SHARP SELECTIVE-RISK ENVELOPES UNDER NORMALIZED DENSITY-RATIO SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Selective prediction controls error by abstaining on uncertain inputs, but an operating point calibrated on source data can lose its guarantee under distribution shift. We characterize the exact worst-case selective risk under a normalized joint density-ratio bound, $1/\Gamma \leq dQ/dP \leq \Gamma$. For binary loss, it depends only on the source coverage and the accepted error rate. The familiar Γ^2 error-odds envelope is sharp precisely when rejected examples can absorb the normalization imbalance; at high coverage, it can discard valid operating points. We derive a closed-form, coverage-dependent error target and use it in HOLM–SHARP-RC, which combines independent score-only coverage estimation with exact binomial tests and Holm correction. Its certificate remains valid after threshold selection, its certified set contains those of the matched odds-envelope and Bonferroni rules, and it recovers every threshold that clears an explicit estimation margin. Independent linear programs match the population formulas to within 10^{-12} on 2,220 instances. In a controlled numerical study with fully specified score–loss distributions, the matched sharp and odds rules coincide in a low-coverage regime, whereas in a high-coverage regime the sharp rule raises mean worst-case coverage from 0.435 to 0.999 with no observed risk violations. These results identify when normalization materially changes calibration and when the simpler odds envelope suffices.

1 INTRODUCTION

A predictor that abstains on unreliable inputs can be more useful than one that always responds. Selective prediction formalizes this tradeoff through *coverage*, the probability of responding, and *selective risk*, the expected loss given a response (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). A confidence score ranks examples, and calibration chooses an operating point along that ranking. Finite-sample procedures can control the source selective risk of the chosen operating point, for example, by multiple testing over candidate thresholds (Angelopoulos et al., 2025). Such guarantees, however, concern the source distribution, whereas deployment may change the relative mass of accepted errors, accepted correct predictions, and rejected examples.

The difficulty is that selective risk is a ratio. A shift can increase the mass of accepted errors, decrease the mass of accepted correct predictions, and change the rejected mass that keeps the distribution normalized. Bounding the first two effects independently yields an appealing rule: if the density ratio lies in $[1/\Gamma, \Gamma]$, the error odds among accepted examples grow by at most a factor of Γ^2. The two extremal weights, however, need not belong to a single normalized deployment distribution. Consider a system that answers every input with source error 10%, and let $\Gamma = 2$. The odds rule reports a worst-case risk of about 0.31. Yet no rejected mass is available to restore normalization: doubling the error mass to 0.2 leaves 0.8 for the correct predictions, a weight of 0.89 rather than $1/2$, so the true worst case is 0.20. A target of $\alpha = 0.25$ is therefore certifiable, although the odds rule rejects it. At high coverage, the odds bound can thus discard certifiable operating points (Figure 1).

We therefore ask: *what risk and coverage can a fixed score guarantee under an unknown reweighting of the source population from a specified family?* We answer this question for normalized, two-sided density-ratio bounds. Because the predictor and score are fixed, the procedure applies to classifiers, verifiers, and generative models alike.

Our contributions are as follows.

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.

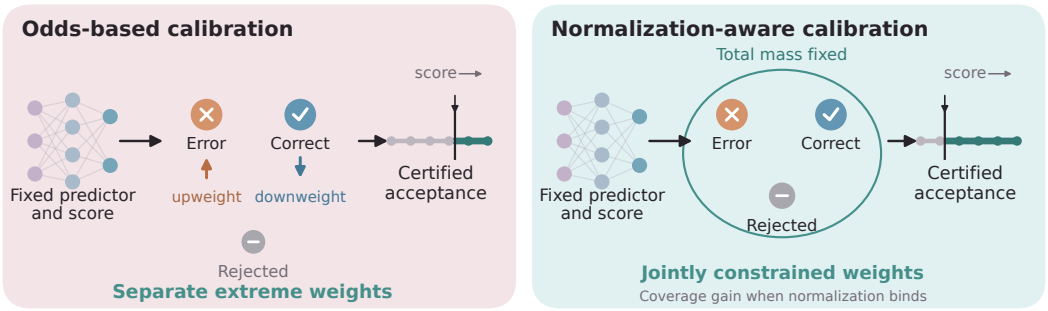


Figure 1: **Normalization couples accepted and rejected outcomes.** The odds bound upweights accepted errors and downweights accepted correct predictions separately. The sharp envelope subjects all three outcome masses to a shared normalization constraint. With the same predictor, score, calibration data, and testing correction, its coverage-dependent target can certify a larger acceptance region when normalization binds. Score ranges are schematic.

- **Sharp population envelope.** We derive the exact worst-case binary selective risk with an attainability proof, decompose it into one odds regime and two normalization-limited regimes, and invert it in closed form into a coverage-dependent error target (Theorem 1, Proposition 1, and Corollary 1). This isolates exactly what is lost by bounding the two accepted outcomes independently.
- **Post-selection calibration.** We give a budget-uniform certificate that holds simultaneously for all Γ (Theorem 3) and the more powerful, target-specific HOLM–SHARP-RC (Theorem 4). The latter is familywise valid, deterministically dominates the matched odds-envelope and Bonferroni rules, and recovers every threshold that clears its sharp target by an explicit margin (Proposition 3).
- **Mechanism-focused validation.** Independent optimization checks the population formulas. Controlled numerical experiments compare the sharp and odds envelopes on identical data, coverage bounds, and testing corrections, and include a regime in which the sharper envelope yields no gain.

The technical contribution lies in the normalized conditional-risk functional and its coverage-dependent calibration target; the confidence bounds and multiple-testing tools are standard ingredients. An upper-tail characterization extends the population result to bounded losses, and the appendix treats contamination and further statistical details.

2 Related Work

Selective prediction and risk control. Selective classification studies how conditional error varies with acceptance (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017); work on selective question answering documents how domain shift affects source-calibrated abstention (Kamath et al., 2020; Liang et al., 2024). Learn then Test provides finite-sample risk control after selection from a finite family (Angelopoulos et al., 2025); our calibration applies this testing framework to a coverage-dependent robust target. Selective Conformal Risk Control jointly calibrates selection and prediction-set thresholds (Xu et al., 2025). SCoRE controls bounded selective deployment risks with e-values, including weighted extensions under covariate shift (Bai & Jin, 2026); its guarantee concerns an expected batch-level selective risk, whereas ours is a high-probability bound on population risk.

Distribution shift and sensitivity analysis. Weighted conformal prediction and PAC prediction sets transfer guarantees using known or uncertain importance weights (Tibshirani et al., 2019; Park et al., 2022). Robust validation protects predictive coverage over divergence neighborhoods (Cauchois et al., 2024), and likelihood-ratio sensitivity analysis gives sharp bounds on event probabilities (Jin et al., 2023); our coverage formula is a constant-bound special case of the latter. Selective risk additionally requires the adverse masses of accepted errors and accepted correct predictions to be jointly attainable after normalization. The rejected cell, which does not enter the risk ratio, supplies

the remaining mass and produces the three regimes in Equation (9). Retaining such compatibility constraints is likewise central to sharp sensitivity bounds for inverse propensity weighting (Dorn & Guo, 2023).

Score construction and calibration. Likelihood-ratio scores can improve selector ranking under covariate shift by exploiting training labels and representation densities (Heng & Soh, 2026). We instead hold the ranking fixed and characterize the operating points it can certify. Accordingly, our main comparator changes only the risk envelope and keeps the score, samples, confidence bounds, and testing correction fixed, which separates the effect of normalization from that of ranking quality.

3 Problem and Scope

Let $Z = (X, Y) \sim P$ be a labeled source example, f a fixed predictor, s a fixed confidence score, and $\ell(Z) = \ell(X, Y, f(X)) \in \{0, 1\}$ the error indicator. Larger scores indicate greater confidence. The deployed selector is $a_\tau(X) = \mathbb{I}\{s(X) \geq \tau\}$; it does not observe Y. For randomized generation or scoring, we include the auxiliary randomness in X and fix its sampling procedure before calibration, so that all probability statements concern the induced joint distribution of score and loss.

We consider normalized reweightings of the *joint* observation:

$$\mathcal{Q}_\Gamma(P) = \{Q \ll P : w = dQ/dP, \quad 1/\Gamma \leq w \leq \Gamma \; P\text{-a.s.}, \quad \mathbb{E}_P w = 1\}, \quad \Gamma \geq 1. \tag{1}$$

Because the ratio is defined on the joint space, the reweighting may depend on whether a prediction is correct. Covariate shift with the same bound on the input density ratio and unchanged label and inference kernels is a subfamily of Equation (1): our bounds remain valid there but need not be attained.

For a selector a, write $p = P(a = 1)$ and, when $p > 0$, $q = P(\ell = 1 \mid a = 1)$. For threshold selectors, we write $p(\tau)$ and $q(\tau)$, and on a grid $p_k = p(\tau_k)$ and $q_k = q(\tau_k)$. Define

$$C_Q(a) = \mathbb{E}_Q a, \quad R_Q(a) = \frac{\mathbb{E}_Q[a\ell]}{\mathbb{E}_Q a}, \quad C_\Gamma(a) = \inf_{Q \in \mathcal{Q}_\Gamma(P)} C_Q(a), \quad R_\Gamma(a) = \sup_{Q \in \mathcal{Q}_\Gamma(P)} R_Q(a). \tag{2}$$

For $p > 0$, every feasible deployment distribution has coverage at least $p/\Gamma > 0$, so R_Q is well defined. If $p = 0$, no answer is issued and the conditional risk is undefined; we then record coverage zero and no risk violation. The reject-all rule is always available as a fallback.

Definition 1 (Score-restricted robust frontier)**.** For a target $\alpha \in (0, 1)$, let

$$\mathcal{C}_{\Gamma,s}(\alpha) = \sup\big(\{C_\Gamma(a_\tau) : p(\tau) > 0, \; R_\Gamma(a_\tau) \leq \alpha\} \cup \{0\}\big). \tag{3}$$

This is a population optimum over the given threshold family, not over all measurable selectors, and it need not be attained. Increasing Γ enlarges the deployment family and therefore cannot increase the frontier. A finite grid and calibration uncertainty incur further coverage losses, which are separate from this population tradeoff.

4 Exact Risk and the Odds Envelope

Let $e = pq$ and $b = p(1 - q)$ denote the source masses of accepted errors and accepted correct predictions. For an event of source mass t, the normalized density-ratio bounds constrain its deployment mass to $[L_\Gamma(t), U_\Gamma(t)]$, where

$$L_\Gamma(t) = \max\{t/\Gamma, \; 1 - \Gamma(1 - t)\}, \qquad U_\Gamma(t) = \min\{\Gamma t, \; 1 - (1 - t)/\Gamma\}. \tag{4}$$

In each expression, the first term bounds the event directly and the second bounds it through its complement. These event-probability envelopes are special cases of existing likelihood-ratio sensitivity bounds (Jin et al., 2023). Selective risk requires an additional step: showing that the upper mass of accepted errors and the lower mass of accepted correct predictions are *jointly attainable*.

Theorem 1 (Exact normalized binary risk)**.** *For $p > 0$ and $q \in [0, 1]$,*

$$R_\Gamma(a) = \Psi_\Gamma(p, q) := \frac{U_\Gamma(pq)}{U_\Gamma(pq) + L_\Gamma(p(1 - q))}. \tag{5}$$

The function Ψ_Γ extends continuously to $p = 0$ by $\Psi_\Gamma(0, q) = \rho_\Gamma(q)$, where

$$\rho_\Gamma(q) = \frac{\Gamma^2 q}{\Gamma^2 q + 1 - q}. \tag{6}$$

For fixed Γ, Ψ_Γ is continuous on $[0, 1]^2$, nondecreasing in q, and nonincreasing in p.

Proof idea. Every feasible distribution assigns at most $U_\Gamma(e)$ to accepted errors and at least $L_\Gamma(b)$ to accepted correct predictions. The remaining mass $1 - U_\Gamma(e) - L_\Gamma(b)$ always lies in the feasible range $[(1 - p)/\Gamma, \Gamma(1 - p)]$ of the rejected cell, so a single distribution attains both extremes. Since $u/(u + v)$ increases in u and decreases in v, this proves Equation (5). The three-cell construction and the monotonicity proof are given in Section A.

Proposition 1 (Odds envelope and its tightness)**.** *For all $(p, q) \in [0, 1]^2$, $\Psi_\Gamma(p, q) \le \rho_\Gamma(q)$. For $0 < p < 1$ and $0 < q < 1$, equality holds if and only if the rejected-cell weight implied by the independent extremes is feasible:*

$$\frac{1 - \Gamma pq - p(1 - q)/\Gamma}{1 - p} \in [1/\Gamma, \Gamma]. \tag{7}$$

Moreover,

$$\rho_\Gamma(q) \le \alpha \quad \Longleftrightarrow \quad q \le q^*(\alpha, \Gamma) := \frac{\alpha}{\alpha + \Gamma^2(1 - \alpha)}. \tag{8}$$

Thus, $q \le q^*$ is sufficient for robust risk control, but it is necessary only in the regime where the envelope is tight.

The sharp envelope also admits an explicit regime form. For $\Gamma > 1$ and $p > 0$,

$$\Psi_\Gamma(p, q) = \begin{cases} \rho_\Gamma(q), & pq \le \frac{1}{\Gamma+1}, \quad p(1-q) \le \frac{\Gamma}{\Gamma+1}, \\ \dfrac{pq + \Gamma - 1}{p + \Gamma - 1}, & pq > \frac{1}{\Gamma+1}, \\ \dfrac{\Gamma pq}{\Gamma p - (\Gamma - 1)}, & p(1-q) > \frac{\Gamma}{\Gamma+1}. \end{cases} \tag{9}$$

The last two cases are mutually exclusive, and for $p < 1$ the condition of the first case is equivalent to Equation (7). In the *odds regime*, the adversary places weight Γ on accepted errors and $1/\Gamma$ on accepted correct predictions, and the rejected cell absorbs the imbalance. In the second case, accepted errors are so heavy that weight Γ on them would force the average weight on the rest of the population below $1/\Gamma$; the extremal distribution instead places weight $1/\Gamma$ everywhere else. In the third case, accepted correct predictions are so heavy that weight $1/\Gamma$ on them would force the average weight on the rest above Γ; the extremal distribution instead places weight Γ everywhere else. In both normalization-limited regimes, the odds envelope is strict for $q \in (0, 1)$. The third case is the most relevant in practice because it describes a high-coverage, low-error selector.

Because $\Psi_\Gamma \le \alpha$ compares a minimum and a maximum of affine functions of q, the sharp envelope can be inverted exactly.

Corollary 1 (Sharp error target)**.** *For $p \in (0, 1]$ and $\alpha \in (0, 1)$, $\Psi_\Gamma(p, q) \le \alpha$ if and only if $q \le q^\dagger(p)$, where*

$$q^\dagger(p) := \sup\{q : \Psi_\Gamma(p, q) \le \alpha\} = \max\left\{q^*,\ \alpha\Big(1 - \frac{\Gamma - 1}{\Gamma p}\Big),\ \alpha - \frac{(1 - \alpha)(\Gamma - 1)}{p}\right\}. \tag{10}$$

The target $q^\dagger$ is nondecreasing in p, satisfies $q^\dagger(p) \ge q^* =: q^\dagger(0)$, and equals $q^\dagger(1) = L_\Gamma(\alpha)$ at full acceptance.

Proof idea. Each branch of Equation (9) gives a linear inequality in q, and their union yields the maximum in Equation (10). The complete argument is in Section A.4.

In the high-coverage branch, robust control is thus equivalent to

$$q \le \alpha\left(1 - \frac{\Gamma - 1}{\Gamma p}\right), \qquad p(1 - q) > \frac{\Gamma}{\Gamma + 1}. \tag{11}$$

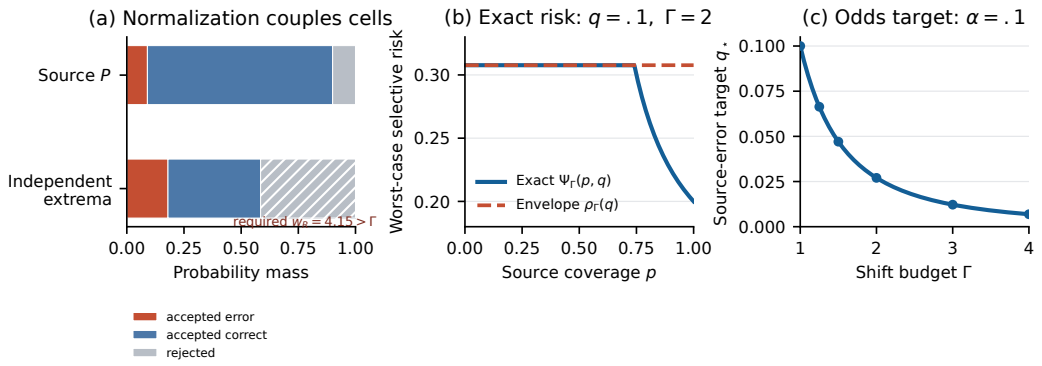


Figure 2: **Why normalization changes selective risk.** (a) With $p = 0.9$, $q = 0.1$, and $\Gamma = 2$, independent extreme weights on the two accepted outcomes would require a rejected-cell weight of 4.15, outside the admissible weight interval $[1/2, 2]$. (b) Enforcing normalization lowers the worst-case risk below the odds envelope at high coverage. (c) The coverage-free odds target becomes more stringent as Γ grows. All panels are analytic evaluations, not benchmark measurements.

At full acceptance and for $\alpha \le \Gamma/(\Gamma+1)$, the sharp target is $q \le \alpha/\Gamma$, whereas the coverage-free requirement $q \le q^*$ is close to $q \le \alpha/\Gamma^2$ for small α. In the introductory example ($p = 1$, $q = 0.10$, $\Gamma = 2$), the exact risk is 0.20, whereas the odds envelope is approximately 0.308; the target $\alpha = 0.25$ is therefore feasible even though $q > q^* \approx 0.0769$.

Proposition 2 (Worst-case coverage)**.** *For any selector with source coverage p, $C_\Gamma(a) = c_\Gamma(p) := L_\Gamma(p)$.*

Consequently, the exact score-restricted frontier is

$$\mathcal{C}_{\Gamma,s}(\alpha) = \sup\left(\{L_\Gamma(p(\tau)) :\ p(\tau) > 0,\ \Psi_\Gamma(p(\tau), q(\tau)) \le \alpha\} \cup \{0\}\right). \tag{12}$$

The worst-risk and worst-coverage distributions may differ: the two quantities are separate uniform guarantees rather than properties of a single adversarial distribution. Replacing Ψ_Γ by ρ_Γ in Equation (12) gives a conservative inner frontier.

Extension to bounded losses. For binary loss, the three outcome masses determine the exact risk. A bounded loss also requires the distribution of the loss within the accepted region, which enters through an upper-tail functional. For an integrable random variable H and $\kappa \in [0, 1]$, define

$$\mathsf{T}_\kappa(H) = \sup\{\mathbb{E}_P[\xi H] : 0 \le \xi \le 1,\ \mathbb{E}_P \xi = \kappa\}. \tag{13}$$

This is the integral of the quantile function of H over its upper κ fraction, with fractional allocation at a tie when needed.

Theorem 2 (Sharp bounded-loss risk)**.** *Let $\ell \in [0, 1]$ and $p = P(a = 1) > 0$. For $\Gamma > 1$, $R_\Gamma(a)$ is the unique $r \in [0, 1]$ satisfying*

$$\frac{1}{\Gamma}\mathbb{E}_P[H_r] + \left(\Gamma - \frac{1}{\Gamma}\right)\mathsf{T}_{1/(\Gamma+1)}(H_r) = 0, \qquad H_r = a(\ell - r). \tag{14}$$

For $\Gamma = 1$, the root is $\mathbb{E}_P[a\ell]/p$. When ℓ is binary, Equation (14) *reduces to Equation* (5)*.*

The extremal reweighting places more mass on the upper tail of $a(\ell - r)$, with rejected examples assigned value zero. Writing $w = \Gamma^{-1} + (\Gamma - \Gamma^{-1})\xi$ makes normalization equivalent to $\mathbb{E}_P \xi = 1/(\Gamma + 1)$, and maximizing $\mathbb{E}_P[wa(\ell - r)]$ gives the left-hand side of Equation (14); its root is the worst-case conditional mean. On a finite population, the tail functional is computed by sorting, and a scalar root search recovers the exact risk. The proof is in Section A.5, and a mean-based finite-sample certificate is given in Section C.

5 FINITE-SAMPLE CALIBRATION

Let $\tau_1 < \cdots < \tau_K$ be a grid of distinct thresholds fixed independently of the calibration data. A separate design sample may determine the score and the grid; all guarantees are conditional on that design.

A budget-uniform certificate. Let $D = \{(X_i, \ell_i)\}_{i=1}^n$ be i.i.d. from P. At threshold τ_k, let m_k and x_k be the numbers of accepted examples and accepted errors, respectively. The one-sided Clopper–Pearson upper limit (Clopper & Pearson, 1934) with failure probability η is

$$\bar{q}(x, m; \eta) = \begin{cases} 1, & m = 0 \text{ or } x = m, \\ \mathrm{Beta}^{-1}(1-\eta; x+1, m-x), & 0 \le x < m. \end{cases} \tag{15}$$

Define

$$\bar{q}_k = \bar{q}(x_k, m_k; \delta/(2K)), \qquad \underline{p}_k = \max\left\{0, \frac{m_k}{n} - \sqrt{\frac{\log(4/\delta)}{2n}}\right\}. \tag{16}$$

The coverage bound uses the two-sided Dvoretzky–Kiefer–Wolfowitz (DKW) inequality (Massart, 1990), which is uniform over thresholds and therefore free of K; clipping at zero keeps the first argument of Ψ_Γ in its domain when few examples are accepted.

Theorem 3 (Uniform post-selection certificate)**.** *With probability at least $1-\delta$ over D, simultaneously for every grid threshold with positive population coverage and every $\Gamma \ge 1$,*

$$R_\Gamma(a_{\tau_k}) \le \overline{R}_k := \Psi_\Gamma(\underline{p}_k, \bar{q}_k), \qquad C_\Gamma(a_{\tau_k}) \ge \underline{C}_k := L_\Gamma(\underline{p}_k). \tag{17}$$

Any rule that selects among the thresholds with $\overline{R}_k \le \alpha$ therefore controls the robust selective risk at level α on this event.

Proof idea. Conditional on m_k, the count x_k is binomial with success probability q_k. A union bound over the K Clopper–Pearson limits, each at level $\delta/(2K)$, spends $\delta/2$, and DKW spends the remaining $\delta/2$ on the event that $p_k \ge \underline{p}_k$ for all k. Monotonicity of Ψ_Γ and L_Γ completes the argument. Because the event involves only source probabilities, it supports all budgets Γ simultaneously; it does not, however, establish that the actual deployment distribution satisfies any chosen budget.

Target-specific step-down calibration. The preceding construction pays a Bonferroni factor of K and estimates coverage from the labeled pairs. Step-down testing can reduce the first cost, and coverage can instead be estimated from an independent score-only source sample $U = \{X_j^u\}_{j=1}^N$ when one is available. Such a sample requires scores but no correctness labels, and its size N is accounted for separately from the labeled size n. Let r_k be its accepted count at τ_k and define

$$\widetilde{p}_k = \max\left\{0, \frac{r_k}{N} - \sqrt{\frac{\log(4/\delta)}{2N}}\right\}, \qquad q_k^\dagger = \sup\{q \in [0,1] : \Psi_\Gamma(\widetilde{p}_k, q) \le \alpha\}. \tag{18}$$

By Corollary 1, $q_k^\dagger = q^\dagger(\widetilde{p}_k)$ is available in closed form; monotone bisection returns the same value up to numerical tolerance. On the labeled sample D, set $\pi_k = 1$ when $m_k = 0$, and otherwise

$$\pi_k = \mathbb{P}\{B \le x_k\}, \qquad B \sim \mathrm{Bin}(m_k, q_k^\dagger). \tag{19}$$

This is the exact binomial p-value for the one-sided null $H_k : q_k \ge q_k^\dagger$. Order the p-values as $\pi_{(1)} \le \cdots \le \pi_{(K)}$ and define the Holm-adjusted values (Holm, 1979)

$$\widetilde{\pi}_{(i)} = \min\left\{1, \max_{j \le i}(K-j+1)\pi_{(j)}\right\}, \tag{20}$$

writing $\widetilde{\pi}_k$ for the adjusted value of threshold τ_k in the original indexing.

Theorem 4 (Holm-optimized post-selection certificate)**.** *Let U and D be independent i.i.d. source samples, and let the grid be fixed independently of both. With probability at least $1-\delta$, every threshold with $m_k > 0$ and $\widetilde{\pi}_k \le \delta/2$ satisfies*

$$R_\Gamma(a_{\tau_k}) \le \alpha, \qquad C_\Gamma(a_{\tau_k}) \ge L_\Gamma(\widetilde{p}_k). \tag{21}$$

Consequently, any data-dependent choice among these thresholds is valid. Moreover, for fixed coverage bounds and labeled counts, the certified set contains (i) the Holm-corrected odds-envelope set, obtained by replacing $q_k^\dagger$ with q^*, and (ii) every threshold that passes the single-step sharp test $\pi_k \le \delta/(2K)$.

Algorithm 1 Holm-optimized sharp calibration (HOLM–SHARP-RC)

Input: fixed predictor and score, independent score-only sample U, labeled calibration sample D, independent grid of distinct thresholds, and (α, Γ, δ).

1. Sort both samples once and compute (r_k, m_k, x_k) for all thresholds by prefix sums.
2. Compute $(\widetilde{p}_k, q_k^\dagger, \pi_k)$ from Equations (18) and (19).
3. Apply Equation (20) and let $\mathcal{I} = \{k : m_k > 0,\ \widetilde{\pi}_k \le \delta/2\}$.
4. If $\mathcal{I}$ is empty, abstain everywhere. Otherwise, return the member of $\mathcal{I}$ maximizing $L_\Gamma(\widetilde{p}_k)$, breaking ties toward greater acceptance.

Proof idea. DKW gives $p_k \ge \widetilde{p}_k$ for all k with failure probability at most $\delta/2$. Conditional on the independent sample U, each target $q_k^\dagger$ is fixed. For each k, $x_k \mid m_k, U \sim \mathrm{Bin}(m_k, q_k)$; averaging over m_k yields a valid p-value for H_k conditional on U. Holm's procedure controls the probability of any false rejection over the grid at level $\delta/2$ under arbitrary dependence, which covers the nested counts x_k. On the DKW event, $R_\Gamma(a_{\tau_k}) > \alpha$ implies $\Psi_\Gamma(\widetilde{p}_k, q_k) > \alpha$ and hence $q_k > q_k^\dagger$ by monotonicity, so every violating threshold is a true null. The containment statements follow because $q_k^\dagger \ge q^*$, binomial lower tails decrease in the success probability, Holm's rejection set is monotone under coordinatewise decreases of the p-values, and Holm rejects every hypothesis that single-step Bonferroni rejects. The full proof is in Section A.

Validity alone does not show that a conservative test returns a useful selector, so we next compare the procedure with a finite-sample oracle. Recall $q^\dagger(p)$ from Corollary 1, let $\epsilon_N = \sqrt{\log(4/\delta)/(2N)}$, and for $\beta \in (0,1)$ set

$$\begin{aligned} p_k^- &= [p_k - 2\epsilon_N]_+, \\ g_k(\beta) &= \sqrt{\frac{\log(K/\beta)}{2m_k}} + \sqrt{\frac{\log(2K/\delta)}{2m_k}}, \\ \mathcal{O}_\beta &= \{k : m_k > 0,\ q_k + g_k(\beta) \le q^\dagger(p_k^-)\}. \end{aligned} \tag{22}$$

The set $\mathcal{O}_\beta$ depends on the data only through the accepted counts m_k, that is, through the calibration scores but not the labels.

Proposition 3 (Margin-dependent oracle recovery)**.** *Let $\widehat{k}$ be the index returned by* HOLM–SHARP-RC*, with reported coverage certificate $L_\Gamma(\widetilde{p}_{\widehat{k}})$, and set the certificate to zero when the method abstains everywhere. With probability at least $1 - \delta - \beta$, the method either abstains everywhere or returns a risk-valid selector, and*

$$L_\Gamma(\widetilde{p}_{\widehat{k}}) \ge \max_{k \in \mathcal{O}_\beta} L_\Gamma(p_k^-), \tag{23}$$

where the maximum over the empty set is zero. Thus, every threshold whose population error rate clears its sharp target by the explicit coverage- and label-estimation margin in Equation (22) *is certified, and the returned certificate is at least the robust coverage of the best such threshold, evaluated at p_k^-.*

Proof idea. On the two-sided DKW event, $\widetilde{p}_k \ge p_k^-$; hence $q_k^\dagger \ge q^\dagger(p_k^-)$ and $L_\Gamma(\widetilde{p}_k) \ge L_\Gamma(p_k^-)$. For each threshold, apply Hoeffding's inequality conditional on its accepted count and then average over that count; a union bound gives $x_k/m_k \le q_k + \sqrt{\log(K/\beta)/(2m_k)}$ for all k with $m_k > 0$, with probability at least $1 - \beta$. For $k \in \mathcal{O}_\beta$, the empirical error rate then lies at least $\sqrt{\log(2K/\delta)/(2m_k)}$ below $q_k^\dagger$, and Hoeffding's bound on the binomial lower tail yields $\pi_k \le \delta/(2K)$. Such a threshold passes the single-step test and is therefore certified by Holm, by Theorem 4(ii).

Use and tradeoffs. HOLM–SHARP-RC is target-specific: changing (α, Γ) changes the hypotheses. The budget-uniform certificate, by contrast, holds for all Γ on a single event and therefore supports post-hoc sensitivity analysis, including misspecification analysis at any larger budget Γ'; exact formulas for budget misspecification and arbitrary contamination are given in Section B. If no

independent score-only pool exists, a predeclared split of a single source sample supplies U and D without affecting Theorem 4; reusing the same observations in both roles is not covered. The computational cost is $O((N+n)\log(N+n) + K\log K)$. Accepted-sample requirements for the budget-uniform baseline are given in Section A.11.

Why step-down rather than fixed-sequence testing. Fixed-sequence testing along the threshold order avoids the multiplicity factor but stops at the first non-rejection. It is attractive when risk is monotone along the sequence and the first hypotheses are well powered. Neither condition holds here in general: the sharp targets $q_k^\dagger$ vary with coverage, so the certifiable set need not be contiguous on the grid, and the most conservative thresholds, where such a sequence would naturally begin, accept the fewest labeled examples. Holm's procedure is order-free and remains valid under the arbitrary dependence induced by nested acceptance sets.

6 Numerical Evaluation

The evaluation asks whether the formulas agree with independent optimization, whether the sharper envelope changes the selected rule in the predicted regime, and whether the returned rules meet the stated risk target. Explicitly specified distributions allow population risk and coverage to be evaluated exactly, without test-sample noise.

Independent optimization. We compare Equation (5) with a three-cell linear-fractional program solved via the Charnes–Cooper transformation. Across 1,720 settings, including boundary cases and 1,000 seeded random cases, the maximum absolute difference is 6.13×10^{-13}. A separate coverage program agrees with Proposition 2 to within 1.11×10^{-16}. For bounded losses, 500 finite-population programs agree with the root of Equation (14) to within 5.44×10^{-14}. These checks validate the implementation; the general claims rest on the proofs.

Two coverage regimes. Let

$$S \sim \mathrm{Uniform}(0,1), \qquad \ell \mid S = s \sim \mathrm{Bernoulli}\big(0.002 + c(1-s)^2\big). \tag{24}$$

Accepting $S \geq \tau$ gives $p = 1 - \tau$ and $q = 0.002 + (c/3)p^2$. We use $c = 0.24$ and $c = 0.12$ to probe the odds and normalization-limited regimes, respectively. Each setting uses 1,000 independent calibration draws of $n = 20{,}000$ labeled pairs and $N = 20{,}000$ independent score-only observations, with $\alpha = 0.10$, $\Gamma = 2$, $\delta = 0.05$, and the fixed grid $\{0, 0.02, \ldots, 1\}$. The four rules cross the error target (odds or sharp) with the multiplicity correction (Bonferroni or Holm) and, within each draw, share the same score-only DKW band and labeled counts. This 2×2 design separates the effect of normalization from that of testing while holding the data and score fixed.

Table 1: **Matched calibration under specified score–loss distributions.** Means over 1,000 draws per regime. R_Γ and C_Γ are exact population quantities of the returned rule; $\underline{C}$ is the reported coverage certificate. Failures count returned rules with $R_\Gamma > 0.10$; every returned rule was nonempty. Within each draw, all rows use the same score-only coverage data and labeled counts.

Regime	Calibration rule	R_Γ	C_Γ	$\underline{C}$	Failures
$c = 0.24$	Bonferroni odds	.07914	.24373	.23850	0/1000
	Bonferroni sharp	.07914	.24373	.23850	0/1000
	Holm odds	.08012	.24551	.24029	0/1000
	HOLM–SHARP-RC	.08012	.24551	.24029	0/1000
$c = 0.12$	Bonferroni odds	.07499	.41893	.39866	0/1000
	Bonferroni sharp	.08388	.99660	.97566	0/1000
	Holm odds	.07469	.43494	.41434	0/1000
	HOLM–SHARP-RC	.08397	**.99920**	**.97826**	0/1000

When the sharp envelope matters. For $c = 0.24$, sharp and odds calibration select the same rule in every draw under either correction. The returned rules lie in the odds regime, where accounting for normalization yields no improvement. For $c = 0.12$, the matched Holm comparison raises mean

worst-case coverage from 0.435 to 0.999, while mean worst-case risk increases from 0.075 to 0.084, still below α. The sharp rule attains greater coverage in every draw, and its mean coverage certificate is 0.978, compared with 0.414 for Holm odds. At full acceptance, this population has source error 0.042, exact robust risk $0.084 \leq \alpha$, and odds envelope $0.149 > \alpha$; normalization alone makes this operating point certifiable. The paired coverage gain is 56.4 percentage points (standard error 0.15), whereas replacing Bonferroni with Holm at the sharp target adds only 0.26 points (standard error 0.045). Nearly all of the gain thus comes from the envelope. Detailed paired contrasts appear in Section D.

Risk control and negative controls. No robust rule returned a risk-violating selector in either regime. For each method and regime, zero failures in 1,000 draws gives a one-sided 95% Clopper–Pearson upper limit of 0.00299 on the failure probability; the 8,000 method evaluations share 2,000 calibration draws, so these limits use the 1,000 independent draws within each regime and are not pooled across methods. Nominal calibration that ignores shift, using either the empirical error rate or a binomial test, selects full acceptance; it violates the robust target in every draw when $c = 0.24$ and satisfies it in every draw when $c = 0.12$. The high-coverage case thus demonstrates recovery of a robustly valid operating point, not an accuracy improvement over nominal selection. Grid-wide coverage-band failures and implementation details are reported in Section D.

7 Discussion and Conclusion

Normalization couples errors, correct predictions, and abstentions. Accounting for this coupling yields an exact selective-risk envelope and identifies a high-coverage regime in which independent error-odds bounds are conservative. The resulting coverage-dependent target gives a finite-sample calibration procedure that is valid after selection and deterministically contains the matched odds-envelope certificate. The numerical results exhibit both the predicted improvement and a regime in which the two envelopes agree.

Limitations. The guarantee applies to a declared joint density-ratio ambiguity set. Under pure covariate shift it remains valid but need not be sharp, and changes of support fall outside the set. The predictor, score, and evaluation loss must be fixed before calibration; when the loss is supplied by a judge, the certificate concerns that judge's evaluations. The numerical study isolates the calibration mechanism through specified score–loss distributions. Whether comparable coverage gains arise for trained predictors is an empirical question whose answer requires fixed prediction records and an independently specified evaluation protocol. The bounded-loss characterization extends the same normalization argument, and the contamination bounds in Section B quantify sensitivity to departures from the reference support.

AI USE STATEMENT

In this work, we used generative AI tools (large language models) solely for language polishing, including improving grammar, clarity, and readability. These tools were not used to generate substantive content or to perform any research tasks, including developing methods or theoretical frameworks, formulating claims or hypotheses, constructing proofs, designing experiments, generating or processing data, analyzing or interpreting results, writing code, searching the literature, translating text, or preparing figures. All AI-assisted revisions were reviewed by the authors to ensure that they preserved the intended meaning and technical accuracy. We take full responsibility for the final content of this work, including all claims, results, and references.

REFERENCES

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.

Tian Bai and Ying Jin. Conformal selective prediction with general risk control. *arXiv preprint arXiv:2603.24704*, 2026. URL https://arxiv.org/abs/2603.24704.

Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C. Duchi. Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association*, 119 (548):3033–3044, 2024.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

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.

Jacob Dorn and Kevin Guo. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, 118(544):2645–2657, 2023. doi: 10.1080/01621459.2022.2069572.

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

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

Alvin Heng and Harold Soh. Know when to abstain: Optimal selective classification with likelihood ratios. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/3fe2a777282299ecb4f9e7ebb531f0ab-Paper-Conference.pdf.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Ying Jin, Zhimei Ren, and Emmanuel J. Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120 (6):e2214889120, 2023. doi: 10.1073/pnas.2214889120. URL https://arxiv.org/abs/2111.12161.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5684–5696, 2020.

Hengyue Liang, Le Peng, and Ju Sun. Selective classification under distribution shifts. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=dmxMGW6J7N.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

Sangdon Park, Edgar Dobriban, Insup Lee, and Osbert Bastani. PAC prediction sets under covariate shift. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=DhP9L8vIyLc.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yunpeng Xu, Wenge Guo, and Zhi Wei. Selective conformal risk control. *arXiv preprint arXiv:2512.12844*, 2025. URL https://arxiv.org/abs/2512.12844.

A Deferred Proofs

A.1 Exact normalized binary risk

Proof. Write $A = \{a = 1\}$ and partition the observation space into accepted errors, accepted successes, and rejections. Their source masses are $e = pq$, $b = p(1-q)$, and $c = 1-p$, respectively. Here the density ratio is defined on the full observation space, including the outcome on which the loss depends. For any feasible deployment distribution, write its masses on these three events as u, v, t. They satisfy

$$\frac{e}{\Gamma} \le u \le \Gamma e, \qquad \frac{b}{\Gamma} \le v \le \Gamma b, \qquad \frac{c}{\Gamma} \le t \le \Gamma c, \qquad u + v + t = 1.$$

Consequently,

$$u \le u_* := \min\{\Gamma e, 1 - (1-e)/\Gamma\}, \qquad v \ge v_* := \max\{b/\Gamma, 1 - \Gamma(1-b)\}.$$

The ratio $u/(u+v)$ increases with u and decreases with v. It remains to show that u_*, v_* are jointly feasible.

For $\Gamma = 1$, the only feasible masses are (e, b, c). Suppose $\Gamma > 1$. There are three cases. If $e \le 1/(\Gamma+1)$ and $b \le \Gamma/(\Gamma+1)$, take $(u, v) = (\Gamma e, b/\Gamma)$; the two displayed inequalities imply $c/\Gamma \le 1 - u - v \le \Gamma c$. If $e > 1/(\Gamma+1)$, then $b < \Gamma/(\Gamma+1)$, and take

$$(u, v, t) = \big(1 - (1-e)/\Gamma, b/\Gamma, c/\Gamma\big).$$

Finally, if $b > \Gamma/(\Gamma+1)$, then $e < 1/(\Gamma+1)$, and take

$$(u, v, t) = \big(\Gamma e, 1 - \Gamma(1-b), \Gamma c\big).$$

Each triple lies in the required intervals and sums to one. A density ratio constant on each event realizes the triple: on an event with positive source mass, divide its deployment mass by its source mass; events of source mass zero require no assignment beyond an arbitrary value in the allowed interval. Thus $R_\Gamma(a) = u_*/(u_* + v_*) = \Psi_\Gamma(p, q)$.

For fixed $p > 0$, u_* increases with q and v_* decreases with q, proving monotonicity in q. For monotonicity in p, divide both masses by p:

$$\frac{u_*}{p} = \min\left\{\Gamma q, \frac{q}{\Gamma} + \frac{\Gamma - 1}{\Gamma p}\right\}, \qquad \frac{v_*}{p} = \max\left\{\frac{1-q}{\Gamma}, \Gamma(1-q) - \frac{\Gamma - 1}{p}\right\}.$$

The first quantity is nonincreasing in p, and the second is nondecreasing. Therefore $\Psi_\Gamma(p, q)$ is nonincreasing in p. For $\Gamma > 1$, letting $p \downarrow 0$ in these expressions gives Γq and $(1-q)/\Gamma$, respectively; hence the extension $\Psi_\Gamma(0, q) = \rho_\Gamma(q)$ is continuous for every $q \in [0, 1]$. For $\Gamma = 1$, $\Psi_1(p, q) = q$ for all $p \ge 0$. □

A.2 Odds envelope and tightness

Proof. With the notation of the preceding proof, every feasible distribution has $u \le \Gamma pq$ and $v \ge p(1-q)/\Gamma$. Therefore

$$R_\Gamma(a) \le \frac{\Gamma pq}{\Gamma pq + p(1-q)/\Gamma} = \rho_\Gamma(q).$$

For $\Gamma > 1$, the simultaneous extremal assignment is feasible exactly when

$$pq \le \frac{1}{\Gamma+1}, \qquad p(1-q) \le \frac{\Gamma}{\Gamma+1},$$

as the three-mass calculation above shows. Under these conditions the bound is attained. If $\Gamma > 1$ and $0 < q < 1$, failure of either condition makes the corresponding mass inequality strict, so the risk bound is strict. At $q = 0$ or $q = 1$, the risk equals q for every feasible distribution irrespective of these conditions. When $\Gamma = 1$, equality always holds. For $0 < \alpha < 1$, rearranging $\Gamma^2 q \le \alpha(\Gamma^2 q + 1 - q)$ gives $\rho_\Gamma(q) \le \alpha$ if and only if $q \le q^*$. This is an equivalence for the envelope, and only a sufficient condition for the exact normalized risk in general. □

A.3 Regime decomposition

Substituting the two branches of $U_\Gamma(pq)$ and $L_\Gamma(p(1-q))$ into Equation (5) gives Equation (9). The switch for the accepted-error mass occurs at $pq = 1/(\Gamma+1)$, and the switch for the accepted-correct mass occurs at $p(1-q) = \Gamma/(\Gamma+1)$. Both strict inequalities cannot hold because their sum would imply $p > 1$. In the accepted-correct saturation regime,

$$\Psi_\Gamma(p,q) = \frac{\Gamma pq}{\Gamma pq + \Gamma p(1-q) - (\Gamma-1)} = \frac{\Gamma pq}{\Gamma p - (\Gamma-1)}.$$

Solving the inequality $\Psi_\Gamma(p,q) \le \alpha$ yields Equation (11). The other two branches follow in the same way from Equation (4).

A.4 Sharp error target

Proof of Corollary 1. The case $\Gamma = 1$ gives $\Psi_1(p,q) = q$ and $q^\dagger(p) = \alpha$. For $\Gamma > 1$ and $p > 0$, put

$$u_1 = \Gamma pq, \qquad u_2 = \frac{pq + \Gamma - 1}{\Gamma}, \qquad v_1 = \frac{p(1-q)}{\Gamma}, \qquad v_2 = \Gamma p(1-q) - (\Gamma-1).$$

Then $U_\Gamma(pq) = \min\{u_1, u_2\}$ and $L_\Gamma(p(1-q)) = \max\{v_1, v_2\}$. The inequalities $u_2 < u_1$ and $v_2 > v_1$ would require $pq > 1/(\Gamma+1)$ and $p(1-q) > \Gamma/(\Gamma+1)$, respectively. They cannot both hold because $p \le 1$. Consequently an active numerator–denominator pair can always be chosen from (u_1, v_1), (u_1, v_2), and (u_2, v_1), including ties. It follows that $\Psi_\Gamma(p,q) \le \alpha$ if and only if at least one of

$$(1-\alpha)u_1 \le \alpha v_1, \qquad (1-\alpha)u_1 \le \alpha v_2, \qquad (1-\alpha)u_2 \le \alpha v_1$$

holds. Necessity uses an active pair; sufficiency follows from $U_\Gamma(pq) \le u_i$ and $L_\Gamma(p(1-q)) \ge v_j$. Solving the three inequalities gives, respectively,

$$q \le q^*, \qquad q \le \alpha\left(1 - \frac{\Gamma-1}{\Gamma p}\right), \qquad q \le \alpha - \frac{(1-\alpha)(\Gamma-1)}{p}.$$

Their union proves Equation (10). Each right-hand side is nondecreasing in p, and the first is independent of p. At $p = 1$, $\Psi_\Gamma(1,q) = U_\Gamma(q)$, whose inverse at α is

$$q^\dagger(1) = \max\{\alpha/\Gamma,\ 1 - \Gamma(1-\alpha)\} = L_\Gamma(\alpha).$$

The extension at $p = 0$ follows from the odds regime for sufficiently small p. □

A.5 Sharp bounded-loss risk

Proof of Theorem 2. Write $l = 1/\Gamma$, $u = \Gamma$, and represent every feasible density ratio as

$$w = l + (u-l)\xi, \qquad 0 \le \xi \le 1.$$

The normalization constraint $\mathbb{E}_P w = 1$ is equivalent to

$$\mathbb{E}_P \xi = \frac{1-l}{u-l} = \frac{1}{\Gamma+1}.$$

Consequently, for any integrable H,

$$\sup_{Q \in \mathcal{Q}_\Gamma(P)} \mathbb{E}_Q H = \frac{1}{\Gamma}\mathbb{E}_P H + \left(\Gamma - \frac{1}{\Gamma}\right) \mathsf{T}_{1/(\Gamma+1)}(H).$$

The variational definition of T_κ also shows that an optimizer puts the larger weight on an upper κ fraction of H, with fractional allocation only at a quantile atom.

Now let $N_Q = \mathbb{E}_Q[a\ell]$ and $D_Q = \mathbb{E}_Q a$. Since $p > 0$, every feasible Q satisfies $D_Q \geq p/\Gamma > 0$. For $H_r = a(\ell - r)$, define

$$\phi(r) = \sup_{Q \in \mathcal{Q}_\Gamma(P)} \mathbb{E}_Q H_r = \sup_{Q \in \mathcal{Q}_\Gamma(P)} (N_Q - r D_Q).$$

Since $0 < D_Q \leq 1$, the function ϕ is 1-Lipschitz and hence continuous in r. If $r < R_\Gamma(a)$, some feasible Q has $N_Q/D_Q > r$, hence $\phi(r) > 0$. If $r > R_\Gamma(a)$, then for every feasible Q,

$$N_Q - r D_Q = D_Q \left(\frac{N_Q}{D_Q} - r \right) \leq \frac{p}{\Gamma} \big(R_\Gamma(a) - r \big) < 0,$$

where the inequality uses $D_Q \geq p/\Gamma$ and the negative sign of $R_\Gamma(a) - r$. Thus ϕ changes sign exactly at $R_\Gamma(a)$. Moreover, $\phi(r_1) - \phi(r_2) \geq (p/\Gamma)(r_2 - r_1)$ for $r_1 < r_2$, so the root is unique. Substituting the preceding robust-expectation identity gives Equation (14). When ℓ is binary, H_r takes the three ordered values $1 - r$, 0, and $-r$ on accepted errors, rejections, and accepted correct predictions; allocating the upper-tail mass across these three cells recovers Equation (5). For $\Gamma = 1$, $Q = P$ and the stated source-risk root is immediate. □

A.6 Worst-case coverage

Proof. For any feasible density ratio w,

$$\mathbb{E}_P[wa] \geq p/\Gamma, \qquad \mathbb{E}_P[wa] = 1 - \mathbb{E}_P[w(1-a)] \geq 1 - \Gamma(1-p).$$

At $p = 0$ and $p = 1$, the claimed values are immediate. If $0 < p < 1$ and $p \leq \Gamma/(\Gamma + 1)$, assign weight $1/\Gamma$ on the accepted region and $(1 - p/\Gamma)/(1 - p)$ on its complement. These weights are feasible and attain p/Γ. If $p > \Gamma/(\Gamma+1)$, assign weight Γ on the rejected region and $(1-\Gamma(1-p))/p$ on the accepted region. This assignment is feasible and attains $1 - \Gamma(1-p)$. The displayed lower bound is therefore exact. □

A.7 Unrestricted-shift observation

Proposition 4 (No guarantee under unrestricted reweighting)**.** *Let* $B = \{a = 1, \ell = 1\}$ *have* $P(B) > 0$*. Then* $\sup_{Q \ll P,\, C_Q(a) > 0} R_Q(a) = 1$.

Proof. Take $Q = P(\cdot \mid B)$. This distribution is absolutely continuous with respect to P because $P(B) > 0$, and under Q both acceptance and error occur with probability one. Thus $R_Q(a) = 1$, which is the largest possible binary selective risk. □

A.8 Uniform risk and coverage certificate

Proof. Condition on any independent data used to fit the generator or score or to select the threshold grid. Fix a grid threshold τ_k with positive source coverage. Conditional on its acceptance indicators, the accepted losses are independent Bernoulli random variables with parameter $q(\tau_k)$. Hence, for every accepted count m_k,

$$\mathbb{P}\left(q(\tau_k) > \bar{q}_k \mid m_k \right) \leq \frac{\delta}{2K}.$$

At $m_k = 0$ this follows from the convention $\bar{q}_k = 1$. Averaging over m_k and applying a union bound gives $q(\tau_k) \leq \bar{q}_k$ simultaneously for all relevant k except on an event of probability at most $\delta/2$.

The two-sided Dvoretzky–Kiefer–Wolfowitz–Massart inequality for the empirical score distribution gives

$$\mathbb{P}\left(\sup_\tau |\hat{p}(\tau) - p(\tau)| > \epsilon \right) \leq 2 \exp(-2n\epsilon^2).$$

The same bound applies to survival functions defined using $s \geq \tau$, including scores with atoms, by taking left limits of the empirical and population distribution functions. With $\epsilon = \sqrt{\log(4/\delta)/(2n)}$, this probability is $\delta/2$. On the complementary event, $p(\tau_k) \geq \underline{p}_k = \max\{0, \hat{p}_k - \epsilon\}$ for all k.

On the intersection of the two events, monotonicity of Ψ_Γ gives

$$R_\Gamma(a_{\tau_k}) = \Psi_\Gamma(p(\tau_k), q(\tau_k)) \le \Psi_\Gamma(\underline{p}_k, \bar{q}_k).$$

The extension at $\underline{p}_k = 0$ makes this statement well defined. Also, monotonicity of c_Γ gives $C_\Gamma(a_{\tau_k}) \ge c_\Gamma(\underline{p}_k)$. Any selector with zero source coverage abstains almost surely under every $Q \ll P$ and has no answered-query risk to certify; it can be treated under the stated abstention convention. A final union bound gives probability at least $1 - \delta$. Since the inequalities hold for every grid point on the same event, arbitrary selection among the certified grid points preserves them. Removing the initial conditioning proves the unconditional result. □

A.9 Proof of the Holm-optimized certificate

Proof. Condition on the independent design data and on the score-only sample U. The quantities $\widetilde{p}_k$ and $q_k^\dagger$ are then fixed. For any threshold with $m_k > 0$, conditional on m_k its accepted-error count has distribution $\mathrm{Bin}(m_k, q_k)$. Under the null $H_k : q_k \ge q_k^\dagger$, the lower tail is largest at the boundary $q_k = q_k^\dagger$. Consequently the value π_k in Equation (19) is super-uniform under H_k, including its discrete boundary cases. Setting $\pi_k = 1$ for $m_k = 0$ preserves this property. Holm's step-down procedure therefore controls the probability of rejecting any true H_k by $\delta/2$, without an independence assumption among the K tests.

Independently, DKW applied to the score distribution of U yields, with probability at least $1 - \delta/2$,

$$p_k = P\{s(X) \ge \tau_k\} \ge \widetilde{p}_k \quad \text{for every } k.$$

On this event, if $R_\Gamma(a_{\tau_k}) > \alpha$, then

$$\Psi_\Gamma(\widetilde{p}_k, q_k) \ge \Psi_\Gamma(p_k, q_k) > \alpha,$$

so monotonicity in q implies $q_k > q_k^\dagger$. Thus any rejected threshold whose robust risk exceeds α would be a false Holm rejection. A union bound over the DKW and familywise-testing events gives the claimed probability $1 - \delta$. On the DKW event, monotonicity of L_Γ also gives $C_\Gamma(a_{\tau_k}) \ge L_\Gamma(\widetilde{p}_k)$ simultaneously.

For the odds-envelope procedure, the admissible source-error target is q^*. Since $\Psi_\Gamma(p, q) \le \rho_\Gamma(q)$, $q_k^\dagger \ge q^*$. A binomial lower-tail probability decreases as its success probability increases, so every sharp raw p-value is no larger than its envelope counterpart. Holm's step-down rejection rule is monotone under coordinatewise decreases of the raw p-values, proving the first nesting statement. Finally, any raw value at most $\delta/(2K)$ and all preceding ordered values pass their corresponding Holm critical values; hence Holm contains every matched single-step Bonferroni rejection. □

A.10 Proof of margin-dependent oracle recovery

Proof of Proposition 3. Let E_{valid} be the event from Theorem 4; it has probability at least $1 - \delta$ and includes the uniform DKW event. On that DKW event, $|r_k/N - p_k| \le \epsilon_N$ for every threshold, hence

$$\widetilde{p}_k = [r_k/N - \epsilon_N]_+ \ge [p_k - 2\epsilon_N]_+ = p_k^- .$$

Because $\Psi_\Gamma(p, q)$ is nonincreasing in p, $q^\dagger(p)$ is nondecreasing, and therefore $q_k^\dagger = q^\dagger(\widetilde{p}_k) \ge q^\dagger(p_k^-)$.

For each fixed k, condition on its own accepted count m_k. The binomial sampling law and Hoeffding's inequality give

$$\mathbb{P}\left\{ m_k > 0,\ \frac{x_k}{m_k} > q_k + \sqrt{\frac{\log(K/\beta)}{2m_k}} \right\} \le \frac{\beta}{K},$$

after averaging over m_k. A union bound therefore gives, with probability at least $1 - \beta$,

$$\frac{x_k}{m_k} \le q_k + \sqrt{\frac{\log(K/\beta)}{2m_k}} \quad \text{for every } k \text{ with } m_k > 0.$$

Fix $k \in \mathcal{O}_\beta$. Combining this display with Equation (22) gives

$$q_k^\dagger - \frac{x_k}{m_k} \geq \sqrt{\frac{\log(2K/\delta)}{2m_k}}.$$

Hoeffding's binomial lower-tail bound at success probability $q_k^\dagger$ consequently yields $\pi_k \leq \exp\{-2m_k(q_k^\dagger - x_k/m_k)^2\} \leq \delta/(2K)$. Every such threshold passes Holm because the step-down procedure contains all single-step Bonferroni rejections. The selector returned by Algorithm 1 maximizes $L_\Gamma(\widetilde{p}_k)$ among the Holm-rejected thresholds, and monotonicity of L_Γ gives

$$\underline{C}_{\widehat{k}} = L_\Gamma(\widetilde{p}_{\widehat{k}}) \geq \max_{k \in \mathcal{O}_\beta} L_\Gamma(\widetilde{p}_k) \geq \max_{k \in \mathcal{O}_\beta} L_\Gamma(p_k^-).$$

On E_{valid} the returned selector also has risk at most α. A union bound with the Hoeffding event proves probability at least $1 - \delta - \beta$. □

A.11 Accepted-sample requirement

Corollary 2 (A sufficient accepted-sample requirement)**.** *With probability at least $1 - \delta/2$, every grid threshold with $m_k > 0$ and*

$$q(\tau_k) + 2\sqrt{\frac{\log(2K/\delta)}{2m_k}} \leq q^*(\alpha, \Gamma) \tag{25}$$

is certified by both budget-uniform variants. Intersecting this event with Theorem 3 gives probability at least $1 - 3\delta/2$ for both certification and certificate validity; the two confidence statements are distinct.

Proof. We first compare the one-sided Clopper–Pearson upper limit with a Hoeffding upper limit at the same failure probability γ. For $m > 0$ and $0 \leq x \leq m$, put

$$h = \min\left\{1, \frac{x}{m} + \sqrt{\frac{\log(1/\gamma)}{2m}}\right\}.$$

If $h = 1$, then $\bar{q}(x, m; \gamma) \leq h$ trivially. Otherwise, Hoeffding's binomial lower-tail inequality gives

$$\mathbb{P}_{Z \sim \mathrm{Bin}(m,h)}(Z \leq x) \leq \exp\{-2m(h - x/m)^2\} = \gamma.$$

The binomial lower tail is decreasing in its success probability, and its inversion defines the Clopper–Pearson upper limit. Consequently $\bar{q}(x, m; \gamma) \leq h$ in this case as well.

Now set $\gamma = \delta/(2K)$ and $t_k = \sqrt{\log(2K/\delta)/(2m_k)}$ when $m_k > 0$. Hoeffding's inequality, conditional on m_k, followed by a union bound shows that, with probability at least $1 - \delta/2$,

$$\frac{x_k}{m_k} \leq q(\tau_k) + t_k \quad \text{for every grid point with } m_k > 0.$$

On this event, $\bar{q}_k \leq x_k/m_k + t_k \leq q(\tau_k) + 2t_k$. Therefore each threshold satisfying the corollary's gap condition has $\bar{q}_k \leq q^*$. Its envelope certificate satisfies $\rho_\Gamma(\bar{q}_k) \leq \alpha$, and the exact certificate is at least as permissive because $\Psi_\Gamma(\underline{p}_k, \bar{q}_k) \leq \rho_\Gamma(\bar{q}_k)$. This proves the stated certification probability $1 - \delta/2$ without changing the algorithm or combining additional confidence bounds. Rearrangement of $2t_k \leq q^* - q$ gives the accepted-sample requirement in the corollary. □

Zero-error specialization. When $x_k = 0$ and $m_k > 0$, the exact one-sided binomial limit is

$$\bar{q}_k = 1 - \left(\frac{\delta}{2K}\right)^{1/m_k}.$$

Thus the envelope certificate is nonempty at that threshold whenever

$$m_k \geq \frac{\log(2K/\delta)}{-\log(1 - q^*)}.$$

This is a sufficient accepted-sample requirement for the exact method as well. It describes the observed zero-error case; it does not assert that a population with nonzero error will produce zero observed errors. For $\alpha = 0.10$, $\Gamma = 2$, $K = 50$, and $\delta = 0.05$, the required integer count is 278.

B Budget Misspecification and Contamination

On the event of Theorem 3, a larger true budget Γ' has the valid certificate $\Psi_{\Gamma'}(\underline{p}_k, \bar{q}_k)$. For an *envelope-certified* threshold, the coverage-independent ceiling is

$$R_{\Gamma'}(a_{\tau_k}) \le \rho_{\Gamma'}(q^*(\alpha,\Gamma)) = \frac{\alpha\Gamma'^2}{\alpha\Gamma'^2 + (1-\alpha)\Gamma^2}. \tag{26}$$

This expression need not hold for a threshold admitted only by the sharp certificate; such a threshold must use the first bound.

Proposition 5 (Arbitrary contamination)**.** *Let $Q = (1-\varepsilon)Q' + \varepsilon Q_{\mathrm{out}}$, where $Q' \in \mathcal{Q}_\Gamma(P)$, $\varepsilon \in [0,1)$, and Q_{out} is arbitrary. For $p > 0$, set $u = U_\Gamma(pq)$ and $v = L_\Gamma(p(1-q))$. Then*

$$R_Q(a) \le T_{\Gamma,\varepsilon}(p,q) := \frac{(1-\varepsilon)u + \varepsilon}{(1-\varepsilon)(u+v) + \varepsilon}. \tag{27}$$

The bound is nondecreasing in q and nonincreasing in p. With $T_{\Gamma,\varepsilon}(0,q) = 1$ for $\varepsilon > 0$ and $T_{\Gamma,0}(0,q) = \rho_\Gamma(q)$, evaluating it at $(\underline{p}_k, \bar{q}_k)$ gives a valid finite-sample bound on the event of Theorem 3.

For $p > 0$, the looser bound $R_Q(a) \le \Psi_\Gamma(p,q) + \varepsilon\Gamma/((1-\varepsilon)p)$, capped at one, shows why small contamination is especially consequential at low coverage. Neither formula identifies a global shift budget or contamination mass from finite target data without additional assumptions.

B.1 Proof of the contamination bound

Proof. Let $u = Q'(a=1,\ell=1)$ and $v = Q'(a=1,\ell=0)$. Let h and j be the corresponding two probabilities under Q_{out}, so $h, j \ge 0$ and $h + j \le 1$. The deployment risk is

$$\frac{(1-\varepsilon)u + \varepsilon h}{(1-\varepsilon)(u+v) + \varepsilon(h+j)}.$$

This is at most its value at $h = 1, j = 0$: decreasing j and then increasing h cannot reduce the ratio. It is then increasing in u and decreasing in v. The exact-mass inequalities proved above therefore give

$$R_Q(a) \le \frac{(1-\varepsilon)u_* + \varepsilon}{(1-\varepsilon)(u_* + v_*) + \varepsilon},$$

where

$$u_* = \min\{\Gamma pq, 1 - (1-pq)/\Gamma\}, \qquad v_* = \max\{p(1-q)/\Gamma, 1 - \Gamma(1-p(1-q))\}.$$

The same feasible in-support distribution attains both masses. If an arbitrary contamination distribution can place all of its mass on an accepted error, the displayed bound is attained. Attainability is unnecessary for its validity.

The bound is nondecreasing in q because u_* increases and v_* decreases with q. To prove it is nonincreasing in $p > 0$, set $\kappa = \varepsilon/(1-\varepsilon)$ and divide its numerator and denominator by $(1-\varepsilon)p$. The numerator is $u_*/p + \kappa/p$, a nonincreasing function of p, while the remaining denominator term v_*/p is nondecreasing, as established in the proof of Theorem 1. At $p = 0$, the safe continuous extension is one when $\varepsilon > 0$, and $\rho_\Gamma(q)$ when $\varepsilon = 0$. Thus substituting the simultaneous confidence bounds $(\underline{p}_k, \bar{q}_k)$ proves the finite-sample consequence.

For comparison, replacing u_* by Γpq and v_* by $p(1-q)/\Gamma$ recovers the simpler upper bound

$$\frac{(1-\varepsilon)\Gamma pq + \varepsilon}{(1-\varepsilon)(\Gamma pq + p(1-q)/\Gamma) + \varepsilon}.$$

Writing its numerator and denominator before adding ε as N and M, respectively, gives $0 \le N \le M$ and $M \ge (1-\varepsilon)p/\Gamma$. Hence

$$R_Q(a) \le \rho_\Gamma(q) + \frac{\varepsilon}{(1-\varepsilon)p/\Gamma},$$

using $(N+\varepsilon)/(M+\varepsilon) \le N/M + \varepsilon/M$. Applying the same inequality directly to the exact masses u_*, v_*, whose sum is at least p/Γ, also yields $R_Q(a) \le \Psi_\Gamma(p,q) + \varepsilon\Gamma/((1-\varepsilon)p)$, as stated in the main text. The additive expression can exceed one and is only an interpretation of sensitivity; the preceding ratio bound remains in $[0,1]$. □

C Bounded Losses

Theorem 2 gives a sharp population characterization for losses in $[0,1]$ but depends on their distribution, not only their mean. The following mean-based certificate is less sharp, but it can be computed from the accepted calibration sample without estimating an upper-tail functional.

Theorem 5 (Bounded-loss certificate)**.** *Let D be an i.i.d. calibration sample from P, let $\mathcal{A}$ be a finite family of M selectors chosen independently of D, and let $\ell \in [0,1]$. For each $a \in \mathcal{A}$, let m_a be its accepted count and, when $m_a > 0$, let $\widehat{\mu}_a$ be its mean accepted loss. For $\delta \in (0,1)$ define*

$$\overline{\mu}_a = \begin{cases} \min\left\{1, \widehat{\mu}_a + \sqrt{\dfrac{\log(M/\delta)}{2m_a}}\right\}, & m_a > 0, \\ 1, & m_a = 0. \end{cases}$$

With probability at least $1-\delta$, simultaneously for all selectors with positive source coverage,

$$R_\Gamma(a) \le \rho_\Gamma(\overline{\mu}_a).$$

Consequently, any selector with $\overline{\mu}_a \le q^*(\alpha, \Gamma)$ is certified at target α. The statement permits selection among these selectors after observing D.

Proof. For a selector with source coverage $p > 0$, write $\mu = \mathbb{E}_P[\ell \mid a = 1]$. For every feasible density ratio, define $u = \mathbb{E}_P[wa\ell]$ and $v = \mathbb{E}_P[wa(1-\ell)]$. Since both loss terms are nonnegative,

$$u \le \Gamma p \mu, \qquad v \ge p(1-\mu)/\Gamma.$$

Their sum is $\mathbb{E}_P[wa]$, so ratio monotonicity gives $R_Q(a) \le \rho_\Gamma(\mu)$ for every $Q \in \mathcal{Q}_\Gamma(P)$. Conditional on the acceptance indicators, the accepted losses are independent, lie in $[0,1]$, and have common mean μ. Hoeffding's inequality therefore gives $\mathbb{P}(\mu > \overline{\mu}_a \mid m_a) \le \delta/M$ for every accepted count. The $m_a = 0$ convention is valid because $\mu \le 1$. After averaging over m_a, a union bound over the M selectors and monotonicity of ρ_Γ prove the claim. □

To report coverage jointly with this bounded-loss certificate, use $\delta/2$ in place of δ in Theorem 5. For threshold selectors of a common fixed score, combine it with the DKW lower coverage bound at failure probability $\delta/2$. For a general finite family of M selectors, use instead

$$\underline{p}_a = \left[\frac{m_a}{n} - \sqrt{\frac{\log(2M/\delta)}{2n}}\right]_+.$$

Hoeffding's inequality and a union bound make these coverage lower bounds simultaneously valid with probability at least $1-\delta/2$; $L_\Gamma(\underline{p}_a)$ then lower-bounds robust coverage. Binary losses admit the tighter Clopper–Pearson construction used in the main text.

D Numerical Protocol and Additional Results

Sampling and pairing. The score and error distributions are exactly those in Equation (24). Labeled streams use seeds 2026091302 and 2026091303 for $c = 0.24$ and $c = 0.12$, respectively. Independent score-only streams use seeds 2026191302 and 2026191303. Each repetition draws $n = 20{,}000$ scores and conditional Bernoulli errors, and $N = 20{,}000$ additional scores. All rules share these draws. The threshold grid is fixed before sampling, and acceptance uses $S \ge \tau$. The smallest threshold wins a tie between equal coverage certificates. The threshold $\tau = 1$ has population coverage zero and supplies a reject-all fallback.

Matched calibration and cost. For Table 1, every rule uses the independent score-only DKW bound with failure probability $\delta/2$, and each labeled sample supplies the same accepted and error counts. The odds target is q^* and the sharp target is $q^\dagger(\widetilde{p}_k)$. Bonferroni rejects a binomial null at p-value $\delta/(2K)$, while Holm applies step-down correction at familywise level $\delta/2$. Each rule consumes 20,000 labeled score–loss pairs and 20,000 score-only observations. Thus neither envelope nor testing comparison changes the information budget. The main table is reconstructed from per-threshold counts recorded during each draw; all selected-rule summaries are then evaluated against the specified probability distribution.

Table 2: Paired coverage differences for Table 1. Gains and standard errors are in percentage points. The standard error is the sample standard deviation of the 1,000 within-draw differences divided by $\sqrt{1000}$.

Regime	Contrast	Gain	Paired SE	Strict gain draws
$c = 0.24$	Sharp − odds, Bonferroni	0.000	0.000	0/1000
	Sharp − odds, Holm	0.000	0.000	0/1000
	Holm − Bonferroni, odds	.178	.012	178/1000
	Holm − Bonferroni, sharp	.178	.012	178/1000
$c = 0.12$	Sharp − odds, Bonferroni	57.767	.143	1000/1000
	Sharp − odds, Holm	56.426	.150	1000/1000
	Holm − Bonferroni, odds	1.601	.062	405/1000
	Holm − Bonferroni, sharp	.260	.045	43/1000

Population evaluation. For each returned threshold, evaluation uses $p = 1 - \tau$, $q = 0.002 + (c/3)p^2$, $R_\Gamma = \Psi_\Gamma(p, q)$, and $C_\Gamma = L_\Gamma(p)$. Risk and coverage optimize separately over the ambiguity set; they need not have the same extremal distribution. A risk failure is a nonempty returned rule with $R_\Gamma > 0.10$; all draws remain in its denominator. An empty rule has coverage zero and undefined conditional risk. The code uses a 10^{-12} tolerance only when comparing numerical population quantities with a target or confidence bound.

Budget-uniform reference and nominal controls. We also evaluate the certificate in Theorem 3, which uses labeled scores for its coverage band. It differs from the matched Bonferroni rows in Table 1, which use an independent sample. Under the budget-uniform certificate, mean worst-case coverage is 0.24373 for both envelopes at $c = 0.24$, and $0.41893/0.99644$ for odds/sharp at $c = 0.12$. These runs use 20,000 labeled pairs. Nominal empirical calibration selects the lowest threshold with empirical accepted error at most α; nominal CP uses simultaneous Clopper–Pearson bounds at target α without a shift adjustment.

Table 3: Nominal calibration controls. Each row summarizes 1,000 draws; both rules select full acceptance in every draw.

Regime	Rule	R_Γ	C_Γ	$\underline{C}$	Risk failures
$c = 0.24$	Nominal empirical	.16400	1.00000	.97907	1000/1000
	Nominal CP	.16400	1.00000	.97907	1000/1000
$c = 0.12$	Nominal empirical	.08400	1.00000	.97907	0/1000
	Nominal CP	.08400	1.00000	.97907	0/1000

Variability and simultaneous events. The standard deviation of worst-case coverage is 0.00969 for both Holm rules when $c = 0.24$, and 0.04735 and 0.01313 for Holm odds and HOLM–SHARP-RC, respectively, when $c = 0.12$. The Holm familywise risk-failure count is zero for each rule in each regime. The score-only DKW band fails at some grid threshold in 6 and 9 draws for $c = 0.24$ and $c = 0.12$, respectively. For the budget-uniform certificates, grid-wide joint risk-or-coverage failures occur in 6/6 draws for odds/sharp when $c = 0.24$ and 19/20 draws when $c = 0.12$. These are counts out of 1,000, and concern the whole grid. A failure at an unselected threshold does not imply a failure of the returned selector. The reported selected-rule counts therefore address a different event from the grid-wide diagnostics.

Optimization checks and artifacts. The optimization checks include $\Gamma = 1$, zero and unit conditional error, full acceptance, and both normalization boundaries. The bounded-loss comparison uses finite atomic distributions and an independently constructed linear-fractional program. The supplementary implementation records software versions, seeds, exact population definitions, selection histograms, numerical summaries, and confidence limits. Per-draw selected-rule records and per-threshold counts permit reconstruction of the matched comparisons.