LABEL-EFFICIENT GROUP RISK CONTROL FOR SELECTIVE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Certifying selective-prediction risk separately for each group requires labeled evidence from every group. Under a fixed label budget, proportional sampling can leave a small group with an overly conservative certified threshold, forcing abstention even when more of its predictions are safe to accept. We call this coverage loss *certification disparity* and cast it as a sequential label-allocation problem. Our method, SPARC, combines prediction-powered loss corrections with simultaneous betting confidence sequences, allowing labels to follow groups that still need evidence. We prove simultaneous, anytime-valid group-wise risk control under predictable allocation, data-dependent stopping and bounded importance sampling. For a fixed-configuration empirical-Bernstein variant, the certification cost separates into residual-variance and bounded-range terms, and passive and active lower bounds identify the corresponding variance orders. For the EB variant, round-robin stopping certifies targets within the sum of group requirements with high probability under the stated conditions; fixed allocation is bottlenecked by the largest requirement per share. On a controlled four-group benchmark, round-robin allocation raises worst-group coverage from 1.6% to 61.7% over proportional allocation while also increasing population coverage, and auxiliary information adds 2.54 percentage points at matched egalitarian allocation. With stopping, round-robin allocation reduces the median labels needed to reach common coverage targets by 37.1% relative to fixed uniform allocation.

1 INTRODUCTION

A selective predictor answers when its confidence exceeds a threshold and abstains otherwise (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). Risk-controlling prediction sets (RCPS) and Learn-then-Test (LTT) use labeled calibration data to choose this threshold with a high-probability risk guarantee (Bates et al., 2021; Angelopoulos et al., 2025). For populations with distinct groups, however, an aggregate guarantee can conceal large differences in error rates, and selective classification can even magnify existing disparities (Jones et al., 2021), motivating group-conditional risk control (Salem et al., 2026).

Group-wise certification introduces a second source of disparity: the evidence that supports the guarantee. A group representing 5% of an i.i.d. calibration sample receives roughly one tenth as many labels as a 50% group. Even at equal acceptance rate, risk margin and residual variance, the smaller group typically obtains a wider confidence bound, which can force additional abstention even when a more permissive threshold would satisfy the risk constraint. We call this evidence-driven coverage loss *certification disparity*; it is distinct from differences in the risk–coverage trade-off that the predictor itself can attain.

Label allocation is therefore part of calibration itself. A group that already holds a useful certificate need not keep consuming the same share of the budget, whereas a poorly certified group may benefit substantially from additional labels. Exploiting this requires certificates that remain valid when group selection and stopping depend on past observations. It also raises a statistical question: which groups need more evidence, and by how much can an auxiliary loss estimate reduce that need?

Approach. SPARC (Sequential Prediction-powered Adaptive Risk Certification) reduces selective-risk control to the sign of the mean of a centered loss. An imputed loss serves as a control variate,

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.

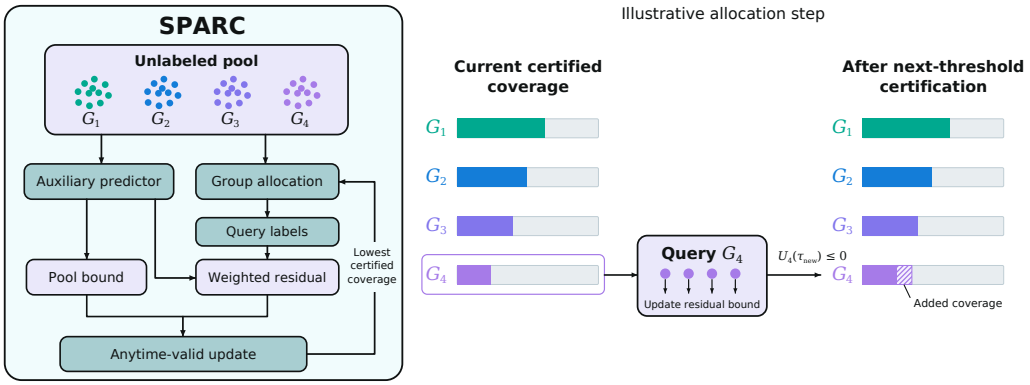

Figure 1: Overview of SPARC. Simultaneous group certificates combine a pool estimate of the imputed loss with labeled residuals. The allocation policy queries the group with the lowest certified coverage. A more permissive threshold is certified once the upper confidence bound on its centered loss drops below zero. Bar lengths are schematic.

and queried labels correct its bias (Angelopoulos et al., 2023a;b). Betting confidence sequences, maintained simultaneously for every group and candidate threshold, remain valid under predictable label allocation and data-dependent stopping. The default policy directs labels to the least-covered group; a target-based stopping policy enables comparison with fixed group allocations.

Xu et al. (2024) develop prediction-assisted sequential risk control with covariate-dependent query policies under an expected labeling-rate constraint. We focus instead on the certification cost incurred when multiple groups share a total query budget. Our contribution is a group-wise cost characterization and allocation comparison, together with a selective-risk construction that accommodates non-monotone risk over a finite threshold grid.

Contributions.

- We prove simultaneous, anytime-valid selective-risk control under predictable allocation and control-variate coefficients, optional stopping and bounded importance sampling (Theorem 1).
- For SPARC-EB, a fixed-configuration empirical-Bernstein variant, $\tilde{O}(\sigma^2/(c\Delta)^2 + W/(c\Delta))$ labels suffice, where c is the acceptance rate, Δ the risk margin, σ^2 the residual variance and W the weight bound. Instance-dependent lower bounds identify the passive and active variance orders under a fixed pre-query information model (Theorems 2–3).
- Under the stated conditions, round-robin stopping for SPARC-EB certifies all group targets within the sum of their label requirements with high probability, without knowing the problem parameters; fixed shares are bottlenecked by the largest requirement per share (Theorem 4).
- Controlled experiments separate auxiliary-signal and allocation effects, report worst-group and population coverage, and measure censoring-aware label costs to reach coverage targets (Section 5).

2 PROBLEM SETUP

Selective prediction. Let $(X, G, Y) \sim P$ with input $X \in \mathcal{X}$, observed group $G \in [K]$, group probability $\pi_g = P(G = g)$ and costly label Y. A fixed predictor f with confidence score s and loss $L = \ell(f(X), Y) \in [0, 1]$ answers at threshold τ_g iff $s(X) \ge \tau_g$. Coverage and selective risk are

$$C_g(\tau) = P(s(X) \ge \tau \mid G = g), \qquad R_g(\tau) = \mathbb{E}[L \mid s(X) \ge \tau, G = g]. \tag{1}$$

Under full abstention, both quantities are set to zero. Given $\alpha, \delta \in (0, 1)$, we seek high-coverage thresholds satisfying the PAC guarantee

$$P\{R_g(\hat{\tau}_g) \le \alpha \text{ for all } g\} \ge 1 - \delta, \tag{2}$$

where the probability is over the calibration randomness. A shared-threshold variant additionally requires all $\hat{\tau}_g$ to coincide.

Candidate region. An independent unlabeled split fixes a finite grid $\Lambda = \{\tau_1 > \cdots > \tau_M\}$. Every candidate abstains on inputs with $s(X) < \tau_M$, so labels need only be drawn from $P_g = P(\cdot \mid G = g, s(X) \geq \tau_M)$; we write $\mathbb{E}_g$ for expectation under P_g. Let $A_\tau = \mathbf{1}\{s(X) \geq \tau\}$, $c_g(\tau) = \mathbb{E}_g A_\tau > 0$ and $r_g = P(s(X) \geq \tau_M \mid G = g)$. Then $C_g(\tau) = r_g c_g(\tau)$ and $R_g(\tau) = \mathbb{E}_g[L \mid A_\tau = 1]$.

Data and label budget. Each group has an unlabeled pool $U_g \overset{\text{iid}}{\sim} P_g$ of size $N_g \geq 2$. At round t, the history $\mathcal{F}_{t-1}$ contains these pools and all previous labeled observations. An allocation rule predictably (i.e., $\mathcal{F}_{t-1}$-measurably) selects a group g_t and a proposal Q_t; a fresh $X_t \sim Q_t$ is drawn and $Y_t \sim P(Y \mid X_t, G = g_t)$ is queried. The density ratio $w_t = dP_{g_t}/dQ_t$ is known and bounded by $W \geq 1$; passive sampling has $Q_t = P_{g_t}$ and $w_t \equiv 1$. The total label budget is B, and $n_g(t)$ counts group-g labels after round t.

Auxiliary information and notation. A fixed imputer $\tilde{\ell} : \mathcal{X} \to [0, 1]$ supplies an imputed loss $\tilde{L} = \tilde{\ell}(X)$ before the label is queried; X includes all information available to the imputer. Its accuracy affects only efficiency; validity comes from the labeled correction. Define

$$Z_\tau = (L - \alpha)A_\tau, \quad \tilde{Z}_\tau = (\tilde{L} - \alpha)A_\tau, \quad \mu_g(\tau) = \mathbb{E}_g Z_\tau, \quad \nu_g(\tau) = \mathbb{E}_g \tilde{Z}_\tau.$$

Both centered losses lie in $[-\alpha, 1 - \alpha]$. In the lower bounds, the conditional label variance is always taken given the full pre-query information X. Logarithms are natural, $a \vee b = \max(a, b)$, and $\tilde{O}$ suppresses logarithmic factors.

3 SPARC: SEQUENTIAL PREDICTION-POWERED ADAPTIVE RISK CERTIFICATION

Figure 1 summarizes the update and allocation cycle. We first construct the group-wise certificates and then use them to select thresholds and allocate labels.

Selective risk as the sign of a mean. The selective risk is a ratio of two expectations and is therefore awkward to bound directly. The following identity, used implicitly by Angelopoulos et al. (2025); Angelopoulos (2026), replaces the ratio with the sign of a single mean.

Lemma 1. *For every g and $\tau \in \Lambda$ with $c_g(\tau) > 0$: $\mu_g(\tau) = c_g(\tau)\big(R_g(\tau) - \alpha\big)$. Hence $R_g(\tau) \leq \alpha \iff \mu_g(\tau) \leq 0$.*

Certification thus needs only the sign of $\mu_g(\tau)$; coverage is estimated from unlabeled data.

Prediction-powered increments with a predictable coefficient. Decompose $\mu_g(\tau) = \nu_g(\tau) + \mathbb{E}_g[Z_\tau - \tilde{Z}_\tau]$. The unlabeled pool estimates the imputed mean $\nu_g(\tau)$, and labels estimate the residual mean (Angelopoulos et al., 2023a). We account for pool uncertainty and tune the correction predictably.

A confidence bound for the pool. Fix $\delta_U, \delta_L > 0$ with $\delta_U + \delta_L = \delta$. For each g, τ, compute the empirical mean $\hat{\nu}_g(\tau)$ and the unbiased sample variance $\hat{V}_g(\tau)$ of $\tilde{Z}_\tau$ over the $N_g \geq 2$ pool inputs. Since $\tilde{Z}_\tau$ has range at most one, the empirical-Bernstein bound of Maurer & Pontil (2009, Theorem 4) gives

$$\bar{\nu}_g(\tau) = \hat{\nu}_g(\tau) + \sqrt{\frac{2\hat{V}_g(\tau)\log(2KM/\delta_U)}{N_g}} + \frac{7\log(2KM/\delta_U)}{3(N_g - 1)}. \tag{3}$$

Define the pool event $E_{\text{pool}} := \{\nu_g(\tau) \leq \bar{\nu}_g(\tau) \text{ for all } g, \tau\}$. A union bound gives $P(E_{\text{pool}}) \geq 1 - \delta_U$. The pool error shrinks as N_g grows; its size relative to $c_g \Delta_g$ determines whether the slack condition of Theorem 2 holds.

A predictable coefficient. Imputation can increase variance when the auxiliary signal is poor. A predictable coefficient lets the statistic fall back toward labeled-only calibration while retaining

validity. Following the power-tuning idea of PPI++ (Angelopoulos et al., 2023b), we use $\tilde{L}$ as a control variate with coefficient $\lambda \in [0,1]$, but we choose the coefficient *predictably*: for the t-th label of group g and each τ, $\lambda_{g,t}(\tau)$ depends only on previous labels (we use the clipped empirical regression coefficient of Z_τ on $\tilde{Z}_\tau$; equation 9). The *prediction-powered increment* for threshold τ is

$$\hat{Z}_t(\tau) \;:=\; w_t(X_t)\Big[Z_{\tau,t} - \lambda_{g_t,t}(\tau)\,\tilde{Z}_{\tau,t}\Big] \;+\; \lambda_{g_t,t}(\tau)\,\bar{\nu}_{g_t}(\tau), \tag{4}$$

where $Z_{\tau,t} = (L_t - \alpha)A_\tau(X_t)$ and $\tilde{Z}_{\tau,t} = (\tilde{L}_t - \alpha)A_\tau(X_t)$. The bracket lies in $[-1,1]$ for every $\lambda \in [0,1]$, so $\hat{Z}_t(\tau)$ is bounded and, on E_{pool}, has conditional mean at least $\mu_{g_t}(\tau)$ (Lemma 2). With $\lambda \equiv 1$ we recover the prediction-powered estimator; with $\lambda \equiv 0$ we obtain the importance-weighted labeled-only statistic $w_t(X_t)Z_{\tau,t}$, which reduces to $Z_{\tau,t}$ under passive sampling. For a fixed proposal Q_g with weights $w = dP_g/dQ_g$, the relevant residual variance is $\mathrm{Var}_{Q_g}\{w(X)(Z_\tau - \lambda\tilde{Z}_\tau)\}$, which is minimized by the clipped weighted regression coefficient

$$\lambda^\star = \mathrm{clip}_{[0,1]}\left(\frac{\mathrm{Cov}_{Q_g}(wZ_\tau, w\tilde{Z}_\tau)}{\mathrm{Var}_{Q_g}(w\tilde{Z}_\tau)}\right),$$

with the convention $\lambda^\star = 0$ when the denominator is zero. Under passive sampling ($w \equiv 1$), this reduces to the usual group-level expression. The coefficient is threshold-specific because the covariance between the loss and the imputed loss varies across acceptance regions.

Anytime-valid certificates by betting. For each pair (g,τ), we test the nulls $H_m : \mu_g(\tau) \geq m$ using the increments of group g only. Let b_t be the known upper endpoint of $\hat{Z}_t(\tau)$. For $m < b_t$, choose a *predictable* bet $\lambda_t^{\text{bet}}(m) \in [0, 1/(b_t - m))$, and set it to zero otherwise. The capital is

$$\mathcal{K}_{g,\tau,t}(m) \;=\; \prod_{i \leq t:\; g_i = g} \Big(1 - \lambda_i^{\text{bet}}(m)\big(\hat{Z}_i(\tau) - m\big)\Big). \tag{5}$$

On E_{pool} and under H_m, each factor has conditional expectation $1 - \lambda_i^{\text{bet}}(m)(\mathbb{E}[\hat{Z}_i(\tau) \mid \mathcal{F}_{i-1}] - m) \leq 1$, so $\mathcal{K}_{g,\tau,\cdot}(m)$ is a non-negative supermartingale with initial value 1. By Ville's inequality (Ville, 1939; Howard et al., 2021), it ever exceeds $1/\delta'$ with probability at most δ'. We define the *upper confidence sequence*

$$U_{g,t}(\tau) \;:=\; \min_{i \leq t}\left[\sup\left\{m \in \mathcal{M} :\; \mathcal{K}_{g,\tau,i}(m) < 1/\delta'\right\} + h\right], \tag{6}$$

where $\mathcal{M} = \{m_1, m_1 + h, \ldots, m_J\}$ is a grid of candidate means with spacing h. We set $\sup\emptyset = m_1 - h$ (so the bracket equals m_1 when every grid point is rejected) and use the value $+\infty$ whenever the largest grid point is not rejected; the $+h$ term keeps the finite-grid bound conservative. Since Ville's inequality is time-uniform, the running minimum satisfies $\mu_g(\tau) \leq U_{g,t}(\tau)$ for all t simultaneously with conditional probability at least $1 - \delta'$ on the pool event; we take $\delta' = \delta_L/(KM)$. The bets are empirical-Kelly (GRAPA) bets fitted to past increments (Waudby-Smith & Ramdas, 2024); Appendix B gives the admissible betting rule and the empirical-Bernstein construction used in the complexity analysis.

Certification and allocation. The certified set of group g is $\mathcal{C}_g(t) = \{\tau \in \Lambda : U_{g,t}(\tau) \leq 0\}$. We return $\hat{\tau}_g(t) = \min \mathcal{C}_g(t)$, or abstain on all inputs if this set is empty. Because the acceptance regions are nested, the smallest certified threshold maximizes coverage among certified candidates, so no confidence bound on coverage is needed. The running minimum in equation 6 ensures that certified sets only grow over time. The shared threshold is the smallest element of $\bigcap_g \mathcal{C}_g(t)$, again with full abstention if the intersection is empty.

Allocation may depend on the entire observed history and uses coverage estimates from unlabeled data: $\hat{C}_g(\tau) = \hat{c}_g(\tau)\hat{r}_g$, where $\hat{c}_g(\tau)$ is the acceptance fraction in the conditional pool and $\hat{r}_g$ estimates r_g from unlabeled inputs drawn from the original group population. Our default *egalitarian* rule queries the group whose current certified threshold has the lowest estimated coverage. When fixed coverage targets are given, *round-robin* cycles over the groups whose certified coverage has not yet reached its target; *proportional* allocation assigns labels according to population shares. Appendix B also describes exploration and a population-coverage rule. Algorithm 1 summarizes the procedure; each label updates $M|\mathcal{M}|$ capital processes, one per threshold and candidate mean.

Algorithm 1 SPARC: sequential prediction-powered adaptive risk certification

Input: threshold grid Λ, mean grid $\mathcal{M}$, level α, error budget $\delta = \delta_U + \delta_L$, pools $\{U_g\}$, fixed imputer $\tilde{\ell}$, label budget B, allocation rule.

For every g, τ, compute $\bar{\nu}_g(\tau)$ at level $\delta_U/(KM)$ and a coverage estimate $\hat{C}_g(\tau)$; set $\mathcal{K}_{g,\tau,0}(m) = 1$, $U_{g,0}(\tau) = +\infty$, and $\mathcal{C}_g(0) = \emptyset$. Initialize every group's decision to full abstention.
for $t = 1, \ldots, B$ **do**
  Select g_t and Q_t from the history, with known $w_t = dP_{g_t}/dQ_t \le W$.
  For every τ, choose $\lambda_{g_t,t}(\tau)$ and admissible bets from past labels only.
  Draw $X_t \sim Q_t$ and compute $s(X_t)$ and $\tilde{L}_t$; query Y_t and observe L_t.
  for $\tau \in \Lambda$ **do**
    Form $\hat{Z}_t(\tau)$ by equation 4; update all $\mathcal{K}_{g_t,\tau,t}(m)$ by equation 5 and $U_{g_t,t}(\tau)$ by equation 6.
  end for
  Set $\mathcal{C}_{g_t}(t) = \{\tau \in \Lambda : U_{g_t,t}(\tau) \le 0\}$; other groups' states are unchanged.
  Set $\hat{\tau}_{g_t}(t) = \min \mathcal{C}_{g_t}(t)$, or full abstention if the set is empty.
  Stop if the stopping condition of the allocation policy is met.
end for

Output: group thresholds $\hat{\tau}_g(t)$ with their coverage estimates, or the shared threshold $\min \bigcap_g \mathcal{C}_g(t)$ (full abstention if the intersection is empty).

4 THEORY

4.1 VALIDITY

Theorem 1 (Simultaneous, anytime-valid group risk control)**.** *Run Algorithm 1 with any allocation rule, any sampling rule with known weights bounded by W, any predictable coefficients and admissible bets as in Section 3, and any fixed imputer. Then*

$$P\Big(\text{for all } t \le B,\ \text{all } g \in [K] \text{ and all } \tau \in \mathcal{C}_g(t):\ R_g(\tau) \le \alpha\Big) \ \ge\ 1 - \delta. \tag{7}$$

In particular, the group-specific thresholds $\hat{\tau}_g(T)$ and the shared threshold satisfy equation 2 at every (possibly data-dependent) stopping time $T \le B$.

The proof uses the global filtration, under which a round allocated to another group multiplies group g's capital by one. Ville's inequality, a union bound over the KM pairs and the pool event then give simultaneous validity (Appendix A.3).

4.2 LABEL COMPLEXITY OF CERTIFICATION

For the efficiency analysis, we define SPARC-EB by replacing the betting bound with the empirical-Bernstein sequence of Appendix B.2 and keeping each coefficient and proposal fixed. Write $u^{\mathrm{EB}}_{g,k}(\tau)$ for its fixed-sample bound after k labels and

$$U^{\mathrm{EB}}_{g,t}(\tau) = \min_{2 \le k \le n_g(t)} u^{\mathrm{EB}}_{g,k}(\tau), \qquad \mathcal{C}^{\mathrm{EB}}_g(t) = \{\tau : U^{\mathrm{EB}}_{g,t}(\tau) \le 0\},$$

with $U^{\mathrm{EB}}_{g,t}(\tau) = +\infty$ before two labels. The returned threshold is $\hat{\tau}^{\mathrm{EB}}_g = \min \mathcal{C}^{\mathrm{EB}}_g$, with full abstention if the set is empty. A union bound over groups, thresholds and sample sizes gives the same simultaneous risk guarantee, and the explicit EB width yields the following efficiency result.

Theorem 2 (Label complexity of SPARC-EB)**.** *Fix a group g and a threshold $\tau \in \Lambda$ with* margin $\Delta := \alpha - R_g(\tau) > 0$*, acceptance rate $c := c_g(\tau)$, and a fixed coefficient $\lambda \in [0, 1]$. Suppose that group g is labeled with a proposal Q_g fixed in advance, with known weights $w = dP_g/dQ_g \le W$ (passive: $Q_g = P_g$, $W = 1$), and that the labeled draws are independent of the pool. Let $\sigma^2 := \mathrm{Var}_{Q_g}\big(w(X)(Z_\tau - \lambda \tilde{Z}_\tau)\big)$ be the* residual variance*. Assume the pool event E_{pool} of Section 3 and the* slack condition *$\lambda(\bar{\nu}_g(\tau) - \nu_g(\tau)) \le c\Delta/2$. Conditional on any realized pool satisfying both conditions and for any $\delta'' \in (0, 1)$, with probability at least $1 - \delta''$ the threshold τ belongs to $\mathcal{C}^{\mathrm{EB}}_g(t)$ for every t with $n_g(t) \ge n^\star$. If the pool satisfies both conditions with probability at least $1 - \delta_U$, the*

unconditional probability is at least $1 - \delta_U - \delta''$*;* E_{pool} *alone does not ensure the slack condition. Here, with* $\ell_n := \log \frac{2n(n+1)KM}{\delta_L} \vee \log \frac{2}{\delta''}$,

$$n^\star := \min\Big\{n \geq 2 :\ 2\sqrt{2\sigma^2 \ell_n/n} + 10\, W \ell_n/(n-1) \leq c\Delta/2\Big\} \\ \leq C\,\kappa \log\big(KM\kappa/(\delta_L \delta'')\big), \qquad \kappa := \frac{\sigma^2}{(c\Delta)^2} + \frac{W}{c\Delta}. \tag{8}$$

for an absolute constant C.

Imputation acts through the *variance term* $\sigma^2/(c\Delta)^2$, via the residual variance of the labeled correction; the *range term* $W/(c\Delta)$ is the finite-sample price of bounded importance-weighted increments. The effective margin is $c\Delta$ because only accepted inputs carry information. This bound assumes a fixed coefficient and proposal; Theorem 1 also covers predictable coefficients and adaptive proposals.

Corollary 1 (Certification price)**.** *Suppose that* $\tau_g^\circ(\Delta)$*, the most permissive grid threshold with* $R_g(\tau) \leq \alpha - \Delta$*, exists and satisfies the conditions of Theorem 2. Conditional on a pool satisfying the pool event and the slack condition, with probability at least* $1 - \delta''$*,*

$$C_g(\hat{\tau}_g^{\mathrm{EB}}(t)) \geq C_g(\tau_g^\circ(\Delta)) \quad \textit{whenever } n_g(t) \geq n^\star.$$

Inverting equation 8 shows that, after n labels, a margin of order $c^{-1}(\sigma\sqrt{\ell_n/n} + W\ell_n/n)$ suffices to certify a fixed candidate; we call this margin the *certification price*. Both terms persist under proportional allocation with $n_g \approx \pi_g n$: subject to the slack condition, a sufficient margin is $\Delta_g \geq Cc_g^{-1}(\sigma_g\sqrt{\ell_{n_g}/n_g} + W\ell_{n_g}/n_g)$ for an absolute constant C. When the variance term dominates and all other quantities agree, a 5% group pays about $\sqrt{10}$ times the margin of a 50% group. Where risk grows slowly with coverage near α, this larger margin can cost substantial coverage.

4.3 Lower bounds: what unlabeled data can and cannot buy

For the lower bounds, the predictor, score, supplied imputer and marginal distribution of every pre-query signal are held fixed across instances; only the conditional label law changes. Fix a group and a threshold, and let the losses be Bernoulli with $p(x) := P(L = 1 \mid X = x)$, where X includes every signal available before the label is queried; $v(x) := p(x)(1 - p(x))$ is then the conditional label variance under this information set. A certifier is δ*-valid* if it certifies τ with probability at most δ under every instance with $R_g(\tau) > \alpha$, and (Δ, δ')*-powerful at* p_0 if it certifies τ with probability at least $1 - \delta'$ under p_0, an instance with $R_g(\tau) = \alpha - \Delta$.

Theorem 3 (Lower bounds)**.** *Let* $\mathcal{X}$ *be countable and* $\delta + \delta' < 1/2$*. Fix an instance* p_0 *with margin* $\Delta := \alpha - R_g(\tau) > 0$ *and acceptance rate* $c := c_g(\tau)$*, write* $v(x) := p_0(x)(1 - p_0(x))$*, and suppose that*

$$\bar{p}_0 := \sup_{x : A_\tau(x) = 1} p_0(x) < 1/2, \qquad 0 < \mathbb{E}_g[v(X)A_\tau(X)] < \infty.$$

Let $\Delta_0 > 0$ *be the constant determined by* (p_0, P_g, A_τ) *in Appendix A.7. If* $\Delta \leq \Delta_0$*:*

(a) (passive*) any* δ*-valid certifier that uses* n *i.i.d. labeled samples from* P_g*, together with an arbitrary amount of unlabeled data, and is* (Δ, δ')*-powerful at* p_0 *satisfies* $n \geq \dfrac{\mathbb{E}_g[v(X)A_\tau(X)]}{4c^2\Delta^2} \log \dfrac{1}{2(\delta + \delta')}$*;*

(b) (active*) any* δ*-valid certifier that is* (Δ, δ')*-powerful at* p_0 *and chooses its* n *labeled inputs adaptively, with expected allocation* $q(x) := \mathbb{E}_{p_0}[N_n(x)]/n \geq \kappa\, P_g(x)$ *on* $\{A_\tau = 1\}$ *for some* $\kappa > 0$*, where* $N_n(x)$ *is the number of labels queried at* x*, satisfies, for* $\Delta \leq \kappa\Delta_0$*,* $n \geq \dfrac{\big(\mathbb{E}_g[\sqrt{v(X)}A_\tau(X)]\big)^2}{4c^2\Delta^2} \log \dfrac{1}{2(\delta + \delta')}$.

The proof perturbs the conditional label law on the accepted region while preserving every pre-query marginal; the active case uses the divergence decomposition of Kaufmann et al. (2016) (Appendix A.7). By total variance, $\mathrm{Var}_g(Z_\tau - \lambda\tilde{Z}_\tau) \geq \mathbb{E}_g[vA_\tau]$ for every imputer and λ. The conditional-risk imputer with $\lambda = 1$ attains equality at the reference instance, although it need not be

correct under the alternative instances. The active floor $(\mathbb{E}_g[\sqrt{v}A_\tau])^2$ never exceeds the passive one and can be much smaller when the variance is concentrated. These are variance-order comparisons under the stated information model; the finite-sample range cost remains separate.

Attaining the active variance order. With the conditional-risk imputer, the proposal $(1-\beta)P_g + \beta Q_v$ with $dQ_v/dP_g \propto \sqrt{v}A_\tau$ satisfies

$$W \le (1-\beta)^{-1}, \qquad \sigma^2 \le \beta^{-1}\big(\mathbb{E}_g[\sqrt{v}A_\tau]\big)^2.$$

Thus any fixed $\beta \in (0,1)$ attains the active variance order up to constants when that term dominates (Proposition 1, Appendix A.9). Since v is unknown, practical proposals use a predictable proxy; they retain validity whenever their weights are known and bounded.

4.4 Adaptive allocation across groups

Theorem 4 (Adaptive versus static allocation)**.** *For each group g, let $\tau_g^\circ \in \Lambda$ be a* pre-specified *target threshold with margin $\Delta_g > 0$, and let $n_g^\star$ be given by equation 8. (a) Run* SPARC-EB *with $B \ge \sum_g n_g^\star$, fix each group's coefficient and proposal as in Theorem 2, and condition on pools satisfying the pool event and the slack condition for every target. The round-robin rule that cycles over the groups with $\tau_g^\circ \notin \mathcal{C}_g^{\mathrm{EB}}(t)$ certifies all K targets within $\sum_g n_g^\star$ one-at-a-time labels with probability at least $1 - K\delta''$, without knowing $(\sigma_g, c_g, \Delta_g)$. With batches of size b, it needs at most $K(b-1)$ additional labels and a correspondingly larger budget. (b) Under the assumptions of Theorem 3(a), consider passive sampling with fixed group counts $a_g B$, or with i.i.d. group draws with probabilities a_g. Any δ-valid certifier that certifies all targets with probability at least $1-\delta'$ requires $B \ge \max_g n_g^{\mathrm{LB}}/a_g$, where n_g^{LB} is the passive lower bound for group g. A uniform split thus requires $K \max_g n_g^{\mathrm{LB}}$, and an i.i.d. calibration set ($a_g = \pi_g$) requires $\max_g n_g^{\mathrm{LB}}/\pi_g$.*

For the conditional-risk imputer with $\lambda = 1$, the variance terms of $n_g^\star$ and n_g^{LB} agree up to constants and logarithms. In this oracle, variance-dominated regime, round-robin with stopping can thus save up to a factor K over a uniform split and $1/\pi_{\min}$ over proportional allocation; for an arbitrary imputer, the bounds need not match. Part (a) is conditional on the pool conditions; an unconditional statement adds their failure probability. The egalitarian and greedy rules are evaluated empirically, without optimality claims. Anytime validity lets each group stop at a data-dependent sample size without losing its guarantee.

5 Experiments

We address two questions: (i) how auxiliary information changes certification under a fixed allocation rule, and (ii) how allocation changes the operating points reached under a common label budget. A controlled population permits exact evaluation of the returned risks and coverages and makes the source of the auxiliary signal explicit.

Benchmark and protocol. The four groups have population shares $0.50/0.30/0.15/0.05$; each is a fixed population of 10^6 records with a noisy confidence score and a binary loss. The auxiliary signal, drawn from $\mathrm{Beta}(10,1)$ on erroneous records and $\mathrm{Beta}(1,10)$ on correct ones, is revealed before the label is queried and acts as a controlled side-information channel. Smaller groups also contain harder examples and noisier scores; comparisons between allocation rules hold these population differences fixed. Appendix D specifies the full generator.

An independent score sample fixes a grid of $M = 25$ thresholds. Each run draws 10^5 unlabeled inputs per group (before restriction to the candidate region) and uses $\alpha = 0.05$, $\delta = 0.1$ and label batches of size 10. The main experiment uses $B = 8000$ and 30 calibration seeds. Risks and coverages are computed exactly over the fixed population; uncertainty reflects variation across calibration seeds, conditional on this population and grid.

Baselines. LTT-CP applies Clopper–Pearson bounds with a Bonferroni correction over groups and thresholds (Angelopoulos et al., 2025), evaluated at the fixed budget under proportional allocation. The labeled-only control sets $\lambda = 0$ in the same betting certifier and spends the entire error budget on

Table 1: Four-group comparison at $B = 8000$. Coverage C (%): mean $\pm$ one standard error over 30 calibration seeds. $R_{\max}$: seed-averaged worst-group risk (%). No row violates the 5% risk target at the final budget in any seed.

Certificate	Allocation	Worst-group C	Population C	$R_{\max}$
LTT-CP	Proportional	12.62 ± 4.27	80.62 ± 0.24	3.51
Labeled-only	Proportional	2.11 ± 2.11	78.77 ± 0.23	3.35
SPARC + auxiliary	Proportional	1.61 ± 1.61	80.54 ± 0.13	3.59
Labeled-only	Egalitarian	63.52 ± 0.44	73.43 ± 0.51	3.47
SPARC + auxiliary	Round-robin	61.70 ± 0.42	83.72 ± 0.09	3.54
SPARC + auxiliary	Egalitarian	66.06 ± 0.22	76.92 ± 0.40	3.87

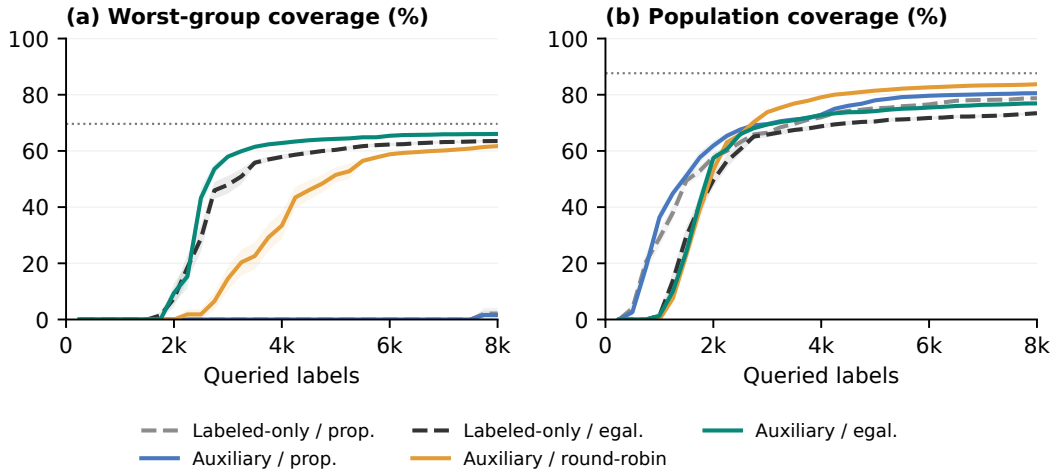

Figure 2: Worst-group and population coverage versus labels. Curves and bands show means and one standard error over 30 seeds, evaluated exactly on the fixed population. Horizontal lines mark the corresponding group-wise risk-feasible grid-oracle coverages.

labeled evidence, whereas prediction-powered runs use $\delta_U = 0.01$ and $\delta_L = 0.09$. The shared betting range, grid and allocation rule provide a matched comparison of labeled-only and prediction-powered certificates. The proportional, round-robin and egalitarian policies differ only in their allocation rule, and sampling within each group is passive ($W = 1$).

Auxiliary information at matched allocation. With the allocator and betting construction held fixed at egalitarian allocation, the auxiliary correction raises worst-group coverage by 2.54 percentage points over the labeled-only control (paired 95% interval $[1.62, 3.47]$) and population coverage by 3.49 points. Under proportional allocation, the auxiliary signal improves population coverage (80.54% versus 78.77%) but does not resolve the lack of certification for the smallest group. Reducing residual variance and reallocating labels thus relieve different constraints.

Allocation and the coverage trade-off. With the auxiliary signal fixed, worst-group coverage rises from 1.61% under proportional allocation to 61.70% under round-robin and 66.06% under egalitarian allocation; population coverage is 80.54%, 83.72% and 76.92%, respectively. Round-robin thus improves both metrics over proportional allocation. Egalitarian allocation spends on average 72.5% of the labels on the 5% group, so its additional worst-group gain over round-robin comes at a cost of 6.8 points in population coverage. Figure 2 traces these operating points as the budget grows. Table 1 reports fixed-budget violations. For SPARC, simultaneous validity over all budgets follows from Theorem 1.

Labels to reach common coverage targets. Before calibration, we fix per-group coverage targets of 84.68/77.79/72.03/63.17%; a target is reached once the returned certified threshold attains at least that coverage. All policies share the auxiliary signal and betting certificate. Across 16 runs,

Table 2: Labels needed to reach all four coverage targets. Kaplan–Meier medians include every run, with unfinished runs right-censored at 16,000 labels. Results use 16 seeds at a resolution of 10-label batches.

Allocation	Median labels	Unfinished (of 16)
Proportional	Not reached	16
Uniform	7200	1
Round-robin with stopping	**4530**	0
Egalitarian with stopping	4700	0

round-robin with stopping reduces the Kaplan–Meier median label cost from 7200 under fixed uniform allocation to 4530, a 37.1% reduction; egalitarian allocation needs 4700 labels, a 34.7% reduction (Table 2). Fixed proportional allocation fails to reach all four targets within 16,000 labels in every run. These experiments probe the budget bottleneck that motivates Theorem 4, whose quantitative bound concerns SPARC-EB.

Additional settings. The allocation effect persists with eight groups (the smallest with a 1.5% share) and at a looser risk level. For eight groups at $B = 12{,}000$, auxiliary-assisted egalitarian allocation reaches $59.51 \pm 0.52\%$ worst-group coverage, whereas proportional allocation with the same signal attains zero. At $\alpha = 0.10$ and $B = 6000$, the corresponding values are $83.95 \pm 0.38\%$ and $56.49 \pm 5.68\%$. Appendix C reports population coverage and fixed-budget risk.

6 Related work

Selective and group-conditional risk control. Selective prediction studies the trade-off between errors and abstention (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). RCPS and LTT provide high-probability calibration guarantees (Bates et al., 2021; Angelopoulos et al., 2025); conformal risk control and its extensions study related expectation guarantees (Angelopoulos et al., 2024; Angelopoulos, 2026; Bai & Jin, 2026). Group disparities motivate conditional and multivalid calibration (Jones et al., 2021; Gibbs et al., 2025; Jung et al., 2023; Salem et al., 2026). We instead study the label cost of obtaining simultaneous selective-risk certificates for specified groups.

Prediction-assisted sequential inference. Prediction-powered inference uses labeled residuals to correct auxiliary predictions; power tuning, stratification and active sampling improve its efficiency (Angelopoulos et al., 2023a;b; Fisch et al., 2024; Zrnic & Candès, 2024). Einbinder et al. (2025) apply this principle to fixed-sample risk control. Xu et al. (2024) combine prediction-assisted variance reduction with active, anytime-valid RCPS and analyze query policies under an expected labeling-rate constraint, and Hultberg et al. (2026) study anytime-valid conformal risk control. Our shared-budget formulation centers on group-specific certification costs and the gap between adaptive allocation with stopping and fixed group shares. Our lower bounds quantify the conditional variance that remains once all pre-query information is used.

7 Discussion

SPARC allocates calibration evidence across groups while preserving simultaneous selective-risk control. The fixed-configuration EB analysis separates residual-variance and bounded-range costs and ties target certification to the sum of the per-group requirements. Experiments with betting certificates separate the gains from auxiliary information and allocation and quantify their population-coverage cost. A shared-threshold variant supports deployments that require a common acceptance rule. The guarantees require observed groups, fresh calibration draws and known bounded importance weights. Controlled signals isolate the mechanism; learned judges on external datasets remain to be evaluated.

AI USE STATEMENT

Generative AI tools assisted with manuscript revision, literature checking, clarification of theoretical statements, code inspection, and preparation of statistical summaries and figures. Empirical results were computed from the accompanying experimental records and executions of the supplied benchmark code. The authors are responsible for the final manuscript, proofs, code, and reported results.

REFERENCES

Anastasios N. Angelopoulos. Conformal risk control for non-monotonic losses. *arXiv preprint arXiv:2602.20151*, 2026.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a.

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *Annals of Applied Statistics*, 19(2):1641–1662, 2025. arXiv:2110.01052.

Tian Bai and Ying Jin. Conformal selective prediction with general risk control. *arXiv preprint arXiv:2603.24704*, 2026.

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

Bat-Sheva Einbinder, Liran Ringel, and Yaniv Romano. Semi-supervised risk control via prediction-powered inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12): 11080–11090, 2025. doi: 10.1109/TPAMI.2025.3594263. First posted as arXiv:2412.11174.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Adam Fisch, Joshua Maynez, R. Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William W. Cohen. Stratified prediction-powered inference for effective hybrid evaluation of language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B*, 87(4):1100–1126, 2025. doi: 10.1093/jrsssb/qkaf008. First posted as arXiv:2305.12616.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *Annals of Statistics*, 49(2):1055–1080, 2021.

Bror Hultberg, Dave Zachariah, and Antônio H. Ribeiro. Anytime-valid conformal risk control. *arXiv preprint arXiv:2602.04364*, 2026.

Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. Selective classification can magnify disparities across groups. In *International Conference on Learning Representations*, 2021.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. In *International Conference on Learning Representations*, 2023.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Conference on Learning Theory*, 2009.

Murilo Salem, Luísa Böhm, Daniel Pontes, and Anderson Ferrugem. Hierarchical group-conditional conformal risk control for selective prediction in language models. *arXiv preprint arXiv:2607.24562*, 2026.

Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer, 2009.

Jean Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, 1939.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B*, 86(1):1–27, 2024.

Ziyu Xu, Nikos Karampatziakis, and Paul Mineiro. Active, anytime-valid risk controlling prediction sets. In *Advances in Neural Information Processing Systems*, volume 37, pp. 60110–60132, 2024. doi: 10.52202/079017-1920. First posted as arXiv:2406.10490.

Tijana Zrnic and Emmanuel J. Candès. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62993–63010. PMLR, 2024.

A PROOFS

Throughout, $(\mathcal{F}_t)_{t\geq 0}$ is the filtration generated by the pools ($\mathcal{F}_0 = \sigma(U_1, \ldots, U_K)$ together with any independent internal randomness) and the labeled rounds. We assume $c_g(\tau) > 0$ for all $g, \tau \in \Lambda$ (every candidate accepts a positive fraction of every group), so that $R_g(\tau)$ is well defined.

A.1 PROOF OF LEMMA 1

$\mu_g(\tau) = \mathbb{E}_g[(L-\alpha)A_\tau] = \mathbb{E}_g[LA_\tau] - \alpha\,\mathbb{E}_g[A_\tau] = c_g(\tau)\,\mathbb{E}_g[L \mid A_\tau = 1] - \alpha c_g(\tau) = c_g(\tau)(R_g(\tau) - \alpha)$. Since $c_g(\tau) > 0$ the sign of $\mu_g(\tau)$ is the sign of $R_g(\tau) - \alpha$. □

A.2 CONDITIONAL MEAN AND RANGE OF THE INCREMENTS

Lemma 2. *Fix a round t with $g_t = g$ and let $\lambda := \lambda_{g,t}(\tau) \in [0,1]$ and $\bar{\nu} := \bar{\nu}_g(\tau)$, both $\mathcal{F}_{t-1}$-measurable. Then $\hat{Z}_t(\tau) \in [-W + \lambda\bar{\nu},\ W + \lambda\bar{\nu}]$ and $\mathbb{E}[\hat{Z}_t(\tau) \mid \mathcal{F}_{t-1}] = \mu_g(\tau) + \lambda\big(\bar{\nu}_g(\tau) - \nu_g(\tau)\big)$, which is at least $\mu_g(\tau)$ on the event E_{pool}.*

Proof. Since $L, \tilde{L} \in [0,1]$ and $\lambda \in [0,1]$, $L - \lambda\tilde{L} - (1-\lambda)\alpha \in [-\lambda - (1-\lambda)\alpha,\ 1] \subseteq [-1, 1]$; multiplying by $A_\tau \in \{0,1\}$ and by $w_t \in [0, W]$ gives the range. Conditionally on $\mathcal{F}_{t-1}$, $X_t \sim Q_t$ and $Y_t \sim P(Y \mid X_t, G = g)$, so for any bounded $h(x,y)$, $\mathbb{E}[w_t(X_t)h(X_t,Y_t) \mid \mathcal{F}_{t-1}] = \mathbb{E}_{X\sim Q_t}[w_t(X)\,\mathbb{E}[h(X,Y) \mid X, G = g]] = \mathbb{E}_{X\sim P_g}[\mathbb{E}[h(X,Y) \mid X, G = g]] = \mathbb{E}_g[h]$ because $w_t = dP_g/dQ_t$. Applying this to $h = Z_\tau$ and $h = \tilde{Z}_\tau$ (the latter does not depend on y) yields $\mathbb{E}[\hat{Z}_t(\tau) \mid \mathcal{F}_{t-1}] = \mu_g(\tau) - \lambda\nu_g(\tau) + \lambda\bar{\nu}_g(\tau)$. On E_{pool}, $\bar{\nu}_g(\tau) \geq \nu_g(\tau)$. □

A.3 PROOF OF THEOREM 1

Fix a pair (g, τ) and write $\mu := \mu_g(\tau)$. Let $\mathcal{M} = \{m_1 < \cdots < m_J\}$ be the grid of candidate values with spacing h. In the definition of $U_{g,t}$, use the endpoint conventions $\sup\emptyset = m_1 - h$ and return $+\infty$ when the largest grid point is not rejected. Thus the returned value is m_1 when every grid point is rejected; otherwise it is the largest unrejected grid point plus h, or $+\infty$ at the upper endpoint. Let $m^\dagger$ be the largest grid point with $m^\dagger \leq \mu$ if one exists ($m^\dagger := m_J$ if $\mu > m_J$). If $\mu < m_1$, there is nothing to prove, since the returned value is at least $m_1 > \mu$. Define, for every round t,

$$\mathcal{K}_t := \prod_{i\leq t} F_i, \qquad F_i := 1 - \mathbf{1}\{g_i = g\}\,\lambda_i^{\mathrm{bet}}(m^\dagger)\big(\hat{Z}_i(\tau) - m^\dagger\big),$$

where $b_i = W + \lambda_{g,i}(\tau)\bar{\nu}_g(\tau)$ is the upper end of the increment range (Lemma 2). The predictable bet belongs to $[0, 1/(b_i - m^\dagger))$ when $m^\dagger < b_i$ and is zero otherwise. Each factor is non-negative: a zero bet gives $F_i = 1$; if $\hat{Z}_i(\tau) \leq m^\dagger$ then $F_i \geq 1$; otherwise the admissible bet gives $\lambda_i^{\mathrm{bet}}(\hat{Z}_i - m^\dagger) < (\hat{Z}_i - m^\dagger)/(b_i - m^\dagger) \leq 1$. This is exactly the capital equation 5 evaluated along the rounds of group g; rounds of other groups multiply it by one. Since g_i is $\mathcal{F}_{i-1}$-measurable, on the event $E_{\mathrm{pool}} \in \mathcal{F}_0$,

$$\begin{aligned}\mathbb{E}[F_i \mid \mathcal{F}_{i-1}] &= 1 - \mathbf{1}\{g_i = g\}\lambda_i^{\mathrm{bet}}(m^\dagger)\big(\mathbb{E}[\hat{Z}_i(\tau) \mid \mathcal{F}_{i-1}] - m^\dagger\big)\\ &\leq 1 - \mathbf{1}\{g_i = g\}\lambda_i^{\mathrm{bet}}(m^\dagger)(\mu - m^\dagger) \leq 1,\end{aligned}$$

using Lemma 2 and $m^\dagger \leq \mu$. Hence $(\mathcal{K}_t \mathbf{1}_{E_{\mathrm{pool}}})_{t\geq 0}$ is a non-negative supermartingale with initial value at most 1, and Ville's inequality gives $P\big(E_{\mathrm{pool}} \cap \{\exists t : \mathcal{K}_t \geq 1/\delta'\}\big) \leq \delta'$. On the complementary event within E_{pool}, $m^\dagger$ is never rejected, so every one-time bound in equation 6 is at least $m^\dagger + h > \mu$ when $m^\dagger < m_J$. If $m^\dagger = m_J$, the bound is $+\infty$, including when $\mu > m_J$. The running intersection over $i \leq t$ preserves this. Therefore $P\big(E_{\mathrm{pool}} \cap \{\exists t :\ U_{g,t}(\tau) < \mu_g(\tau)\}\big) \leq \delta' = \delta_L/(KM)$.

Let $E_L := \{\mu_g(\tau) \leq U_{g,t}(\tau)\ \forall g, \tau, t\}$. A union bound over the KM pairs gives $P(E_{\mathrm{pool}} \cap E_L^c) \leq \delta_L$, and the pool bounds give $P(E_{\mathrm{pool}}^c) \leq \delta_U$, so $P(E_{\mathrm{pool}} \cap E_L) \geq 1 - \delta$. On $E_{\mathrm{pool}} \cap E_L$, every $\tau \in \mathcal{C}_g(t)$ satisfies $\mu_g(\tau) \leq U_{g,t}(\tau) \leq 0$, hence $R_g(\tau) \leq \alpha$ by Lemma 1. Nothing in the argument depends on how g_i, Q_i, $\lambda_{g,i}$ or the bets are chosen, as long as they are $\mathcal{F}_{i-1}$-measurable, nor on the imputer, nor on when the algorithm stops. □

A.4 THE LABELED SEQUENCE OF A GROUP IS I.I.D. UNDER A FIXED PROPOSAL

For Theorems 2 and 4 we use the following standard device. Let $(\xi_{g,k})_{k\ge1}$, $\xi_{g,k} = (X_{g,k}, Y_{g,k})$, be i.i.d. draws from $Q_g \otimes P(Y \mid X, G = g)$, independent across g and of the pools, generated before the algorithm runs; the k-th time the allocation rule picks group g it reveals $\xi_{g,k}$. This realization has the same law as the sequential protocol of Section 2, and the labeled sequence of group g is $(\xi_{g,k})_{k\le n_g(t)}$: an i.i.d. prefix whose (random) length is determined by the allocation. Consequently, a fixed-sample inequality that holds for every k with failure probability δ_k holds simultaneously for all k, hence at $k = n_g(t)$, with failure probability $\sum_k \delta_k$ (the union-bound construction of Appendix B.2).

A.5 PROOF OF THEOREM 2

Write $D_i := w(X_{g,i})\big(Z_\tau(\xi_{g,i}) - \lambda\tilde{Z}_\tau(\xi_{g,i})\big) \in [-W, W]$ for the labeled part of the i-th increment of group g, which for a fixed λ and fixed proposal are i.i.d. with mean $\mathbb{E}D = \mu_g(\tau) - \lambda\nu_g(\tau)$ (Lemma 2) and variance σ^2. The one-time empirical-Bernstein upper bound of Appendix B.2 satisfies, after $k \ge 2$ labels,

$$u^{\mathrm{EB}}_{g,k}(\tau) \le \lambda\bar{\nu}_g(\tau) + \bar{D}_k + \sqrt{2\hat{V}_k\ell_k/k} + \tfrac{14W\ell_k}{3(k-1)}, \qquad \ell_k = \log\tfrac{2k(k+1)KM}{\delta_L} \vee \log\tfrac{2}{\delta''},$$

with $\bar{D}_k$ the sample mean and $\hat{V}_k$ the unbiased sample variance, and $U^{\mathrm{EB}}_{g,t}(\tau) = \min_{2\le k\le n_g(t)} u^{\mathrm{EB}}_{g,k}(\tau)$. Let $k = n^\star$ and $L := \log(2/\delta'')$. By Bernstein's inequality for i.i.d. variables with $|D_i - \mathbb{E}D| \le 2W$, with probability at least $1 - \delta''/2$, $\bar{D}_k \le \mathbb{E}D + \sqrt{2\sigma^2L/k} + 4WL/(3k)$; by Theorem 10 of Maurer & Pontil (2009) (rescaled to range $2W$), with probability at least $1 - \delta''/2$, $\sqrt{\hat{V}_k} \le \sigma + 2W\sqrt{2L/(k-1)}$. On the intersection of the two events, using $\ell_k \ge L$ and $\sqrt{2\ell_k/k}\cdot 2W\sqrt{2L/(k-1)} \le 4W\ell_k/(k-1)$,

$$\begin{aligned} u^{\mathrm{EB}}_{g,k}(\tau) &\le \mu_g(\tau) + \lambda\big(\bar{\nu}_g(\tau) - \nu_g(\tau)\big) + 2\sqrt{2\sigma^2\ell_k/k} + \frac{10W\ell_k}{k-1} \\ &\le -c\Delta + \frac{c\Delta}{2} + 2\sqrt{\frac{2\sigma^2\ell_k}{k}} + \frac{10W\ell_k}{k-1} \le 0, \end{aligned}$$

where we used $\mu_g(\tau) = -c\Delta$ (Lemma 1), the pool condition, and the definition equation 8 of $n^\star$. Hence $U^{\mathrm{EB}}_{g,t}(\tau) \le u^{\mathrm{EB}}_{g,n^\star}(\tau) \le 0$ for every t with $n_g(t) \ge n^\star$, i.e., $\tau \in \mathcal{C}^{\mathrm{EB}}_g(t)$.

For the explicit bound, note that equation 8 holds as soon as $2\sqrt{2\sigma^2\ell_n/n} \le c\Delta/4$ and $10W\ell_n/(n-1) \le c\Delta/4$, i.e., as soon as $n \ge 1 + 128\kappa\ell_n$ with $\kappa = \sigma^2/(c\Delta)^2 + W/(c\Delta)$. Since $\ell_n \le \log(bn^2)$ with $b := 8KM/(\delta_L\delta'')$, it suffices that $n \ge a(\log(bn^2) + 1)$ with $a := 128\kappa \vee 1$. Lemma 3 below shows that this holds for all $n \ge 8a\log(8ab)$, so $n^\star \le 8a\log(8ab) \le C\kappa\log(KM\kappa/(\delta_L\delta''))$ for an absolute constant C. □

Lemma 3. *Let $a \ge 1$ and $b \ge e$. Then $n \ge a(\log(bn^2) + 1)$ for every $n \ge n_0 := 8a\log(8ab)$.*

Proof. The map $n \mapsto n - a\log(bn^2)$ is increasing for $n \ge 2a$, and $n_0 \ge 8a\log 8 \ge 2a$, so it suffices to check $n = n_0$. Since $\log\log x \le \log x - 1$ for $x \ge e$, $\log n_0 = \log(8a) + \log\log(8ab) \le 2\log(8ab) - 1$, hence $a(\log b + 2\log n_0 + 1) \le a(\log(8ab) + 4\log(8ab) - 1) \le 5a\log(8ab) < n_0$. □

A.6 PROOF OF COROLLARY 1

Conditional on a pool satisfying the stated conditions, Theorem 2 certifies $\tau^\circ_g(\Delta)$ at every time with $n_g(t) \ge n^\star$, on an event of probability at least $1 - \delta''$. On that event, the chosen threshold is no larger than this certified candidate and hence has at least its coverage. Inverting the two terms in equation 8 gives the stated sufficient-margin scaling. □

A.7 PROOF OF THEOREM 3

Fix the group, threshold and instance p_0, and drop the indices. Let $\Delta := \alpha - R(p_0) > 0$ and $v(x) := p_0(x)(1 - p_0(x))$. Write $\pi(x) := P_g(X = x)$, $\mathcal{A} := \{x : A_\tau(x) = 1,\ \pi(x) > 0\}$, $c = \pi(\mathcal{A})$, $\bar{p}_0 := \sup_{x\in\mathcal{A}} p_0(x) < 1/2$, $v_{\max} := \sup_{x\in\mathcal{A}} v(x) \le 1/4$, $\mathbb{E}[vA] := \sum_{x\in\mathcal{A}} \pi(x)v(x) > 0$ and

$\mathbb{E}[\sqrt{v}A] := \sum_{x\in\mathcal{A}} \pi(x)\sqrt{v(x)} > 0$. Zero-variance terms in the divergence bounds below contribute zero. Define

$$\Delta_0 := \frac{(1/2 - \bar{p}_0)\,(\mathbb{E}[\sqrt{v}A])^2}{2\,c\,v_{\max}}.$$

An instance is a function $p : \mathcal{X} \to [0,1]$; under instance p the loss of an input x is Bernoulli$(p(x))$ and the risk is $R(p) = \sum_{x\in\mathcal{A}} \pi(x)p(x)/c$. We use the Bernoulli divergence bound $\mathrm{kl}(p \,\|\, p') \le (p-p')^2/(p'(1-p'))$ (the χ^2 bound) and the fact that $p \mapsto p(1-p)$ is increasing on $[0, 1/2]$.

(a) Passive. Let $\theta := 2\Delta c/\mathbb{E}[vA]$ and define the alternative $p_1 := p_0 + \theta v \mathbf{1}_{\mathcal{A}}$. Then $R(p_1) = R(p_0) + \theta\mathbb{E}[vA]/c = \alpha + \Delta > \alpha$. On $\mathcal{A}$, $\theta v(x) \le 2\Delta c v_{\max}/\mathbb{E}[vA] \le 2\Delta c v_{\max}/(\mathbb{E}[\sqrt{v}A])^2 \le 1/2 - \bar{p}_0$ for $\Delta \le \Delta_0$ (using $(\mathbb{E}[\sqrt{v}A])^2 \le c\,\mathbb{E}[vA] \le \mathbb{E}[vA]$ by Cauchy–Schwarz), so $p_0 \le p_1 \le 1/2$ on $\mathcal{A}$ and $p_1(1-p_1) \ge v$. Let P_0, P_1 be the laws of everything the certifier observes (n labeled pairs, any number of unlabeled inputs, internal randomness) under p_0, p_1. The marginal of X is π under both, so unlabeled inputs and internal randomness contribute nothing, and

$$\mathrm{KL}(P_0 \,\|\, P_1) = n \sum_x \pi(x)\,\mathrm{kl}\big(p_0(x) \,\|\, p_1(x)\big) \le n \sum_{x\in\mathcal{A}} \pi(x) \frac{\theta^2 v(x)^2}{v(x)} = n\theta^2\,\mathbb{E}[vA] = \frac{4n\Delta^2 c^2}{\mathbb{E}[vA]}.$$

Let $S = \{\text{the certifier does not certify } \tau\}$. Power at p_0 gives $P_0(S) \le \delta'$ and validity at p_1 gives $P_1(S^c) \le \delta$. By the Bretagnolle–Huber inequality (Tsybakov, 2009, Lemma 2.6), $P_0(S) + P_1(S^c) \ge \frac{1}{2}\exp(-\mathrm{KL}(P_0\|P_1))$, so $\mathrm{KL}(P_0\|P_1) \ge \log\frac{1}{2(\delta+\delta')}$, which is positive because $\delta + \delta' < 1/2$. Combining the two displays proves (a).

(b) Active. Now the certifier chooses X_i as a function of $(X_j, L_j)_{j<i}$ and internal randomness, and observes $L_i \sim$ Bernoulli$(p(X_i))$. Let $N_n(x)$ be the number of queries at x and $q(x) = \mathbb{E}_{p_0}[N_n(x)]/n$, assumed to satisfy $q \ge \kappa\pi$ on $\mathcal{A}$. The divergence decomposition for adaptively collected data (Kaufmann et al., 2016, Lemma 1) gives, for any alternative p_1, $\mathrm{KL}(P_0\|P_1) = \sum_x \mathbb{E}_{p_0}[N_n(x)]\,\mathrm{kl}(p_0(x)\|p_1(x)) = n\sum_x q(x)\,\mathrm{kl}(p_0(x)\|p_1(x))$. Let $S_q := \sum_{x\in\mathcal{A}} \pi(x)^2 v(x)/q(x)$ and define $p_1 := p_0 + \theta\,\pi v/q$ on $\mathcal{A}$ with $\theta := 2\Delta c/S_q$, so that again $R(p_1) = \alpha + \Delta$. By Cauchy–Schwarz, $(\mathbb{E}[\sqrt{v}A])^2 = \big(\sum_{\mathcal{A}} \frac{\pi\sqrt{v}}{\sqrt{q}}\sqrt{q}\big)^2 \le S_q \sum_{\mathcal{A}} q \le S_q$. Feasibility: on $\mathcal{A}$, $\theta\pi v/q \le 2\Delta c\,v_{\max}/(\kappa S_q) \le 2\Delta c v_{\max}/(\kappa(\mathbb{E}[\sqrt{v}A])^2) \le 1/2 - \bar{p}_0$ for $\Delta \le \kappa\Delta_0$, so $p_1(1-p_1) \ge v$ on $\mathcal{A}$ and

$$\mathrm{KL}(P_0\|P_1) \le n \sum_{x\in\mathcal{A}} q(x) \frac{\theta^2\pi(x)^2 v(x)^2/q(x)^2}{v(x)} = n\theta^2 S_q = \frac{4n\Delta^2c^2}{S_q} \le \frac{4n\Delta^2c^2}{(\mathbb{E}[\sqrt{v}A])^2}.$$

The Bretagnolle–Huber step is identical to (a). □

A.8 Proof of Theorem 4

(a) Fix the coefficients and proposals and condition on pools satisfying the stated coverage and slack conditions. By Appendix A.4 the labeled sequence of every group is an i.i.d. prefix, so Theorem 2 applies to each group's EB certificate with its own $(\sigma_g, c_g, \Delta_g)$: with conditional probability at least $1 - K\delta''$ (union bound), group g is certified at its target as soon as it has received $n_g^\star$ labels. Under round-robin over groups whose target is outside the EB certified set, a group stops receiving labels the moment its target is certified, so it receives at most $n_g^\star$ labels (at most $n_g^\star + b - 1$ if labels are bought in batches of size b), and the total is at most $\sum_g n_g^\star$ (plus $K(b-1)$). (b) Use the alternative from Theorem 3(a) for group g, changing only the conditional label law on its accepted region. Let d_g be the per-label divergence for a passive draw from group g. Other groups and unlabeled inputs contribute no divergence. For fixed counts the total divergence is $a_g B d_g$. For i.i.d. group draws it is $\mathbb{E}_{p_0}[n_g(B)]d_g = a_g B d_g$, by conditioning on the group sequence; the counts need not be deterministic. Certifying all targets with probability at least $1 - \delta'$ implies this power for group g, so the same Bretagnolle–Huber argument gives $a_g B \ge n_g^{\mathrm{LB}}$ for every g. □

A.9 Oracle importance proposal

A proposal attaining the active variance order. The following construction links the active lower bound to an admissible sampling design.

Proposition 1 (Importance proposals)**.** *For a fixed target τ, let $P_g = P_g$ and let $A = A_\tau$. Assume $\mathbb{E}_g[\sqrt{v(X)}A(X)] > 0$. Let $\tilde{L} = p$ be the conditional-risk imputer, $\lambda = 1$, and define the full-support defensive mixture*

$$\mathrm{d}Q_{g,\tau}(x) = \left[(1-\beta) + \beta \frac{\sqrt{v(x)}A(x)}{\mathbb{E}_g[\sqrt{v(X)}A(X)]}\right] \mathrm{d}P_g(x), \qquad \beta \in (0,1).$$

The first component gives full support, while the second concentrates labels where the conditional residual variance is large. The density ratio $w = dP_g/dQ_{g,\tau}$ satisfies $w \le 1/(1-\beta)$ everywhere and $w \le \mathbb{E}_g[\sqrt{v}A]/(\beta\sqrt{v(X)})$ when $A = 1$ and $v(X) > 0$. Consequently,

$$\sigma^2 = \mathbb{E}_{Q_{g,\tau}}[w^2 vA] \le \frac{\left(\mathbb{E}_g[\sqrt{v}A]\right)^2}{\beta}.$$

With $\beta = 1/2$, the variance term has the order of the active lower bound in Theorem 3(b), up to constants and logarithms, when v is known. This is a variance-order comparison in the conditional-mean-oracle and variance-dominated regime; it does not imply universal optimality, and the range term still uses $W = 1/(1-\beta)$.

The mixture exposes the trade-off in Theorem 2: increasing β strengthens the variance bound but enlarges the range bound $W = 1/(1-\beta)$. Concentrating queries can therefore reduce the variance component of certification cost while increasing its bounded-range component. With unknown v, a predictable proxy may define a proposal, provided its density ratio is known and bounded; the simultaneous validity theorem continues to apply.

Proof of Proposition 1.

Proof. The density of $Q_{g,\tau}$ relative to P_g is $(1-\beta) + \beta\sqrt{v(X)}A(X)/\mathbb{E}_g[\sqrt{v}A]$. Hence on $A = 1$ with $v(X) > 0$, w is bounded by both $1/(1-\beta)$ and $\mathbb{E}_g[\sqrt{v}A]/(\beta\sqrt{v(X)})$. With $\tilde{L} = p$ and $\lambda = 1$, the residual is $(L - p(X))A(X)$ and has conditional mean zero. Its conditional second moment is $\mathbb{E}[(Z_\tau - \tilde{Z}_\tau)^2 \mid X] = v(X)A(X)$, and therefore

$$\mathbb{E}_{Q_{g,\tau}}[w^2 vA] = \mathbb{E}_g[wvA] \le \frac{\mathbb{E}_g[\sqrt{v}A]}{\beta}\,\mathbb{E}_g[\sqrt{v}A].$$

□

B Algorithmic details

Algorithm 1 appears in Section 3. This appendix gives the confidence-bound constructions and implementation conditions.

B.1 Calibration quantities

After n labels of group g, let $a_i = (L_i - \alpha)A_\tau(X_i)$ and $b_i = (\tilde{L}_i - \alpha)A_\tau(X_i)$. A predictable regression coefficient is

$$\lambda_{g,n+1}(\tau) = \mathrm{clip}_{[0,1]}\left(\frac{\widehat{\mathrm{Cov}}(wa, wb)}{\widehat{\mathrm{Var}}(wb)}\right), \tag{9}$$

with a fixed convention when the denominator vanishes. Both this convention and a warm-up coefficient preserve predictability; Appendix D gives the experimental choices. Under passive sampling $w_i = 1$. Under adaptive sampling this regression, or an unweighted historical regression, remains predictable; the variance-minimizing interpretation applies to a fixed proposal.

The pool bound and coverage estimates are specified in Section 3. The two confidence-sequence constructions follow.

B.2 Empirical-Bernstein confidence sequence

For fixed coefficients and proposals, let $D_1, D_2, \ldots \in [-W, W]$ denote the i.i.d. residual increments of a group. Set $\delta' = \delta_L/(KM)$ and $\delta_k = \delta'/(k(k+1))$ for $k \geq 2$. Applying the fixed-sample bound of Maurer & Pontil (2009, Theorem 4) at every k gives

$$u^{\mathrm{EB}}_{g,k}(\tau) = \lambda\bar{\nu}_g(\tau) + \bar{D}_k + \sqrt{\frac{2\hat{V}_k \log(2/\delta_k)}{k}} + \frac{14W \log(2/\delta_k)}{3(k-1)}.$$

Since $\sum_{k\geq 2} \delta_k \leq \delta'$, these bounds hold simultaneously; take their running minimum, with $+\infty$ before the second label. This construction supplies the explicit width used in Theorem 2. Its i.i.d. argument applies to fixed coefficients and proposals. The betting construction below accommodates predictable updates.

B.3 Betting confidence sequence

For each group and threshold, let $\mathcal{M} = \{m_1, m_1 + h, \ldots, m_J\}$ be a fixed grid of candidate means. Capital processes start at one and use equation 5. For $m < b_t$, the current bet lies in $[0, c_{\max}/(b_t - m)]$ for a fixed $c_{\max} \in (0, 1)$; for $m \geq b_t$ it is zero. This makes every factor strictly positive. The choices of bet and coefficient precede the current label.

An empirical-Kelly rule chooses the bet to maximize historical log capital,

$$\sum_{i<t:\, g_i=g} \log\Big(1 - \beta\big(\hat{Z}_i(\tau) - m\big)\Big),$$

over the current admissible interval, additionally restricting the search to positive historical factors (Waudby-Smith & Ramdas, 2024). At initialization, with no labeled history, a zero bet is admissible. A histogram approximation or periodic refitting reduces computation without changing the validity argument. Between refits, cached bets must be clipped at every round to the current admissible interval and set to zero when $m \geq b_t$: a changing control-variate coefficient changes the current upper endpoint. This clipping is a condition of the stated algorithm.

For each time t, form the largest unrejected candidate mean plus h. Use $\sup \emptyset = m_1 - h$ so that complete rejection returns m_1, and return $+\infty$ whenever m_J is unrejected. Take the running minimum over time as in equation 6. These endpoint conventions preserve coverage even when the true mean lies outside the chosen grid; a poorly placed grid can reduce power without invalidating the guarantee.

B.4 Allocation and stopping rules

Egalitarian allocation prioritizes the group with the smallest estimated certified coverage $\hat{C}_g(\hat{\tau}_g)$, assigning zero coverage to groups with no certified candidate. Ties may be broken toward the group with fewer labels. A fixed exploration probability can mix this rule with uniform sampling; both choices are predictable and satisfy the validity theorem.

Round-robin allocation in Theorem 4 removes a group when its specified target belongs to the EB certified set. The empirical coverage-target variant instead removes a group once $\hat{\tau}_g \leq \tau^\circ_g$, so a certified operating point reaches the target coverage. *Proportional* and *uniform* allocation track fixed group shares π_g and $1/K$, respectively. A population-coverage policy can instead prioritize the largest estimated weighted coverage gain per label. These rules share the certificate but pursue different coverage objectives.

B.5 Extensions and sampling scope

Shared threshold. A candidate is eligible when $U_{g,t}(\tau) \leq 0$ for every group. Returning the smallest eligible threshold maximizes coverage among eligible candidates because their acceptance regions are nested.

Predictable imputer selection. For a fixed finite family $\{\tilde{\ell}^{(1)}, \ldots, \tilde{\ell}^{(J)}\}$, compute a pool bound for every member at level $\delta_U/(KMJ)$ and select the member predictably. A union bound then gives the

same simultaneous-validity argument. Refitting an unrestricted imputer on the certification labels requires a separate construction.

The population guarantees use fresh draws, with known importance weights when the proposal changes. They do not cover sampling without replacement from the certification pool without an additional finite-population argument. The fixed-coefficient complexity bounds further require proposals fixed in advance and labeled draws independent of the pool.

C Evaluation and additional results

Fixed-budget metrics. For each run r, we compute $C_{\min,r} = \min_g C_g(\hat{\tau}_{g,r})$, $C_{\text{pop},r} = \sum_g \pi_g C_g(\hat{\tau}_{g,r})$ and $R_{\max,r} = \max_g R_g(\hat{\tau}_{g,r})$. All three quantities use the full fixed reservoir. Group extrema are taken within a run before averaging. A fully abstaining group has zero coverage and risk. The violation count is $\sum_r \mathbf{1}\{R_{\max,r} > \alpha\}$ at the stated budget. LTT-CP is evaluated only as a fixed-budget comparator; the paper makes no optional-stopping guarantee for that baseline.

We report the standard error across independent calibration seeds. The paired interval for the imputation comparison is a Student-t interval for the seed-matched difference, with 29 degrees of freedom. These intervals describe calibration randomness at the specified population, rather than variation across independently generated datasets.

Coverage-target attainment and censoring. For targets fixed before calibration, define

$$T_r = \inf\{t : \hat{\tau}_{g,r}(t) \le \tau_g^\circ \text{ for every } g\}. \tag{10}$$

Because acceptance regions are nested, this condition means that every group has a certified operating point with at least its target coverage. The target candidate itself need not be a member of its certified set. Recorded group-local first-passage counts and the deterministic deficit allocator recover the global attainment time for static allocation. For stopping allocation, the final spent-label count is the attainment time. Both are measured at batch boundaries. A run unfinished at $B_{\max} = 16{,}000$ contributes $(B_{\max}, 0)$; a completed run contributes $(T_r, 1)$. The Kaplan–Meier median is the first time the estimated unfinished probability is at most one half. All 16 runs contribute, and median reductions compare method-specific medians.

Table 3: Additional controlled populations. Entries are percentages; coverage is mean $\pm$ one standard error. $R_{\max}$ is the mean worst-group risk. All rows have zero fixed-budget violations. Eight-group results use 12 seeds; results at $\alpha = 0.10$ use 16.

Method	Worst-group C	Population C	$R_{\max}$
Eight groups, $\alpha = 0.05$, $B = 12{,}000$			
LTT-CP, proportional	0.00 ± 0.00	77.16 ± 0.45	3.56
SPARC, auxiliary, proportional	0.00 ± 0.00	75.97 ± 0.30	3.57
SPARC, auxiliary, egalitarian	59.51 ± 0.52	72.22 ± 0.74	3.79
Four groups, $\alpha = 0.10$, $B = 6000$			
LTT-CP, proportional	66.09 ± 1.95	92.90 ± 0.20	7.44
SPARC, auxiliary, proportional	56.49 ± 5.68	93.46 ± 0.30	7.79
SPARC, auxiliary, egalitarian	83.95 ± 0.38	90.50 ± 0.28	8.53

The eight-group setting reduces the smallest group share to 1.5%, intensifying the evidence bottleneck under proportional allocation. At the higher risk target, all three procedures attain greater coverage. Across both settings, egalitarian allocation raises the minimum group coverage while allocating less evidence to population-dominant groups, so population coverage must be assessed separately.

D Controlled benchmark and implementation

Finite population. The benchmark fixes a reservoir of 10^6 records per group, generated with seed 0. For each record in group g, draw $H \sim \text{Bernoulli}(\zeta_g)$, $p \mid H = 0 \sim \text{Beta}(1, 150)$, $p \mid H = 1 \sim \text{Beta}(2, 2)$, and $\eta \sim \mathcal{N}(0, 1)$ independently conditional on H. The confidence score

and realized binary loss are

$$s = \mathrm{sigmoid}(\mathrm{logit}(1-p) + b_g + \kappa_g \eta)\,, \qquad L \mid p \sim \mathrm{Bernoulli}(p).$$

The record's score and realized loss are fixed thereafter. Within each group, the uniform distribution over this reservoir is the evaluation population; labels are not redrawn on repeated visits to a record. The group parameters are

g	1	2	3	4
π_g	0.50	0.30	0.15	0.05
ζ_g	0.15	0.20	0.25	0.30
b_g	0	0.5	1.0	1.5
κ_g	2.0	2.2	2.4	2.6

Thus population imbalance, error prevalence and confidence quality vary across groups. The eight-group experiment replaces each group by two groups with shares $0.7\pi_g$ and $0.3\pi_g$. The larger subgroup retains its parameters; the smaller uses $(\zeta_g + 0.03, b_g + 0.2, \kappa_g + 0.1)$, with 5×10^5 records per subgroup.

Pre-query auxiliary signal. Each record also receives one independent $u \sim \mathrm{Uniform}(0,1)$ at reservoir construction. The controlled auxiliary signal is

$$\tilde{\ell} = \begin{cases} u^{1/a_g}, & L = 1, \\ 1 - u^{1/a_g}, & L = 0. \end{cases}$$

Consequently, $\tilde{\ell} \mid L = 1 \sim \mathrm{Beta}(a_g, 1)$ and $\tilde{\ell} \mid L = 0 \sim \mathrm{Beta}(1, a_g)$. We set $a_g = 10$ in the four-group experiments and use $a_g = 10$ and 9 in each larger/smaller subgroup pair. The signal is exposed before the costly query of L, remains fixed for each record, and is correlated with the realized loss by construction. It represents a controlled side-information channel, not the output of a trained judge. The confidence-only signal is $\tilde{\ell} = 1 - s$; the constant signal is $\tilde{\ell} = \alpha$. All available pre-query signals are part of X in the information model of Section 2. The latent p specifies reservoir generation and is not interpreted as the conditional error probability given this enlarged information set.

Thresholds, pools and sampling. A separate score sample, with $\lfloor 200{,}000\pi_g \rfloor$ independent draws with replacement from each group and seed 12345, fixes $M = 25$ thresholds. They are empirical quantiles of this mixture score sample at equally spaced probability levels from 0.02 to 0.70, ordered decreasingly. Each trial then draws 100,000 records per group with replacement and retains those with $s \geq \tau_M$ for the imputed-loss pool. Thus 100,000 is the pool size *before* conditioning; the retained count is the N_g used in the pool bound. Conditional on this retained count, the pool records are i.i.d. from P_g, so the pool bound applies with that N_g. Queried records are independent draws with replacement from the candidate-region reservoir, independent of the pool draws. The fixed-budget and allocation comparisons use passive within-group sampling, $w = 1$. An independently sampled record may occur in both samples because the sampling model permits replacement.

Certificate implementation. Prediction-powered trials use $\delta = 0.1$, split as $\delta_U = 0.01$ and $\delta_L = 0.09$, with Bonferroni allocation across the KM group–threshold pairs. The pool bound is empirical Bernstein. The sequential certificate uses a histogram approximation to empirical-Kelly betting (Waudby-Smith & Ramdas, 2024): 40 bins for the residual after mapping $[-W, W]$ to $[0, 1]$, with one pseudo-observation at each endpoint used only to choose bets. Bet fitting is refreshed every 20 group labels, using three projected Newton steps per refresh and eight at initialization. The cap is $c_{\max} = 0.9$; cached bets are clipped to the current admissible interval at every update. We use 1501 candidate means spanning $[-0.5, 0.5]$ on the centered-loss scale, with the grid padding and endpoint conventions of Section 3. Each capital factor uses the observed increment; histogram binning affects bet selection only.

The adaptive control-variate coefficient starts at 1 for the first 30 group labels. Thereafter it is the clipped empirical regression coefficient of Z_τ on $\tilde{Z}_\tau$, computed from preceding labels; the implementation uses 1 when the empirical variance is at most 10^{-12}. This convention is admissible for predictable coefficients. The explicit labeled-only comparison fixes $\lambda = 0$, uses $\delta_L = 0.1$ and $\delta_U = 0$, and sets the unused pool offset to zero; the constant-signal implementation with its pool

bound is a separate configuration. Labels are acquired in batches of 10, with a sequential coefficient and capital update for each label inside the batch. Egalitarian allocation uses uniform exploration among eligible groups with probability 0.1 and otherwise chooses the group with the lowest estimated certified coverage, breaking ties by label count. Static allocation tracks the desired shares using their cumulative label deficits.

Repetitions and evaluation. The four-group experiment uses $\alpha = 0.05$, $B = 8000$ and seeds 1000–1029; the eight-group experiment uses $\alpha = 0.05$, $B = 12{,}000$ and seeds 4000–4011; the higher-risk experiment uses $\alpha = 0.10$, $B = 6000$ and seeds 5000–5015. The reservoir and threshold grid stay fixed across repetitions. The trial's unlabeled-pool coverage estimates guide allocation. Reported coverage and selective risk are computed exactly from the returned thresholds and the accepted reservoir records, using their fixed losses. Group extrema are formed within a run before aggregation; population coverage uses the specified π_g.

Coverage-target allocation experiment. The allocation comparison uses seeds 3000–3015, $\alpha = 0.05$ and $B_{\max} = 16{,}000$. Before calibration, the benchmark fixes each target three grid positions more conservative than the largest index with reservoir risk at most α, truncated at the first candidate. The resulting target indices are $(20, 18, 16, 12)$ under one-based indexing. A group reaches its target when its returned certified index is at least the target index: it has certified an operating point with at least the target coverage. This criterion does not require the particular target candidate itself to belong to the certified set. Round-robin and egalitarian allocation remove groups once this criterion is met; static rules retain their fixed shares. Completion cost is the first 10-label batch boundary at which all groups reach their targets. Adaptive rules record this boundary when they stop. For static rules, it is recovered from each group's recorded local first-passage label count and the deterministic deficit-allocation schedule. Unfinished runs are right-censored at $B_{\max}$. We report Kaplan–Meier medians, defined as the first time the estimated survival probability is at most one half; unfinished runs remain in the analysis.