q_{1-\beta}(s+1, C-s), & s < C, \\ 1, & s = C. \end{cases} \tag{10}$$

We certify candidates whose upper bound on $q(k)$ is small enough to imply $R(k) \le \delta$ through Theorem 2. The certified set and mean-gain selection rule are

$$\widehat{\mathcal{S}} = \{k \in \{1, \dots, K\} : u_{\mathrm{CP}}(S_k, C, \alpha/K) \le (1-\eta)\delta\}, \tag{11}$$

$$\widehat{k} \in \operatorname*{arg\,max}_{k \in \widehat{\mathcal{S}} \cup \{0\}} \widehat{U}(k), \quad \widehat{U}(k) = \frac{1}{C}\sum_{c=1}^{C} \widehat{\Delta}_c(k), \quad \widehat{U}(0) = 0. \tag{12}$$

Ties favor the baseline and then a fixed candidate order. A nonbaseline selection therefore requires both a risk certificate and positive estimated gain. The simultaneous construction permits the same calibration batches to be used for certification and subsequent mean-gain selection (Angelopoulos et al., 2025).

Theorem 3 (Finite-sample safety after selection)**.** *Under Assumptions 2 and 3, with conditional probability at least $1-\alpha$, every $k \in \widehat{\mathcal{S}}$ has $R(k) \le \delta$. Consequently,* (4) *holds for* (12)*. The selected policy produces an allocation in $\mathcal{Z}(H)$ for every deployment batch.*

Proof is given in Appendix A.4. Exact binomial coverage controls each $q(k)$, and the α/K allocation yields a simultaneous event over the K candidates. Combining that event with (9) gives the risk statement after selection. Feasibility is inherited from $\pi_k(H) \in \mathcal{Z}(H)$. Algorithm 1 summarizes the procedure.

Algorithm 1 Future-batch risk certification and mean-gain policy selection

Require: Training-fixed policies $\pi_0, \ldots, \pi_K$, outcome predictions, calibration logs, $\epsilon, \delta, \eta, \alpha$.

1: **for** each calibration batch c **do**
2: Compute feasible allocations $\pi_k(H_c)$ before using its logged treatments or outcomes.
3: Compute paired gains $\widehat{\Delta}_c(k)$ and conditional lower bounds $L_c(k)$.
4: Record $F_c(k) = \mathbf{1}\{L_c(k) < -\epsilon\}$ for each candidate.
5: Form $\widehat{\mathcal{S}}$ using (10) and (11), then select $\widehat{k}$ by (12).
6: On a future batch H, deploy $\pi_{\widehat{k}}(H)$.

4.3 Certification probability and retained mean gain

Certification probability depends on the flag probability of a fixed candidate. Define the largest admissible flag count

$$b_C = \max\{s \in \{0, \ldots, C\} : u_{\mathrm{CP}}(s, C, \alpha/K) \leq (1-\eta)\delta\}, \tag{13}$$

with $b_C = -1$ when the set is empty. Let $X_{C,q} \sim \mathrm{Binomial}(C, q)$. Then

$$\mathbb{P}\{k \in \widehat{\mathcal{S}} \mid \mathcal{T}\} = \mathbb{P}\{X_{C,q(k)} \leq b_C\}. \tag{14}$$

Equation (14) separates the two sample sizes. Within-batch sample size affects the lower bound and therefore $q(k)$. The number of calibration batches C determines how precisely that flag probability can be certified. Appendix A.5 gives the zero-flag sample requirement.

Define the best candidate satisfying the population risk constraint by

$$U_\delta^\star = \max_{k \in \{0,\ldots,K\}: R(k) \leq \delta} U(k). \tag{15}$$

Finite calibration can certify only part of this safe set. When $b_C \geq 0$, let q_ρ satisfy $\mathbb{P}\{X_{C,q_\rho} \leq b_C\} = 1 - \rho$ for $\rho \in (0, 1-\alpha)$, and define

$$U_\rho^{\mathrm{cert}} = \max_{k \in \{0\} \cup \{j : q(j) \leq q_\rho\}} U(k).$$

This is the largest mean gain among candidates whose certification probability is at least $1 - \rho$. Theorem 5 in Appendix A.6 shows that, for $\zeta \in (0, 1-\alpha-\rho)$ and

$$a_C = \underline{e}^{-1}\sqrt{2\log(2K/\zeta)/C},$$

the selected policy satisfies $R(\widehat{k}) \leq \delta$ and $U(\widehat{k}) \geq U_\rho^{\mathrm{cert}} - 2a_C$ with probability at least $1-\alpha-\rho-\zeta$. Thus finite calibration can reduce mean gain through estimation error and through safe candidates that are not certified with high probability.

5 Experiments

We first evaluate the complete selection procedure with learned resource-constrained policies. We then isolate within-batch causal evaluation and across-batch calibration in controlled experiments.

Datasets and protocol. The semi-synthetic environment uses 12 Criteo covariates (Diemert et al., 2021). Untreated response probabilities come from an outcome model fitted to source control observations. The fully synthetic environment generates all covariates and specifies untreated response probabilities directly. In both environments, a generated batch-level context W, observed by the learner, changes the covariate distribution and enters the response probabilities. The signal parameter $\gamma \in \{1, 2\}$ scales heterogeneous log-odds shifts. Conditional on each batch, treatment and outcome pairs are independent across individuals, with known treatment probability $e = 0.5$ and binary outcomes. Every policy treats exactly 20% of a batch, with at least 6% of the full population treated within a synthetic quota group. Batch sizes are $N = 500{,}000$ for Criteo and $N = 2{,}000{,}000$ for Synthetic. At each γ, 20 independent repetitions refit the baseline, candidates, and evaluation models from separate histories, followed by $C = 160$ calibration batches and $T = 256$ independent future batches. Known response probabilities are used only for post-selection evaluation. Each experimental campaign is one realization of the batch H in the theory. Appendices C and F give the full protocols.

Table 1: End-to-end policy selection with the same 18-candidate family in both environments. Gains are mean ± sample standard deviation over 20 independent repetitions per setting, including fallback zeros. Retention is the ratio of average gains to the same-setting MeanOnly. Selected is the number of repetitions in which a candidate policy is chosen instead of Production. Failures pool 256 future batches per fitted policy and are descriptive. Production's zero is exact.

Signal	Method	Gain ↑ (pp)	Retained ↑	Selected	Failures
Criteo, $N = 500{,}000$					
$\gamma = 1$	Production	0 (exact)	0.00%	0/20	0/5,120
	MeanOnly	0.05625 ± 0.01624	100.00%	20/20	0/5,120
	NoisyCP	0.05625 ± 0.01624	100.00%	20/20	0/5,120
	Ours	0.05139 ± 0.01778	91.37%	20/20	0/5,120
$\gamma = 2$	Production	0 (exact)	0.00%	0/20	0/5,120
	MeanOnly	0.03057 ± 0.00936	100.00%	20/20	0/5,120
	NoisyCP	0.03057 ± 0.00936	100.00%	20/20	0/5,120
	Ours	0.03022 ± 0.00920	98.88%	20/20	0/5,120
Synthetic, $N = 2{,}000{,}000$					
$\gamma = 1$	Production	0 (exact)	0.00%	0/20	0/5,120
	MeanOnly	0.07989 ± 0.02560	100.00%	20/20	0/5,120
	NoisyCP	0.07989 ± 0.02560	100.00%	20/20	0/5,120
	Ours	0.07874 ± 0.02664	98.55%	20/20	0/5,120
$\gamma = 2$	Production	0 (exact)	0.00%	0/20	0/5,120
	MeanOnly	0.07025 ± 0.01216	100.00%	20/20	0/5,120
	NoisyCP	0.07025 ± 0.01216	100.00%	20/20	0/5,120
	Ours	0.07025 ± 0.01216	100.00%	20/20	0/5,120

Methods and metrics. Production is a depth-three XGBoost T-learner trained on 256,000 records (Chen & Guestrin, 2016; Künzel et al., 2019). The main end-to-end comparison uses the same $K = 18$ candidate family in both environments. It contains 12 T-learners and doubly robust (DR) learners (Kennedy, 2023) across three training sizes and two depths, plus six baseline-DR score blends. MeanOnly selects by estimated mean gain. NoisyCP replaces the causal lower-bound flag with the point-estimate flag and retains the same binomial calibration step. Our procedure selects by estimated mean gain among certified candidates and Production. We report future mean causal gain, retention relative to MeanOnly, selection counts, and violations of the $-\epsilon = -0.05$ percentage-point (pp) threshold. Future-risk intervals are simultaneous within each prespecified evaluation family. The calibration parameters are $\delta = \alpha = 0.05$ and $\eta = 0.01$. Appendix C.2 defines the comparison rules.

Main results. Table 1 shows that our procedure selects a positive-gain candidate in all 40 repetitions per environment. It retains 91.37% and 98.88% of MeanOnly's average gain on Criteo, and 98.55% and 100% on Synthetic. For each environment, the $20 \times 2 \times 18 = 720$ fitted candidate instances form the prespecified future-risk evaluation family. Every selected candidate has a simultaneous 95% future-risk upper bound of 3.93%, below the 5% target. MeanOnly and NoisyCP also have zero observed future violations, so the end-to-end comparison quantifies the mean gain retained after certification. The paired Criteo results in Appendix D show that certification changes the selected candidate in 9 of 20 repetitions for $\gamma = 1$ and 4 of 20 for $\gamma = 2$, consistent with the larger retention loss in the weaker-signal setting.

Candidate screening. Figure 1 compares Criteo certification decisions with independent future performance. Among 720 fitted candidates, simultaneous future intervals lie above the 5% risk limit for 279 candidates, below it for 426, and overlap it for 15. None of the 279 candidates whose future interval lies above 5% is certified. All 228 certified candidates have positive future mean gain and zero observed violations. Certification also excludes 198 candidates whose future intervals lie below 5%, including 124 with positive future mean gain. These candidates have calibration flag frequencies that are too high for certification with $C = 160$. In this benchmark, every candidate whose future interval lies above 5% also has negative mean gain, so MeanOnly does not select them.

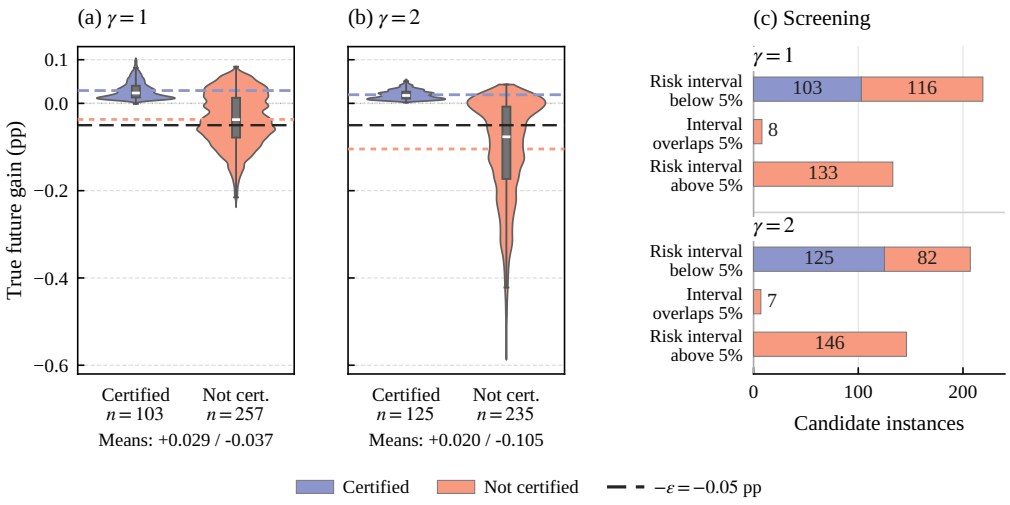

Figure 1: Candidate screening on Criteo. Panels (a) and (b) pool future gains for certified and non-certified fitted candidates at the two signal settings, with n denoting the number of fitted candidates. Inner boxes show interquartile ranges and medians. Colored dashed lines mark means, listed below each panel in the same group order. The black dashed line is the loss threshold $-\epsilon = -0.05$ pp. Panel (c) reports candidate counts by simultaneous future-risk interval classification and certification status.

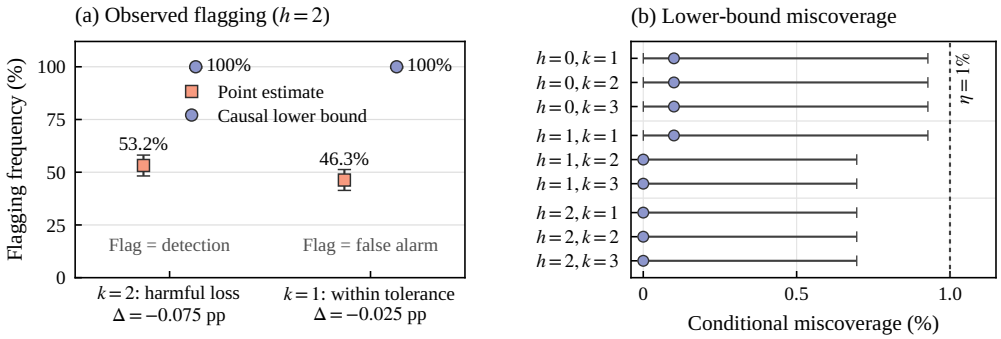

Figure 2: Fixed-mechanism logging with 1,000 independent logs per batch type. Panel (a) compares point-estimate and causal-lower-bound flagging for a harmful gain and a nearby within-tolerance gain. Panel (b) reports conditional lower-bound miscoverage for the nine batch-type and candidate conditions. Error bars are simultaneous 95% Clopper–Pearson intervals over the 27 proportions defined in Appendix E.2.

Within-batch causal evaluation. To isolate logging uncertainty, we fix three batch types with $N = 50{,}000$, a Production policy, and three feasible candidate policies. Each policy treats one of five equally sized groups under the same capacity and quota constraints. The response probabilities give exact candidate gains. We draw 1,000 independent randomized logs per batch type with $e = 0.5$, keeping the policies and response mechanism fixed. At the harmful gain -0.075 pp, the point-estimate and lower-bound flag rates are 53.2% and 100%. At the within-tolerance gain -0.025 pp, the corresponding rates are 46.3% and 100%. Missed harmful batches weaken risk calibration, while conservative flags on safe batches reduce certification probability. Across all nine batch-type and candidate conditions, lower-bound miscoverage occurs zero or once, with simultaneous upper endpoints below $\eta = 1\%$. Figure 2 is consistent with the conditional coverage guarantee and illustrates the conservatism that can reduce certification probability. Appendix E.2 gives all nine conditions.

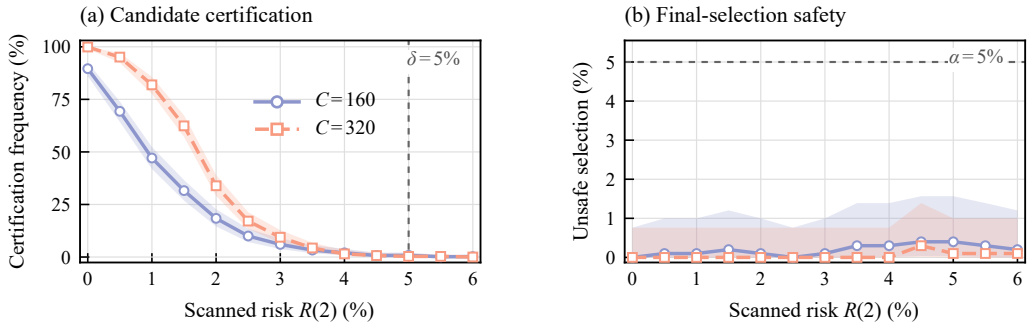

Figure 3: Risk scan over $r = R(2)$. Panel (a) reports candidate-two certification against the risk limit $\delta = 5\%$. Panel (b) reports the frequency of selecting a policy with risk above δ against the error limit $\alpha = 5\%$. Each point uses 1,000 independent calibration sequences with simultaneous 95% Clopper–Pearson intervals.

Across-batch calibration. Next, we vary how often the three batch types occur while keeping the policies, response probabilities, and $N = 50{,}000$ fixed. Candidate two exceeds the loss tolerance only for batch type $h = 2$, whose probability is $r \in \{0, 0.5, \ldots, 6\}\%$. Its exact risk is $R(2) = r$, while the other candidate risks remain 0 and 6%. Each setting uses 1,000 independent sequences of 320 batches, with a batch type and fresh randomized log drawn for each batch. Certification and selection use either the first 160 batches or all 320 batches. As shown in Figure 3, doubling C raises candidate-two certification from 89.6% to 99.9% at $r = 0$ and from 47.1% to 81.9% at $r = 1\%$, while certification becomes rare near the 5% risk limit. This behavior is consistent with (14). A larger C sharpens inference on the flag probability and increases certification probability for candidates sufficiently below the boundary. Panel (a) reports candidate-level certification. Panel (b) evaluates the target of Theorem 3. Across the scanned settings, unsafe selections occur in 0 to 4 of 1,000 sequences, and all simultaneous upper endpoints remain below $\alpha = 5\%$. Appendix E.3 reports the complete scan.

6 Conclusion

In this paper, we study policy selection for resource-constrained treatment allocation when deployment must control the probability of a material batch-level loss relative to a fixed baseline. The loss event is counterfactual and must be inferred from finite randomized logs. Our procedure combines conditionally valid lower bounds for batch causal gains with simultaneous binomial calibration across batches, yielding finite-sample risk control after selection from a fixed candidate set. We also characterize certification probability and the mean gain retained by finite calibration. Experiments with learned policies and controlled repeated-log and repeated-calibration designs illustrate the predicted finite-sample behavior and show that the main candidate family can retain most of the gain achieved by mean-based selection. These results connect randomized causal evaluation with finite-sample policy selection under shared resource constraints.

Limitations and future work. The analysis assumes that the candidate set is fixed before calibration. Extending the guarantee to candidates updated during certification requires valid calibration under data-dependent candidate generation. The theory also assumes that calibration batches and future deployment batches follow the same distribution, leaving certification under distribution shift for future work. Tighter conditionally valid lower bounds for the batch gain may improve certification probability while keeping the deployment target and selection rule fixed.

AI use statement

Generative AI tools were used for language polishing and manuscript editing. The authors verified the resulting proofs, code, numerical results, and the final manuscript.

REFERENCES

Meng Ai, Biao Li, Heyang Gong, Qingwei Yu, Shengjie Xue, Yuan Zhang, Yunzhou Zhang, and Peng Jiang. LBCF: A large-scale budget-constrained causal forest algorithm. In *Proceedings of the ACM Web Conference 2022*, 2022. doi: 10.1145/3485447.3512103.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021. doi: 10.3982/ECTA15732.

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005. doi: 10.1111/j.1541-0420.2005.00377.x.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):43:1–43:34, 2021. doi: 10.1145/3478535.

Eli Ben-Michael, D James Greiner, Kosuke Imai, and Zhichao Jiang. Safe policy learning through extrapolation: Application to pre-trial risk assessment. *Journal of the American Statistical Association*, 120(551):1386–1399, 2025.

Artem Betlei, Mariia Vladimirova, Mehdi Sebbar, Nicolas Urien, Thibaud Rahier, and Benjamin Heymann. Maximizing the success probability of policy allocations in online systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10):11061–11068, 2024. doi: 10.1609/aaai.v38i10.28982.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Eustache Diemert, Artem Betlei, Christophe Renaudin, Massih-Reza Amini, Théophane Gregoir, and Thibaud Rahier. A large scale benchmark for individual treatment effect prediction and uplift modeling. *arXiv preprint arXiv:2111.10106*, 2021. URL https://arxiv.org/abs/2111.10106.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1097–1104, 2011. URL https://arxiv.org/abs/1103.4601.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Zeyang Jia, Eli Ben-Michael, and Kosuke Imai. Bayesian safe policy learning with chance constrained optimization: Application to military security assessment during the Vietnam War. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 189(3):1448–1472, 2026. doi: 10.1093/jrsssa/qnaf122.

Ying Jin and Naoki Egami. Conformal policy learning with distribution-free safety guarantees. *arXiv preprint arXiv:2609.17296*, 2026. URL https://arxiv.org/abs/2609.17296.

Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023. doi: 10.1214/23-EJS2157.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018. doi: 10.3982/ECTA13288.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. doi: 10.1073/pnas.1804597116.

Ilja Kuzborskij, Claire Vernade, Andras Gyorgy, and Csaba Szepesvari. Confident off-policy evaluation and selection through self-normalized importance weighting. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 640–648, 2021. URL https://proceedings.mlr.press/v130/kuzborskij21a.html.

Romain Laroche, Paul Trichelair, and Rémi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3652–3661, 2019. URL https://proceedings.mlr.press/v97/laroche19a.html.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), 2015. doi: 10.1609/aaai.v29i1.9541.

Hao Zhou, Rongxiao Huang, Shaoming Li, Guibin Jiang, Jiaqi Zheng, Bing Cheng, and Wei Lin. Decision focused causal learning for direct counterfactual marketing optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6368–6379, 2024. doi: 10.1145/3637528.3672353.

A Proofs of the Main Results

Throughout this appendix, condition on the complete training state $\mathcal{T}$. Candidate policies, nuisance predictions, numerical tolerances, and tie-breaking rules are then fixed. Randomness in a calibration observation consists of the batch information H_c and the logged treatment assignments and outcomes conditional on that information. The conditional statements below can be integrated over $\mathcal{T}$ when their assumptions hold for almost every training state.

A.1 Conditional evaluation of complete allocations

Proof of Proposition 1. Fix a batch H_c and an evaluated action a. The prediction $m_{ci}(a)$ and assignment probability $e_{ci}(a)$ are fixed under this conditioning. By consistency and randomized assignment,

$$\mathbb{E}[\widehat{\psi}_{ci}(a) \mid H_c, \mathcal{T}] = m_{ci}(a) + \frac{1}{e_{ci}(a)}\mathbb{E}[\mathbf{1}\{A_{ci} = a\}\{Y_{ci} - m_{ci}(a)\} \mid H_c, \mathcal{T}] \tag{16}$$

$$= m_{ci}(a) + \mathbb{E}[Y_{ci}(a) - m_{ci}(a) \mid H_c, \mathcal{T}] \tag{17}$$

$$= \mu_i(a, H_c). \tag{18}$$

Every selected action a^k_{ci} is fixed conditional on $H_c, \mathcal{T}$. Applying this identity to the candidate and baseline actions and subtracting gives

$$\mathbb{E}[D_{ci}(k) \mid H_c, \mathcal{T}] = \mu_i(a^k_{ci}, H_c) - \mu_i(a^0_{ci}, H_c).$$

Linearity of expectation yields the conclusion after summing and dividing by N_c. The expectation calculation uses randomization and consistency. Conditional independence across individuals enters the concentration construction in Appendix B. □

Conditional on the complete batch information, the resource-constrained optimizer returns a fixed action vector. Proposition 1 therefore applies to the complete allocation whenever the policy is fixed before its evaluation log is observed.

A.2 The connection between flags and underperformance

Proof of Theorem 2. Fix candidate k and define $A_k(H) = \mathbf{1}\{\Delta_H(k) < -\epsilon\}$. This event depends on the deployment population, including its constraints, and not on the additional logging randomness used for evaluation. If $A_k(H) = 1$ and $L_H(k) \le \Delta_H(k)$, then $L_H(k) < -\epsilon$. Therefore

$$\mathbb{P}\{F_H(k) = 1 \mid H, \mathcal{T}\} \ge A_k(H)\mathbb{P}\{L_H(k) \le \Delta_H(k) \mid H, \mathcal{T}\} \ge (1 - \eta)A_k(H).$$

Taking expectations over H gives $q(k) \ge (1 - \eta)R(k)$. Dividing by $1 - \eta > 0$ and using $R(k) \le 1$ proves (9).

For sharpness, consider $0 \le \epsilon < 1$ and choose $d \in (\epsilon, 1]$. Let $\Delta_H = -d$ on a subset of batches with probability $r \in [0, 1]$, and let $\Delta_H = 0$ elsewhere. On a harmful batch, set $L_H = \Delta_H$ with probability $1 - \eta$ and $L_H = 0$ with probability η. On any other batch, set $L_H = 0$. The lower bound has conditional coverage at least $1 - \eta$ for every batch, and its flag probability is exactly $(1 - \eta)r$. Thus a larger universal coefficient multiplying R is impossible under the stated coverage requirement alone. This construction establishes sharpness of the statistical implication under conditional coverage alone. □

A.3 Binomial upper confidence bounds

For completeness, we prove the coverage property used by the algorithm. Let $S \sim \mathrm{Binomial}(C, q)$, where $q \in [0, 1]$, and fix $\beta \in (0, 1)$. Define $u(s) = u_{\mathrm{CP}}(s, C, \beta)$ as in (10). For $s < C$, the beta-binomial identity gives

$$\mathbb{P}\{X_{C,u(s)} \le s\} = \beta. \tag{19}$$

It follows by differentiating the binomial cumulative distribution, or directly by integrating the beta density. The left side is continuous and strictly decreasing in its probability argument on $(0, 1)$.

Lemma 4 (One-sided exact coverage)**.** *For every $q \in [0, 1]$, $\mathbb{P}_q\{q > u(S)\} \le \beta$.*

Proof. The upper bound $u(s)$ is increasing in s. If no integer s satisfies $u(s) < q$, the failure probability is zero. Otherwise let s_q be the largest such integer. Since $u(C) = 1$, necessarily $s_q < C$. The failure event is $\{S \le s_q\}$, and (19) and monotonicity imply

$$\mathbb{P}_q\{q > u(S)\} = \mathbb{P}\{X_{C,q} \le s_q\} \le \mathbb{P}\{X_{C,u(s_q)} \le s_q\} = \beta.$$

At $q = 0$ the event is empty. At $q = 1$, $S = C$ almost surely and $u(C) = 1$, so the event is also empty. □

This is the finite-sample binomial construction of Clopper & Pearson (1934). Its coverage can exceed the nominal level because the count is discrete.

A.4 Safety of the certified set and the selected policy

Proof of Theorem 3. The candidate-specific flags are functions of independent and identically distributed calibration batches and their logs. Thus, for each fixed k, $S_k \sim \mathrm{Binomial}(C, q(k))$. Lemma 4 at $\beta = \alpha/K$ and a union bound give

$$\mathbb{P}\left\{q(k) \le u_{\mathrm{CP}}(S_k, C, \alpha/K) \text{ for every } k = 1, \dots, K \mid \mathcal{T}\right\} \ge 1 - \alpha.$$

Call this event E. The union bound applies directly even when the candidate counts are dependent. On E, any certified k satisfies

$$R(k) \le \frac{q(k)}{1-\eta} \le \frac{u_{\mathrm{CP}}(S_k, C, \alpha/K)}{1-\eta} \le \delta.$$

If the selection is a certified candidate, it is consequently safe on E. If the selection is the baseline, it has risk zero by definition. Therefore

$$\mathbb{P}\{R(\widehat{k}) > \delta \mid \mathcal{T}\} \le \mathbb{P}(E^c \mid \mathcal{T}) \le \alpha.$$

Every candidate and the baseline map every H into $\mathcal{Z}(H)$, so the selected map does too. □

Selection by estimated mean gain may reuse the calibration data because the simultaneous event covers every candidate in the fixed family.

A.5 Certification probability and sample-size requirements

Monotonicity of u_{CP} implies $k \in \widehat{\mathcal{S}}$ if and only if $S_k \le b_C$. Consequently,

$$\mathbb{P}(k \in \widehat{\mathcal{S}} \mid \mathcal{T}) = \sum_{s=0}^{b_C} \binom{C}{s} q(k)^s \{1 - q(k)\}^{C-s},$$

where the sum is zero when $b_C = -1$. This proves (14). For $s = 0$, (19) becomes $(1-u)^C = \alpha/K$, so

$$u_{\mathrm{CP}}(0, C, \alpha/K) = 1 - (\alpha/K)^{1/C}.$$

Certification with zero flags is possible exactly when this is at most $(1-\eta)\delta$. Taking logarithms, with both logarithms negative, gives

$$C \ge \left\lceil \frac{\log(\alpha/K)}{\log\{1 - (1-\eta)\delta\}} \right\rceil. \tag{20}$$

Assume $b_C \ge 0$. Since $(1-\eta)\delta < 1$, $b_C < C$. The binomial cumulative distribution at b_C is continuous and strictly decreasing from one to zero as q increases from zero to one. Thus q_ρ exists and is unique for $\rho \in (0, 1)$. If $\rho < 1 - \alpha/K$, then $q_\rho < (1-\eta)\delta$. To see this, the definition of b_C and (19) give

$$\mathbb{P}\{X_{C,(1-\eta)\delta} \le b_C\} \le \alpha/K < 1 - \rho.$$

Monotonicity yields the claim. The candidates in (21) are therefore safe under Theorem 2 for this range of ρ.

A.6 MEAN-GAIN COMPARISON

For $\rho \in (0, 1-\alpha)$ and $b_C \geq 0$, let q_ρ be the unique probability satisfying $\mathbb{P}\{X_{C,q_\rho} \leq b_C\} = 1-\rho$, and define

$$U_\rho^{\mathrm{cert}} = \max_{k \in \{0\} \cup \{j : q(j) \leq q_\rho\}} U(k). \tag{21}$$

This is the largest mean gain among the baseline and candidates whose certification probability is at least $1-\rho$.

Theorem 5 (Mean-gain comparison for certifiable candidates)**.** *Suppose Assumptions 1, 2, and 3 hold and $b_C \geq 0$. For $\zeta \in (0, 1-\alpha-\rho)$, set*

$$a_C = \underline{e}^{-1} \sqrt{\frac{2\log(2K/\zeta)}{C}}. \tag{22}$$

With probability at least $1-\alpha-\rho-\zeta$, the selected candidate satisfies

$$R(\widehat{k}) \leq \delta, \qquad U(\widehat{k}) \geq U_\rho^{\mathrm{cert}} - 2a_C. \tag{23}$$

The proof follows Lemma 6 below. The theorem compares mean gain within the portion of the fixed candidate family that is certified with probability at least $1-\rho$. Relative to (15), the corresponding gap is $U_\delta^\star - U_\rho^{\mathrm{cert}} + 2a_C$.

Lemma 6 (Uniform mean-gain estimation)**.** *Under Assumptions 1 and 2, with a_C from* (22)*,*

$$\mathbb{P}\{\max_{1 \leq k \leq K} |\widehat{U}(k) - U(k)| > a_C \mid \mathcal{T}\} \leq \zeta.$$

Proof. If the two selected actions agree, $D_{ci}(k) = 0$. Otherwise write $a = a_{ci}^k$ and $b = a_{ci}^0$. When $A_{ci} = a$, the paired score is

$$D_{ci}(k) = \{1 - e_{ci}(a)^{-1}\} m_{ci}(a) - m_{ci}(b) + Y_{ci}/e_{ci}(a).$$

Since predictions and outcomes lie in $[0,1]$, this expression belongs to $[-e_{ci}(a)^{-1}, e_{ci}(a)^{-1}]$. When $A_{ci} = b$, the same argument gives the interval $[-e_{ci}(b)^{-1}, e_{ci}(b)^{-1}]$. For any other logging action, the score is the difference of two predictions and belongs to $[-1,1]$. Consequently,

$$|D_{ci}(k)| \leq M, \quad |\widehat{\Delta}_c(k)| \leq M, \qquad M = \underline{e}^{-1}.$$

The batch estimates are independent across c, and their expectations equal $U(k)$ by Proposition 1 and iterated expectation. Hoeffding's inequality (Hoeffding, 1963) yields

$$\mathbb{P}\{|\widehat{U}(k) - U(k)| > a \mid \mathcal{T}\} \leq 2\exp\{-Ca^2/(2M^2)\}.$$

Substituting $a = a_C$ and summing over K candidates proves the result. The baseline is exact and adds no estimation error. □

Proof of Theorem 5. Choose a deterministic maximizer k_ρ in (21), conditional on $\mathcal{T}$, with ties broken by the fixed candidate order. If $k_\rho = 0$, it is always available. Otherwise $q(k_\rho) \leq q_\rho$, so (14) gives

$$\mathbb{P}\{k_\rho \in \widehat{\mathcal{S}} \mid \mathcal{T}\} \geq 1-\rho.$$

On this event and the uniform mean-estimation event of Lemma 6, the selection rule implies

$$U(\widehat{k}) \geq \widehat{U}(\widehat{k}) - a_C \geq \widehat{U}(k_\rho) - a_C \geq U(k_\rho) - 2a_C = U_\rho^{\mathrm{cert}} - 2a_C. \tag{24}$$

The inequalities also hold if either candidate is the baseline, because its estimation error is zero. Intersect with the safety event of Theorem 3 and apply a union bound to obtain probability at least $1-\alpha-\rho-\zeta$. □

The bounded-score constant can make the mean-gain bound conservative at small C. Empirical gain retention compares two selected fitted systems, whereas $U_\delta^\star$ is the largest mean gain among candidates satisfying the population risk constraint.

B CONSTRUCTIVE CONDITIONAL BOUNDS FOR PAIRED CAUSAL GAINS

This appendix establishes Assumption 3 under the randomized-logging conditions in Assumption 1. We first give a Hoeffding reference construction for general finite actions, then derive the tighter binary-treatment bound used in the experiments. Both preserve the paired AIPW estimator and its full-batch normalization.

B.1 CONDITIONAL SUPPORT OF THE PAIRED SCORE

Fix H, training state $\mathcal{T}$, candidate k, and individual i. Suppress the batch subscript and write $a = a_i^k(H)$ and $b = a_i^0(H)$. If $a = b$, define $\ell_i = u_i = 0$. Otherwise, let $c_i = m_i(a) - m_i(b)$ and, for any logging action d of positive probability and endpoint $y \in \{0, 1\}$, define

$$v_i(d, y) = c_i + \left\{ \frac{\mathbf{1}\{d = a\}}{e_i(a)} - \frac{\mathbf{1}\{d = b\}}{e_i(b)} \right\} \{y - m_i(d)\}. \tag{25}$$

Set

$$\ell_i = \min_{d:e_i(d)>0,\ y\in\{0,1\}} v_i(d, y), \qquad u_i = \max_{d:e_i(d)>0,\ y\in\{0,1\}} v_i(d, y), \qquad h_i = u_i - \ell_i. \tag{26}$$

Only $d = a$, $d = b$, and the constant value c_i for other actions need to be evaluated. The paired score is affine in the observed outcome for each logging action. Since outcomes lie in $[0, 1]$, it follows that

$$D_i(k) \in [\ell_i, u_i] \quad \text{almost surely conditional on } H, \mathcal{T}.$$

All endpoints are determined by the training predictions, logging design, and chosen actions.

B.2 A VALID ONE-SIDED RADIUS

Define

$$r_H^{\mathrm{H}}(k) = \frac{1}{N(H)} \sqrt{\frac{\log(1/\eta)}{2} \sum_{i=1}^{N(H)} h_i^2}, \qquad L_H^{\mathrm{H}}(k) = \widehat{\Delta}_H(k) - r_H^{\mathrm{H}}(k). \tag{27}$$

If the sum of squared widths is zero, set the radius to zero.

Proposition 7 (Conditional validity of the reference bound)**.** *Under Assumption 1, the lower bound* (27) *satisfies Assumption 3. The baseline lower bound remains zero.*

Proof. Condition on $H, \mathcal{T}$. Each action and prediction is fixed, and the individual scores are independent. If the squared widths sum to zero, every score is a constant and conditional unbiasedness gives $\widehat{\Delta}_H(k) = \Delta_H(k)$ almost surely. Otherwise Hoeffding's inequality gives, for $s \geq 0$,

$$\mathbb{P}\left\{ \sum_i \left(D_i(k) - \mathbb{E}[D_i(k) \mid H, \mathcal{T}] \right) > s \mid H, \mathcal{T} \right\} \leq \exp\left\{ -\frac{2s^2}{\sum_i h_i^2} \right\}.$$

Substitute $s = N r_H^{\mathrm{H}}(k)$ and use Proposition 1. This bounds $\mathbb{P}\{L_H^{\mathrm{H}}(k) > \Delta_H(k) \mid H, \mathcal{T}\}$ by η. □

The proof gives candidate-specific coverage conditional on H. Simultaneous control across the fixed candidate set is applied in the across-batch certification step.

Dependence on changed assignments. Let $d_H(k) = \sum_i \mathbf{1}\{a_i^k(H) \neq a_i^0(H)\}$ be the number of changed assignments. Since $|D_i(k)| \leq M = \underline{e}^{-1}$, $h_i \leq 2M$ on a changed assignment and $h_i = 0$ otherwise. Thus

$$r_H^{\mathrm{H}}(k) \leq \frac{M\sqrt{2 d_H(k) \log(1/\eta)}}{N(H)}. \tag{28}$$

This is a property of the baseline-paired estimator. At fixed N and $\underline{e}$, fewer changed assignments give a smaller displayed upper bound. The actual radius also depends on the support widths of the changed assignments. The estimand continues to average over the entire batch.

B.3 BINARY-TREATMENT BOUND USED IN THE EXPERIMENTS

Specialize to binary treatments and outcomes, as in the experiments, and fix $H, \mathcal{T}, k$. Write $N = N(H)$, $e_i = e_i(1)$, $m_{ai} = m_i(a)$, $\tau_i = \mu_i(1, H) - \mu_i(0, H)$, and $\nu_i = m_{1i} - m_{0i}$. Define

$$\begin{aligned} d_i &= a_i^k(H) - a_i^0(H), \qquad I = \{i : d_i \neq 0\}, \\ Z_i &= \frac{A_i(Y_i - m_{1i})}{e_i} - \frac{(1 - A_i)(Y_i - m_{0i})}{1 - e_i}, \qquad i \in I. \end{aligned} \tag{29}$$

For $i \in I$, $D_i(k) = d_i(\nu_i + Z_i)$ and $\mathbb{E}[Z_i \mid H, \mathcal{T}] = \tau_i - \nu_i$. For $i \notin I$, $D_i(k) = 0$. If I is empty, the true gain, estimate, and radius are all zero. Otherwise, for each $i \in I$, the four residual support endpoints are

$$\mathcal{R}_i = \left\{ \frac{m_{0i}}{1 - e_i}, -\frac{1 - m_{0i}}{1 - e_i}, -\frac{m_{1i}}{e_i}, \frac{1 - m_{1i}}{e_i} \right\}.$$

Let $\omega_i = \max \mathcal{R}_i - \min \mathcal{R}_i$ and define

$$\begin{aligned} a_\star &= \max_{i \in I,\, z \in \mathcal{R}_i} |z|, \qquad b_\star = \max_{i \in I} \omega_i, \\ V_H &= \frac{1}{4} \sum_{i \in I} \omega_i^2, \qquad Q = \sum_{i \in I} Z_i^2. \end{aligned} \tag{30}$$

The support constants use all possible residual endpoints. They are fixed conditional on $H, \mathcal{T}$. For the centered sum

$$T_\Delta = N(\widehat{\Delta}_H(k) - \Delta_H(k)) = \sum_{i \in I} \{d_i Z_i - \mathbb{E}[d_i Z_i \mid H, \mathcal{T}]\},$$

write $V = \sum_{i \in I} \mathrm{Var}(d_i Z_i \mid H, \mathcal{T})$ and $W = \mathbb{E}[Q \mid H, \mathcal{T}]$. The support bound on variance and $d_i^2 = 1$ on I imply

$$V \leq V_H, \qquad W = V + \sum_{i \in I} (\tau_i - \nu_i)^2 \geq V. \tag{31}$$

Thus Q is the residual second-moment sum used to construct the variance upper bound. The quantities V, W, τ_i appear only in the proof.

An observable variance upper bound. For $x > 0$, set

$$\overline{W}_x(Q) = Q + a_\star^2 x + \sqrt{a_\star^4 x^2 + 2 a_\star^2 x Q}. \tag{32}$$

We claim $\mathbb{P}\{W > \overline{W}_x(Q) \mid H, \mathcal{T}\} \leq e^{-x}$. Indeed, $Q_i = Z_i^2 \in [0, a_\star^2]$ satisfies $Q_i^2 \leq a_\star^2 Q_i$. For $t \geq 0$, the inequality $e^{-z} \leq 1 - z + z^2/2$ for $z \geq 0$, followed by conditional independence, yields

$$\mathbb{E}[e^{t(W - Q)} \mid H, \mathcal{T}] \leq \exp(t^2 a_\star^2 W / 2).$$

For $W > 0$, optimizing the Chernoff parameter gives $\mathbb{P}\{W - Q > \sqrt{2 a_\star^2 W x} \mid H, \mathcal{T}\} \leq e^{-x}$. On its complement, solving the quadratic inequality in W gives $W \leq \overline{W}_x(Q)$. The case $W = 0$ is immediate. Consequently, with

$$x = \log(2/\eta), \qquad \overline{V} = \min\{V_H, \overline{W}_x(Q)\}, \tag{33}$$

we have $\mathbb{P}\{\overline{V} < V \mid H, \mathcal{T}\} \leq \eta/2$. Thus $\overline{V}$ is an upper bound on the variance sum with conditional probability at least $1 - \eta/2$. The corresponding variance bound for $\widehat{\Delta}_H(k)$ is $\overline{V}/N^2$.

Radius and conditional coverage. The radius is

$$\begin{aligned} r_H(k) &= \frac{1}{N} \min \left\{ \sqrt{2 V_H x}, \sqrt{2 \overline{V} x} + \frac{b_\star x}{3} \right\}, \\ L_H(k) &= \widehat{\Delta}_H(k) - r_H(k). \end{aligned} \tag{34}$$

Proposition 8 (Conditional validity of the experimental lower bound)**.** *Under Assumption 1, for binary treatments and outcomes,* (34) *satisfies* $\mathbb{P}\{L_H(k) > \Delta_H(k) \mid H, \mathcal{T}\} \leq \eta$ *for every fixed candidate and almost every* H*. Thus it satisfies Assumption 3.*

Proof. All probabilities in this proof condition on $H, \mathcal{T}$. The case $I = \varnothing$ is exact. Otherwise Hoeffding's inequality gives $\mathbb{P}\{T_\Delta > \sqrt{2V_H x}\} \le e^{-x}$. Each centered summand $\xi_i = d_i Z_i - \mathbb{E}[d_i Z_i]$ obeys $|\xi_i| \le b_\star$. The inequalities $|\mathbb{E}[\xi_i^j]| \le b_\star^{j-2}\mathbb{E}[\xi_i^2]$ and $j! \ge 2 \cdot 3^{j-2}$ for $j \ge 2$ give

$$\log \mathbb{E}[e^{tT_\Delta}] \le \frac{Vt^2}{2(1 - b_\star t/3)}, \qquad 0 < t < 3/b_\star.$$

For $V > 0$, take $t = \sqrt{2x/V}/\{1 + (b_\star/3)\sqrt{2x/V}\}$ in Chernoff's inequality to obtain $\mathbb{P}\{T_\Delta > \sqrt{2Vx} + b_\star x/3\} \le e^{-x}$. For $V = 0$, $T_\Delta = 0$ almost surely.

The two thresholds $\sqrt{2V_H x}$ and $\sqrt{2Vx} + b_\star x/3$ are deterministic under the conditioning. Their exceedance events concern the same scalar T_Δ and are nested. Hence

$$\mathbb{P}\left\{T_\Delta > \min\left(\sqrt{2V_H x}, \sqrt{2Vx} + b_\star x/3\right)\right\} \le e^{-x} = \eta/2.$$

On $E_V = \{\overline{V} \ge V\}$, $Nr_H(k)$ is at least this deterministic minimum. A coverage failure is therefore contained in E_V^c or the displayed tail event. A union bound gives total probability at most $\eta/2 + \eta/2 = \eta$. □

Both deterministic tail bounds control the same centered sum. Their minimum therefore has tail probability at most $\eta/2$. The remaining $\eta/2$ error budget controls estimation of the variance upper bound. The radius is candidate-specific through the disagreement set I and its support quantities.

C Experimental Protocol and Statistical Reporting

C.1 Semi-synthetic Criteo environment

The semi-synthetic experiment uses 12 Criteo covariates (Diemert et al., 2021). A fixed partition of 2,796,600 source records assigns 400,000 records to Production training, 400,000 to nuisance training, 600,000 to candidate training, and 1,200,000 to batch generation. These pools are disjoint. Repetitions draw new training histories and evaluation batches from the corresponding pools.

Let (z_1, z_2, z_3, z_4) be standardized versions of source features (f_0, f_3, f_6, f_9). A synthetic group indicator G thresholds $0.6z_1 - 0.4z_2 + 0.2z_3$ at its source-population 70th percentile. It is used only to define the quota constraint. The response components are

$$s = \tanh(0.5z_1 - 0.3z_2 + 0.4G),$$
$$v = \tanh(0.8z_1 - 0.6z_3 + 0.3z_4), \qquad h = \tanh(0.5z_1 - 0.3z_3 + 0.2z_4).$$

The untreated probability μ_0 is produced by five cross-fitted XGBoost outcome models fit to source control observations. Their predictions are averaged and logit-calibrated using out-of-fold control predictions. This response model is fixed before the experiment and is separate from Production and from the nuisance model used for AIPW evaluation.

Within each source pool, records are ranked by h and divided into ten equal-frequency strata. Let p_j and c_j denote the empirical mass and mean h of stratum j. Each batch draws $W \sim \mathrm{Unif}[-1, 1]$ and assigns stratum weights proportional to $p_j \exp(0.35 W c_j)$. Integer stratum counts sum to the requested batch size, and records are sampled without replacement within strata. A sampled batch is retained only when it contains at least $\lfloor 3N/50 \rfloor$ members of group G, which ensures feasibility of the quota constraint.

Let $\sigma(u) = (1 + e^{-u})^{-1}$, $\ell_0 = \mathrm{logit}(\mu_0)$, and $b = 0.05 + 0.10s + 0.05(2G - 1)$. For $\gamma \in \{1, 2\}$, the treated probability is

$$\begin{aligned}\mu_1 = {} & 0.9\,\sigma(\ell_0 + \gamma\{b + (0.30 + 0.05\tanh W)v\}) \\ & + 0.1\,\sigma(\ell_0 + \gamma\{b + (-0.45 + 0.05\tanh W)v\})\,.\end{aligned} \tag{35}$$

The value of μ_0 is clipped to $[10^{-6}, 1 - 10^{-6}]$ before applying the logit. Conditional on the complete observed batch, logging draws $A_i \sim \mathrm{Bernoulli}(1/2)$ and $Y_i \mid A_i \sim \mathrm{Bernoulli}(\mu_{A_i,i})$ independently across individuals. The known propensity is therefore $e_i = 1/2$.

C.2 ALLOCATION, FITTED POLICIES, AND COMPARISON RULES

Every learned policy solves

$$\max_{a\in\{0,1\}^N} \sum_i a_i s_i, \qquad \sum_i a_i = \lfloor N/5 \rfloor, \qquad \sum_i G_i a_i \geq \lfloor 3N/50 \rfloor. \tag{36}$$

The exact optimizer treats the highest-scoring $\lfloor N/5 \rfloor$ individuals subject to the group quota. Ties at a cutoff follow a fixed record order.

Table 2: Configuration of the semi-synthetic Criteo experiment.

Component	Specification
Inputs	12 Criteo covariates, batch context W, and synthetic quota group G
Logging and outcome	Known $e = 0.5$, binary response from (35)
Allocation	Exactly 20% treated, at least 6% of the full population treated within G
Signal settings	$\gamma \in \{1, 2\}$
Repetitions	20 per setting
Batch size	$N = 500{,}000$
Calibration and future batches	$C = 160$ and $T = 256$ per fitted system
Baseline	Depth-three XGBoost T-learner, 256,000 training records
Evaluation outcome model	Separate depth-three XGBoost T-learner, 256,000 records
Raw candidates	16,000/64,000/256,000 records, T and DR learners, depths 3/5
Score-blend candidates	Two 256,000-record DR learners, three blend weights each
Candidate family	$K = 18$ per repetition
Loss and risk tolerances	$\epsilon = 0.0005$, $\delta = 0.05$
Error probabilities	$\eta = 0.01$, $\alpha = 0.05$
Certificate	One-sided paired-AIPW lower bound in (34), conditional risk conversion, exact binomial upper bound
Selection	Highest estimated mean gain among certified candidates and Production

Each repetition fits Production and the evaluation nuisance T-learner on separate histories of 64 batches with 4,000 individuals each. Candidate histories also contain 64 batches. Random-order prefixes of 250, 1,000, and 4,000 individuals yield nested training sets of 16,000, 64,000, and 256,000 records. The 12 raw candidates cross these sizes with T-learners and DR learners at depths three and five. T-learners fit separate treatment-specific binary-response models. DR learners regress the known-propensity AIPW pseudo-outcome using the separately fitted nuisance model. All models use the 14 observed inputs.

Let s_0 denote the Production score and s_j a fitted DR score. A blend candidate uses $(1-w)s_0 + ws_j$ with $w \in \{0.1, 0.25, 0.5\}$ before solving (36). The candidate family is fixed before calibration. MeanOnly selects the policy with the largest estimated mean gain among the 18 candidates and Production. NoisyCP uses $\mathbf{1}\{\widehat{\Delta}_c(k) < -\epsilon\}$ as its flag, applies the same exact binomial calibration, and then uses the same mean-gain selection rule. Our procedure uses the lower-bound flag in (8).

At $C = 160$, $K = 18$, $\alpha = 0.05$, and $\eta = 0.01$,

$$u_{\mathrm{CP}}(0, 160, \alpha/K) = 0.0361197, \qquad u_{\mathrm{CP}}(1, 160, \alpha/K) = 0.0494790,$$
$$u_{\mathrm{CP}}(2, 160, \alpha/K) = 0.0609371.$$

The certification threshold $(1 - \eta)\delta = 0.0495$ therefore permits at most one calibration flag.

For independent future evaluation, each environment uses the prespecified family of $M_f = 20 \times 2 \times 18 = 720$ fitted candidate instances. If $J_r(k)$ is the number of true underperformance events among $T = 256$ future batches, the simultaneous two-sided Clopper–Pearson interval uses per-tail error $0.05/(2M_f)$. With zero failures, its upper endpoint is 0.0393161. Future batches are independent of calibration and are used only to evaluate the fitted policy.

D Additional End-to-End Results

Figure 4 shows the paired Criteo mean gains of MeanOnly and our procedure across the 20 independent repetitions at each signal setting. Certification changes the selected candidate in 9 of 20 repetitions for $\gamma = 1$ and 4 of 20 for $\gamma = 2$.

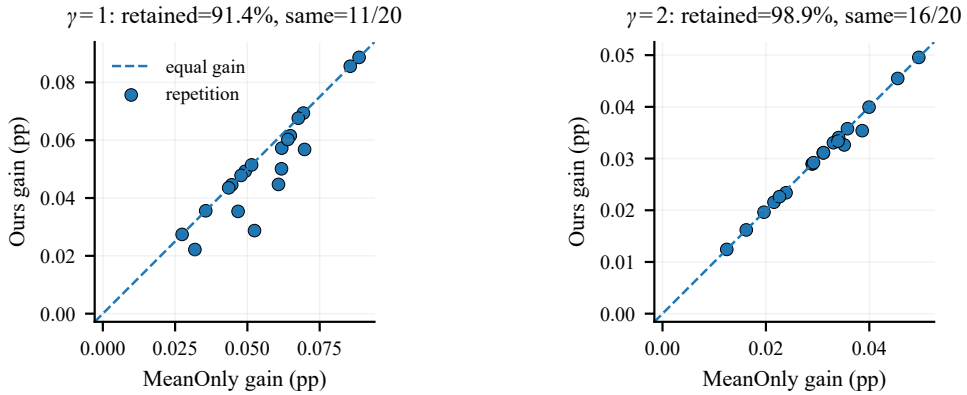

Figure 4: Paired future mean gains across 20 Criteo repetitions per signal setting. The diagonal denotes equal gain.

Table 3: Criteo uncertainty summaries for our procedure. Intervals are approximate 95% paired-bootstrap intervals from 10,000 resamples of the 20 repetitions. Retention is the ratio of mean gains to the corresponding MeanOnly mean.

Signal	Gain [interval] (pp)	Retention [interval] (%)	Same selection as MeanOnly
$\gamma = 1$	0.05139 [0.04410, 0.05906]	91.37 [85.40, 96.20]	11/20
$\gamma = 2$	0.03022 [0.02638, 0.03417]	98.88 [97.53, 99.91]	16/20

E Controlled Logging and Calibration Experiments

E.1 Fixed feasible allocation mechanism

Each batch contains $N = 50{,}000$ individuals divided equally among five observed groups indexed by $j \in \{0, 1, 2, 3, 4\}$. Exactly half of each group belongs to the quota group. Every policy treats exactly one full group, so each allocation satisfies the same 20% capacity and 6% quota requirements. Production treats group 0. Candidates 1, 2, and 3 treat groups 1, 2, and 3.

There are three batch types indexed by $h \in \{0, 1, 2\}$. Untreated response probabilities equal 0.30. The treatment response probabilities are chosen so that the exact full-batch gains of candidates 1, 2, and 3 are

Batch type	$\Delta(1)$	$\Delta(2)$	$\Delta(3)$
$h = 0$	0.02000	0.01500	0.02000
$h = 1$	0.02000	0.01500	-0.01000
$h = 2$	-0.00025	-0.00075	0.02000

Logging uses independent $A_i \sim \text{Bernoulli}(1/2)$ and binary outcomes. The fixed nuisance predictions are $m_i(0) = 0.30$ for every group and $m_i(1) = (0.35, 0.45, 0.425, 0.45, 0.35)$ for groups 0 through 4.

E.2 Repeated randomized logs

For each of the nine batch-type and candidate conditions, we generate 1,000 independent randomized logs. We record lower-bound miscoverage, point-estimate flagging, and lower-bound flagging. The 27 reported proportions use simultaneous 95% Clopper–Pearson intervals.

Table 4: Repeated-log results. Gain is the exact candidate gain in pp. The displayed flag error is a false-positive rate for safe conditions and a false-negative rate for harmful conditions.

h	Candidate	Gain (pp)	Point error (%)	Lower-bound error (%)	Miscoverage
0	1	2.000	0.0	0.0	1/1000
0	2	1.500	0.0	0.7	1/1000
0	3	2.000	0.0	0.0	1/1000
1	1	2.000	0.0	0.0	1/1000
1	2	1.500	0.0	0.6	0/1000
1	3	-1.000	0.0	0.0	0/1000
2	1	-0.025	46.3	100.0	0/1000
2	2	-0.075	46.8	0.0	0/1000
2	3	2.000	0.0	0.0	0/1000

A zero miscoverage count has simultaneous upper endpoint 0.696%, and one miscoverage has upper endpoint 0.928%. Both are below $\eta = 1\%$.

E.3 Controlled risk scan

Let the batch-type probabilities be $(p_0, p_1, p_2) = (0.94 - r, 0.06, r)$ for $r \in \{0, 0.005, \ldots, 0.06\}$. Candidate risks are

$$R(1) = 0, \qquad R(2) = r, \qquad R(3) = 0.06,$$

and their mean gains are

$$U(1) = 0.02 - 0.02025r, \qquad U(2) = 0.015 - 0.01575r, \qquad U(3) = 0.0182.$$

For each r, we generate 1,000 independent sequences of 320 batches. We run certification and selection using the first $C = 160$ batches and using all $C = 320$ batches. The corresponding admissible flag counts are $b_{160} = 2$ and $b_{320} = 7$. Simultaneous intervals cover the 52 reported unsafe-selection proportions.

Table 5: Controlled risk scan. Counts are out of 1,000 sequences at each setting and calibration size. Certification refers only to candidate two. Unsafe selection refers to any selected policy with $R > 0.05$. Upper limits are percentages from the simultaneous family over the 52 unsafe-selection proportions. The two sizes use paired sequences.

	Certified $k = 2$		Unsafe selected		Unsafe-selection upper (%)	
r (%)	$C = 160$	$C = 320$	$C = 160$	$C = 320$	$C = 160$	$C = 320$
0.0	896	999	0	0	0.7611	0.7611
0.5	693	951	1	0	0.9997	0.7611
1.0	471	819	1	0	0.9997	0.7611
1.5	316	624	2	0	1.2037	0.7611
2.0	184	339	1	0	0.9997	0.7611
2.5	100	171	0	0	0.7611	0.7611
3.0	60	94	1	0	0.9997	0.7611
3.5	33	43	3	0	1.3906	0.7611
4.0	20	15	3	0	1.3906	0.7611
4.5	7	7	4	3	1.5669	1.3906
5.0	8	3	4	1	1.5669	0.9997
5.5	1	3	3	1	1.3906	0.9997
6.0	2	0	2	1	1.2037	0.9997

At $r = 0$, candidate-two certification increases from 896/1000 to 999/1000 when C increases from 160 to 320. At $r = 1\%$, it increases from 471/1000 to 819/1000. Across all settings and both calibration sizes, unsafe final selections occur 0 to 4 times in 1,000 sequences. The largest simultaneous upper endpoint is 1.5669%, below $\alpha = 5\%$.

F Fully Synthetic End-to-End Experiment

For each batch, draw $W \sim \mathrm{Unif}[-1, 1]$. Conditional on W, draw $|f_0| \sim \mathrm{Unif}[0, 1]$ and assign its positive sign with probability $0.5 + 0.25W$. Draw $f_1, \ldots, f_{11}$ independently from $\mathrm{Unif}[-1, 1]$. A random permutation assigns exactly half of the individuals to the synthetic quota group G, independently of their features. The learner inputs are $(f_0, \ldots, f_{11}, W, G)$.

The untreated probability is $\mu_0 = \sigma((f_0 + f_1)/2)$. Set $(z_1, z_2, z_3, z_4) = \sqrt{3}(f_0, f_3, f_6, f_9)$ and use the same definitions of s and v as in Appendix C.1. Equation (35) then defines μ_1. Conditional on the observed batch, logging again uses independent $A_i \sim \mathrm{Bernoulli}(1/2)$ and binary outcomes.

Each signal setting has 20 independent repetitions. Production, the evaluation nuisance model, and the same $K = 18$ candidate family are trained from separate randomized histories using the same training sizes and learner specifications as in the Criteo experiment. Each repetition then generates $C = 160$ calibration batches and $T = 256$ future batches of size $N = 2{,}000{,}000$. Every allocation treats exactly 400,000 individuals, including at least 120,000 quota-group members.

Table 6: Synthetic uncertainty summaries for our procedure. Intervals are approximate 95% paired-bootstrap intervals from 10,000 resamples of 20 repetitions.

Signal	Gain [interval] (pp)	Retention [interval] (%)
$\gamma = 1$	$0.07874\ [0.06789, 0.09073]$	$98.55\ [96.51, 100.00]$
$\gamma = 2$	$0.07025\ [0.06496, 0.07556]$	$100.00\ [100.00, 100.00]$

Our procedure selects a candidate in all 20 repetitions at both signal settings. MeanOnly and NoisyCP also select candidates in every repetition. All three methods have zero threshold violations among their 5,120 future batches per setting. The future-risk interval uses the same prespecified $M_f = 720$ family as the Criteo experiment, so a zero-failure fitted candidate has simultaneous 95% upper endpoint 3.93%. Our procedure selects the same candidate as MeanOnly in 17/20 repetitions for $\gamma = 1$ and 20/20 for $\gamma = 2$.