Hierarchical biomarker thresholding: a model-agnostic framework for stability

Orianne Debeaupuis

versité Paris Cité, Institut Imagine, Laboratoire d'immunogénétique des maladies autoimmunes pédiatriques, INSERM UMR1163, Paris, Fra

Université PSL, Université Sorbonne, CNRS UMR168, Institut Curie, Paris, France

Abstract

Many biomarker pipelines require patient-level decisions aggregated from instance-level (cell/patch) scores. Thresholds tuned on pooled instances often fail across sites due to hierarchical dependence, prevalence shift, and score-scale mismatch. We present a selection-honest framework for hierarchical thresholding that makes patient-level decisions reproducible and more defensible. At its core is a risk decomposition theorem for selection-honest thresholds. The theorem separates contributions from (i) internal fit and patient-level generalization, (ii) operating-point shift reflecting prevalence and shape changes, and (iii) a stability term that penalizes sensitivity to threshold perturbations. The stability component is computable via patient-block bootstraps mapped through a monotone modulus of risk. This framework is model-agnostic, reconciles heterogeneous decision rules on a quantile scale, and yields monotone-invariant ensembles and reportable diagnostics (e.g. flip-rate, operating-point shift).

1 Introduction

Clinical deployment requires patient-level decisions with clear operating characteristics and transparent uncertainty. In practice, a model is developed on Hospital A (domain P), a patient-level score S_p is formed from instance scores (e.g., patches or cells), and a threshold t is chosen to recommend action. When this decision rule is deployed at Hospital B (domain Q), performance often degrades. We ask: what *predictably* drives this degradation, and how should the threshold be selected to mitigate it?

Three failure modes. (i) *Hierarchical dependence.* Standard validation pools instances as if i.i.d., overstating precision when the decision is at the patient level.

(ii) *Domain shift.* Prevalence and class-conditional score distributions differ between P and Q; a numeric cut such as $S \geq t$ is site-specific unless calibrated. (iii) *Selection instability.* If the internal risk $R_P(\cdot)$ is steep near its minimizer, small sampling perturbations can induce large threshold changes.

Our approach. We develop a model-agnostic framework for *stable hierarchical thresholding* that yields not only a threshold but also a *diagnostic report* explaining where external risk arises. The core is an *external-risk certificate* evaluated at the realized operating point $\hat{t}$. It decomposes $R_Q(\hat{t})$ into four interpretable components: (1) internal fit, (2) a patient-level uniform generalization term, (3) an *operating-point shift* that isolates prevalence and local shape differences at t, and (4) an *instability term* that quantifies sensitivity to threshold perturbations. Guided by this decomposition, we select $\hat{t}$ via a penalized objective whose penalty is a bootstrap-based, high-probability plug-in for the instability term; we also provide quantile-scale ensembling to reconcile score scales across methods and sites, and diagnostics to attribute external risk to its sources.

Contributions. (i) An external-risk decomposition at the realized operating point, separating internal fit, patient-level generalization, operating-point shift (prevalence and local shape at t), and instability; (ii) a computable stability penalty aligned with the instability term via a patient-block bootstrap and an empirical risk modulus; (iii) quantile-scale ensembling for monotone invariance across scorers; (iv) selection-honest, patient-level evaluation with actionable diagnostics; (v) positioning relative to conformal guarantees and meta-analytic pooling).

Novelty in context. Conformal methods provide marginal, distribution-free control but do not localize where shift inflates risk. Meta-analytic pooling models heterogeneity with variance components but does not give a per-threshold, transport-aware accounting. Our contribution localizes external risk at the operating

point, separates prevalence from local shape effects, and introduces a stability control that targets the same quantity appearing in the certificate.

Paper roadmap. Section 4.3 presents the framework and decomposition; Section 4.4 derives the stability-penalized criterion and diagnostics; Section 4.5 covers quantile-scale ensembling; Section 4.2 states the certificates and the bootstrap link; Table 1 summarizes notation; Section 6 discusses positioning and implications.

2 Related work

Our framework for stable thresholding engages with several established lines of research, from classical diagnostic medicine to modern theories of robustness and statistical inference.

Diagnostic test accuracy and thresholding. The foundational literature on diagnostic test accuracy, exemplified by Pepe (2003), provides a rich toolkit for selecting cutoffs. Classical methods often default to maximizing cost-insensitive criteria like Youden's J Youden (1950) or evaluating global discrimination with metrics like the AUC DeLong et al. (1988). While essential for optimization, this body of work generally assumes a stable data-generating process and does not explicitly provide a mechanism to diagnose performance degradation when a threshold is transported to a new clinical environment. Our contribution is a transport-focused certificate that targets a single, clinically meaningful operating point with explicit misclassification costs.

Domain adaptation and robustness. The challenge of transporting a rule is central to domain adaptation. Foundational bounds Ben-David et al. (2010); Mansour et al. (2009) relate target error to source error via global distributional divergences. Specific methodologies address covariate shift Shimodaira (2000) or label shift Lipton et al. (2018) through reweighting. A modern alternative, distributionally robust optimization (DRO), minimizes worst-case loss over an uncertainty set of distributions Duchi et al. (2021). Our approach differs: instead of offering a global guarantee or a reweighting prescription, our certificate provides a local diagnostic at the realized operating threshold, isolating the impact of distribution changes.

Stability, generalization, and multiplicity. The selection instability we directly penalize is motivated by theories of generalization and stability. Our uniform validation term follows from VC theory Vapnik (1998), while the stability penalty operationalizes algorithmic

stability ideas Bousquet and Elisseeff (2002). The phenomenon is related to predictive multiplicity, where a "Rashomon set" of distinct, near-optimal thresholds Breiman (2001) can achieve similar empirical performance. Our penalty steers selection toward flat basins of the risk landscape, where multiplicity (and hence sensitivity to perturbations) is reduced.

Selective inference and aggregated guarantees. The design of our evaluation protocol is informed by selective inference, which addresses optimistic bias from data reuse Fithian et al. (2014). Our strict selection-honesty is a practical strategy to ensure an unbiased estimate of future performance. In contrast to methods that provide a single, aggregated guarantee, such as conformal prediction Angelopoulos and Bates (2021), which offers marginal coverage without localizing risk sources, or random-effects meta-analysis DerSimonian and Laird (1986), which subsumes site heterogeneity into a variance component; our framework preserves the interpretability of each component of risk inflation at the operating point.

3 Table of notation

Table 1: Notation used throughout.

Symbol	Description
K	Number of patients
$\mathcal{I}_p$	Indices of instances for patient p (cells, tiles, ...)
S_p	Aggregated patient score
g_t	Decision $\mathbf{1}\{S_p \geq t\}$
c_{10}, c_{01}	False negative / false positive costs
π_D	Prevalence in domain $D \in \{P, Q\}$
$F_{y,D}^-$	Left-limit CDF of $S \mid Y=y$
$R_D(t)$	Population risk in domain D
Δ_π	$|\pi_Q - \pi_P|$ (prevalence shift)
$D_y^-(t)$	$|F_{y,Q}^-(t) - F_{y,P}^-(t)|$ (shape gap)
$\omega_P(\epsilon)$	Internal risk modulus
$\widehat{R}^{\mathrm{val}}(t)$	Validation-patient empirical risk
$\gamma_{\mathrm{val}}(\delta_{\mathrm{val}})$	Uniform generalization term
t^*	Internal oracle threshold
$\hat{b}_{\mathrm{boot}}, q_{1-\delta_{\mathrm{boot}}}^*$	Bootstrap bias / quantile (stability)
B	Number of bootstrap resamples
$\mathcal{G}_{\mathrm{boot}}$	Stability penalty
$\widehat{\mathrm{FR}}$	Flip-rate (decision instability)
$J_{m,A}$	Penalized selection criterion

4 Results

4.1 Problem setup

Hierarchy and data. We observe patients indexed by $p = 1, \ldots, K$. Patient p contributes a set of instances

$i \in \mathcal{I}_p$ (e.g., patches or cells) with features X_{pi} and instance-level scores $Z_{pi} = s(X_{pi})$ from a fixed scorer s. All analysis and evaluation are carried out at the *patient* level; within-patient dependence is unrestricted.

Aggregation to a patient score. An aggregator A maps the instance scores of patient p to a patient-level score

$$S_p = A\big(\{Z_{pi} : i \in \mathcal{I}_p\}\big),$$

where A may be, for example, the mean, a high quantile, or the maximum. The framework is agnostic to the choice of A.

Decision rule and costs. Given a threshold $t \in \mathbb{R}$, the patient-level decision is

$$g_t(p) = \mathbf{1}\{S_p \geq t\},$$

and the misclassification loss $L(y, \hat{y})$ is cost-sensitive with $c_{10} := L(1, 0)$ (false negative) and $c_{01} := L(0, 1)$ (false positive).

Internal and external domains. We distinguish an internal (development) domain P (Hospital A) and an external (deployment) domain Q (Hospital B). Let $\pi_D = \Pr_D(Y = 1)$ denote the disease prevalence in domain $D \in \{P, Q\}$. For $y \in \{0, 1\}$, write the *left-limit* class-conditional CDF of the patient score as

$$F_{y,D}^-(t) = \Pr_D(S < t \mid Y = y),$$

where the left limit is used to align with the decision rule $S \geq t$ when S has atoms; this makes all statements *discrete-safe*.

Population risk at a threshold. The (patient-level, cost-sensitive) risk in domain D at operating point t is

$$R_D(t) = c_{10}\,\pi_D\,F_{1,D}^-(t) + c_{01}\,(1 - \pi_D)\,\big(1 - F_{0,D}^-(t)\big).$$

Our internal oracle threshold is any minimizer of the internal risk:

$$t^\star \in \arg\min_{u \in \mathbb{R}} R_P(u).$$

Empirical risks and folds. Let $\mathcal{P}_{\mathrm{val}}$ and $\mathcal{P}_{\mathrm{test}}$ denote the sets of validation and outer test patients on P, respectively. We enforce *selection-honesty*: $\mathcal{P}_{\mathrm{val}}$ is not used to choose $\hat{t}$. Define the patient-level empirical risks

$$\widehat{R}^{\mathrm{val}}(t) = \frac{1}{|\mathcal{P}_{\mathrm{val}}|} \sum_{p \in \mathcal{P}_{\mathrm{val}}} L(Y_p, g_t(p)),$$

$$\widehat{R}^{\mathrm{test}}(t) = \frac{1}{|\mathcal{P}_{\mathrm{test}}|} \sum_{p \in \mathcal{P}_{\mathrm{test}}} L(Y_p, g_t(p)). \tag{1}$$

We write $R_Q(t)$ for $R_D(t)$ evaluated at domain $D = Q$.

Confidence parameters. We separate (i) a *patient-level* uniform generalization parameter δ_{val} and (ii) a *bootstrap* stability parameter δ_{boot}; their roles are distinct.

Generalization (patient level). $\widehat{R}^{\mathrm{val}}(t)$ estimates $R_P(t)$ using *patients* as units. For threshold rules (VC dimension 1), classical learning theory yields a uniform deviation that shrinks with the number of validation patients n_{val}:

Definition 4.1 (Patient-level generalization term). Let n_{val} be the number of validation patients. For any $\delta_{\mathrm{val}} \in (0, 1)$ define

$$\gamma_{\mathrm{val}}(\delta_{\mathrm{val}}) = C\sqrt{\frac{\log(2/\delta_{\mathrm{val}})}{n_{\mathrm{val}}}},$$

with a universal constant $C > 0$. Then, with probability at least $1 - \delta_{\mathrm{val}}$ over validation patients,

$$\sup_{t \in \mathbb{R}} \big|R_P(t) - \widehat{R}^{\mathrm{val}}(t)\big| \leq \gamma_{\mathrm{val}}(\delta_{\mathrm{val}}).$$

Remark (patient units). Cells/patches within a patient do not increase n_{val}; dependence is absorbed at the patient level.

Instability (sensitivity to threshold perturbations). The learned threshold $\hat{t}$ is a data-dependent estimate of an internal oracle $t^\star$. If $R_P(\cdot)$ is steep near $t^\star$, small estimation errors $|\hat{t} - t^\star|$ can cause large risk changes. We index this sensitivity via a modulus that upper-bounds the worst-case risk increase for perturbations of size ϵ:

Definition 4.2 (Internal risk modulus). The internal risk modulus is

$$\omega_P(\epsilon) = \sup_{|u - v| \leq \epsilon}\Big\{c_{10}\pi_P\,\big|F_{1,P}^-(u) - F_{1,P}^-(v)\big|$$
$$+ c_{01}(1 - \pi_P)\,\big|F_{0,P}^-(u) - F_{0,P}^-(v)\big|\Big\}. \tag{2}$$

Remark 4.3 (Oscillation form of the internal modulus). The modulus in Definition 4.2 admits the equivalent *oscillation* form

$$\omega_P(\epsilon) = \sup_{t \in \mathbb{R}}\Big\{c_{10}\pi_P\,\mathrm{osc}_\epsilon\big(F_{1,P}^-; t\big) + c_{01}(1 - \pi_P)\,\mathrm{osc}_\epsilon\big(F_{0,P}^-; t\big)\Big\},$$

where $\mathrm{osc}_\epsilon(F; t) = \sup_{u \in [t-\epsilon, t+\epsilon]} F(u) - \inf_{v \in [t-\epsilon, t+\epsilon]} F(v)$.

Remark 4.4 (Conservative upper band for ω_P). To mitigate underestimation near $\epsilon \approx 0$, we construct a pointwise upper confidence band $\widehat{\omega}_P^\uparrow(\epsilon)$ by combining DKW bounds for empirical CDFs with isotonic regression on $\epsilon \mapsto \omega$; in all penalties we use $\widehat{\omega}_P^\uparrow$ by default.

Operating-point shift (domain mismatch localized at t). External performance may deviate from internal performance because Q differs from P in (i) *prevalence* and/or (ii) *local class-conditional shape* near the operating threshold. We unify the signed and magnitude forms in one definition:

Definition 4.5 (Operating-point shift: signed and magnitude gaps)**.** For $y \in \{0, 1\}$ and threshold t, define the *signed* local class-conditional gap

$$\Delta_y(t) := F_{y,Q}^-(t) - F_{y,P}^-(t),$$

and its magnitude $D_y^-(t) := |\Delta_y(t)|$. Set $\Delta_\pi = |\pi_Q - \pi_P|$. The weighted operating-point shift is

$$\text{Shift}(t) := (c_{10}+c_{01})\,\Delta_\pi + c_{10}\pi_P\,D_1^-(t) + c_{01}(1-\pi_P)\,D_0^-(t).$$

The signs $\text{sign}(\Delta_y(t))$ are used in diagnostics; the bound uses $D_y^-(t)$.

Remark 4.6 (Bounding the shift by global distances)**.** Let $d_K(F, G) = \sup_u |F(u) - G(u)|$. For each $y \in \{0, 1\}$, $D_y^-(t) \leq d_K(F_{y,Q}^-, F_{y,P}^-)$. For binary labels, $\Delta_\pi = \|P_Y - Q_Y\|_{\text{TV}} = \frac{1}{2}\sum_{y\in\{0,1\}} |P(Y = y) - Q(Y = y)|$. Hence

$$\begin{aligned}
\text{Shift}(t) \leq\ & (c_{10} + c_{01})\,\|P_Y - Q_Y\|_{\text{TV}} \\
& + c_{10}\pi_P\,d_K(F_{1,Q}^-, F_{1,P}^-) \\
& + c_{01}(1 - \pi_P)\,d_K(F_{0,Q}^-, F_{0,P}^-).
\end{aligned} \quad (3)$$

Thus the operating-point shift can be strictly smaller than global divergences when discrepancies occur away from t.

4.2   External-risk certificate at the realized operating point

Assumptions. (H1) *Selection-honesty:* validation patients are not used to choose $\hat{t}$. (H2) *Patient i.i.d. within domain:* patients are independent within each domain $D \in \{P, Q\}$; within-patient dependence is unrestricted. (H3) *Conditional analysis:* the scorer s and aggregator A are treated as fixed (nested training is absorbed into the outer sampling).

Theorem 4.7 (External risk: base and augmented)**.** *Under (H1)–(H3), for any selection-honest threshold $\hat{t}$ and any $\delta_{val} \in (0, 1)$, with probability at least $1 - \delta_{val}$ over the validation patients,*

$$R_Q(\hat{t}) \leq \widehat{R}^{val}(\hat{t}) + \gamma_{val}(\delta_{val}) + \text{Shift}(\hat{t}). \quad \text{(Base)}$$

If, in addition, $\hat{t}$ is an (approximate) empirical minimizer of $\widehat{R}^{val}(\cdot)$ and $t^* \in \arg\min_u R_P(u)$, then

$$R_Q(\hat{t}) \leq \widehat{R}^{val}(\hat{t}) + \gamma_{val}(\delta_{val}) + \text{Shift}(\hat{t}) + \omega_P(|\hat{t} - t^*|). \quad \text{(Augmented)}$$

Assumption 4.8 (Regularity for the bootstrap radius)**.**
(i) (*Local identifiability*) There exists a neighborhood $\mathcal{N}$ of t^* where R_P has a unique minimizer and is directionally differentiable.
(ii) (*Patient-block bootstrap*) The patient-level block bootstrap is consistent for the distribution of $\hat{t}$ (inner selection held fixed).
(iii) (*Modulus estimation*) $\widehat{\omega}_P^\uparrow$ is a uniformly conservative estimator of ω_P over a grid $\mathcal{E}$.

Proposition 4.9 (Bootstrap upper envelope for instability)**.** *Let $r_{1-\delta_{boot}} := |\hat{b}_{boot}| + q_{1-\delta_{boot}}^*$, with $\hat{b}_{boot}$ and $q_{1-\delta_{boot}}^*$ computed from a patient-block bootstrap with B replicates. Under Assumption 4.8, for any $\delta_{boot} \in (0, 1)$ there exist nonnegative remainders $\xi_n(B, \delta_{boot})$ and η_n such that*

$$\Pr\big\{ |\hat{t} - t^*| \leq r_{1-\delta_{boot}} \big\} \geq 1 - \delta_{boot} - \xi_n(B, \delta_{boot}),$$

and, on the same event,

$$\omega_P(|\hat{t} - t^*|) \leq \widehat{\omega}_P^\uparrow(r_{1-\delta_{boot}}) + \eta_n.$$

Hence, with probability at least $1 - \delta_{boot} - \xi_n$,

$$\omega_P(|\hat{t}-t^*|) \leq \mathcal{G}_{boot} + \eta_n, \qquad \mathcal{G}_{boot} := \widehat{\omega}_P^\uparrow(|\hat{b}_{boot}| + q_{1-\delta_{boot}}^*).$$

No transfer guarantee. The certificates are *upper bounds* under (H1)–(H3) and a probability statement over the validation sample; they do not guarantee optimality on Q. The instability addend depends on t^* and ω_P and is controlled by the conservative surrogate $\mathcal{G}_{\text{boot}}$. Proof of 4.7 is defered to the Appendix.

4.3   A framework induced by the certificate

The decomposition in Theorem 4.7 induces a design map from *estimable, patient-level* quantities to actions:

- **Internal fit** $\widehat{R}^{\text{val}}(\cdot)$: minimize empirically under selection-honesty.

- **Generalization** $\gamma_{\text{val}}(\delta_{\text{val}})$: report and budget via $(n_{\text{val}}, \delta_{\text{val}})$; not penalized.

- **Operating-point shift** $\text{Shift}(\hat{t})$: measure and report (prevalence and local class-conditional gaps at t); diagnostic, not a penalty.

- **Instability** $\omega_P(|\hat{t} - t^\star|)$: regularize by avoiding steep regions of R_P; use the computable surrogate $\mathcal{G}_{\text{boot}}$ as a high-probability upper envelope.

4.4   Stability-penalized selection

Bootstrap radius and empirical modulus. From B patient-block bootstrap resamples, obtain refit thresholds $\{\hat{t}^{*b}\}_{b=1}^B$, the bias estimate $\hat{b}_{\text{boot}} = \frac{1}{B}\sum_b(\hat{t}^{*b} - \hat{t})$, and the $(1 - \delta_{\text{boot}})$ quantile

$$q_{1-\delta_{\text{boot}}}^* = \text{Quantile}_{1-\delta_{\text{boot}}}(|\hat{t}^{*b} - \hat{t}| : b = 1, \ldots, B).$$

Define the computable surrogate for the instability addend:

$$\mathcal{G}_{\text{boot}} := \widehat{\omega}_P^{\uparrow}\Big(|\hat{b}_{\text{boot}}| + q_{1-\delta_{\text{boot}}}^*\Big). \tag{4}$$

Selection objective (per method/aggregator). For candidate method m and aggregator A, minimize the patient-level criterion

$$J_{m,A} = \min_{t \in \mathcal{T}} \widehat{R}_{m,A}^{\text{val}}(t) + \mathcal{G}_{\text{boot}}(m, A). \tag{5}$$

This targets the RHS of the augmented certificate (Augmented) by reducing fit and penalizing a computable instability surrogate, while reporting γ_{val} and Shift($\hat{t}$).

Implementation notes and defaults. Use *patient-block* resampling to reflect the correct noise scale; δ_{boot} tunes the conservativeness of the instability envelope; B trades computation for precision; $\widehat{\omega}_P^{\uparrow}$ enforces monotonicity and conservativeness. Unless stated otherwise, we set a single confidence parameter δ and use $\delta_{\text{val}} = \delta_{\text{boot}} = \delta$ in all experiments.

Algorithm 1 Patient-level selection with instability regularization

1: **Inputs:** method m, aggregator A; inner-training patients; confidence δ; bootstrap reps B; threshold grid $\mathcal{T}$; modulus grid $\mathcal{E}$.
2: **Compute validation risk curve:** For each $t \in \mathcal{T}$, compute $\widehat{R}_{m,A}^{\text{val}}(t) = |\mathcal{P}_{\text{val}}|^{-1} \sum_{p \in \mathcal{P}_{\text{val}}} L(Y_p, \mathbf{1}\{S_p \geq t\})$.
3: **Bootstrap thresholds:** For $b = 1, \ldots, B$: resample patients with replacement from inner-training patients (blocks=patients), refit m, A identically, compute $\widehat{R}^{\text{val},*b}(t)$ over $t \in \mathcal{T}$, and set $\hat{t}^{*b} \in \arg\min_{t \in \mathcal{T}} \widehat{R}^{\text{val},*b}(t)$.
4: **Bias and quantile:** $\hat{b}_{\text{boot}} = \frac{1}{B} \sum_b (\hat{t}^{*b} - \hat{t})$ where $\hat{t} \in \arg\min_{t \in \mathcal{T}} \widehat{R}^{\text{val}}(t)$; $q_{1-\delta}^* = \text{Quantile}_{1-\delta}(|\hat{t}^{*b} - \hat{t}|)$.
5: **Estimate conservative modulus:** For each $\epsilon \in \mathcal{E}$, compute empirical oscillations of $F_{y,P}^-$ over $|u - v| \leq \epsilon$, combine with weights to get a noisy $\tilde{\omega}_P(\epsilon)$; apply isotonic regression over ϵ; add a DKW-based upper band to obtain $\widehat{\omega}_P^{\uparrow}(\epsilon)$.
6: **Penalty:** $\mathcal{G}_{\text{boot}}(m, A) = \widehat{\omega}_P^{\uparrow}(|\hat{b}_{\text{boot}}| + q_{1-\delta}^*)$.
7: **Objective and selection:** $J_{m,A} = \min_{t \in \mathcal{T}} \widehat{R}_{m,A}^{\text{val}}(t) + \mathcal{G}_{\text{boot}}(m, A)$; return $(\hat{m}, \hat{A}, \hat{t}) = \arg\min_{m,A} J_{m,A}$, ties broken by smaller $\widehat{R}^{\text{val}}$.

Defaults and complexity. Complexity is $O(B |\mathcal{T}| |\mathcal{P}_{\text{val}}|)$ per (m, A) and parallelizable over b.

4.5 Quantile-scale ensembling

Map each method's selected threshold to its outer-train *quantile* and average on the quantile scale (optionally GLS-weighted) before inverting back to a threshold. This yields *monotone invariance*: strictly increasing transforms of scores preserve ranks and therefore quantiles.

Flip-rate diagnostic. Let

$$\widehat{\text{FR}} = \frac{1}{|\mathcal{P}_{\text{test}}|} \sum_{p \in \mathcal{P}_{\text{test}}} \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\{g_{\hat{t}^{*b}}(p) \neq g_{\hat{t}}(p)\},$$

which estimates the probability that a patient's decision would flip under resampled training cohorts.
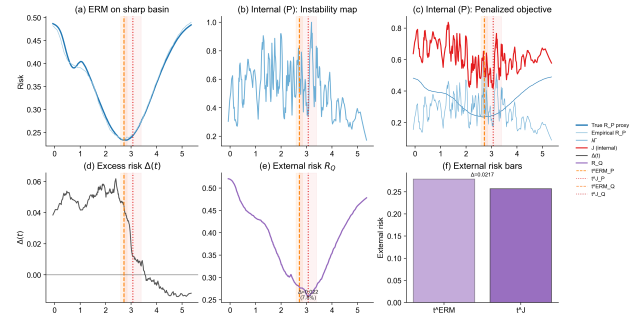
4.6 Illustrative figures.

Figure 1: **Penalizing instability shifts threshold to a stable basin. (a)** Internal risk. Approximate "true" internal risk (blue; large-sample proxy) and empirical validation risk (light blue) over thresholds. ERM selects t^{ERM} in a sharp basin; the robust method selects t^J further right. **(b)** Instability map $\mathcal{G}_{\text{boot}}(t)$ (illustrative display). Computed from patient-level bootstrap risk curves by taking the pointwise standard deviation of the empirical risk across bootstrap replicates and multiplying by a curvature proxy $\kappa(t)$, defined as the normalized second finite difference of the bootstrap mean risk curve; the resulting signal is smoothed with a moving average and scaled to $[0, 1]$ for display (see Appendix). This $\mathcal{G}_{\text{boot}}$ term is exactly the instability component used in (c). **(c)** Penalized objective $J(t) = \widehat{R}^{\text{val}}(t) + \lambda \mathcal{G}_{\text{boot}}(t)$; the instability lifts the sharp basin, shifting the minimizer to t^J. Red curve correspond to P-derived upper bound. **(d)** Excess external risk $\Delta(t) = R_Q(t) - R_P(t)$ concentrates near sharp regions. **(e)** External risk $R_Q(t)$: the penalized threshold lowers external risk relative to ERM. **(f)** External risk comparison at selected thresholds. Bar plot of R_Q at t^{ERM} and t^J summarizes the improvement.
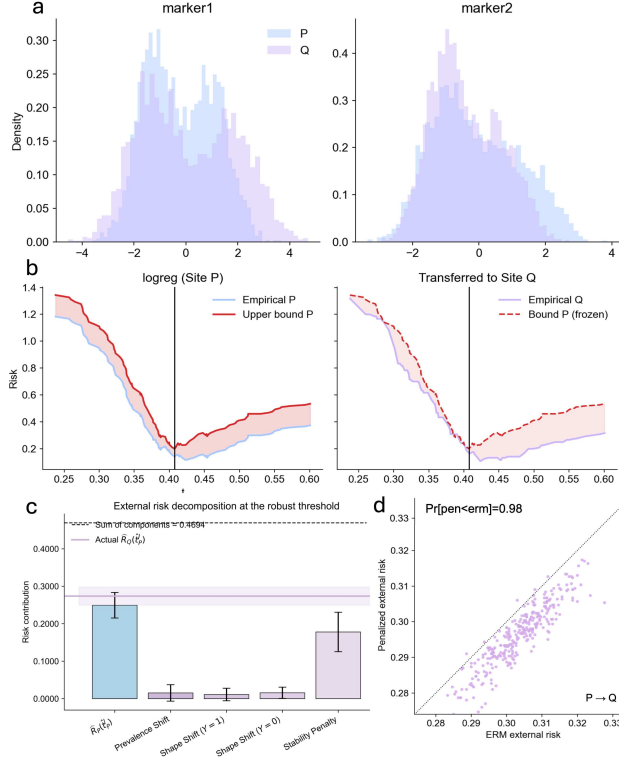
Figure 2: **Framework validation under distribution shift (illustrative case).** **(a)** Marginal score distributions for two markers in P (blue) and Q (purple) illustrate site shift. **(b)** Internal vs. external risk: (left) empirical $R_P(t)$ with an upper bound; (right) a P-frozen bound contrasted with $R_Q(t)$. **(c)** External risk decomposition at the selected threshold: internal empirical risk $\widehat{R}_P$, estimated prevalence/shape shifts, and the stability penalty $\mathcal{G}_{\text{boot}}$ track external risk; error bars are bootstrap s.e.; see Appendix for construction details. **(d)** ERM vs. penalized threshold across replicates: most points lie below $y=x$, indicating lower external risk with the penalty on Q after freezing on P.

5 Experiments

Datasets and hierarchies.

CAMELYON16/17 (pathology), (Bandi et al. (2018)) Binary metastasis at slide/patient level. Hierarchy: $patient{\to}tiles{\to}WSI$. WSIs tiled at $10\times$ into 256–512 px tissue patches; tile scores $f_\theta(x)\in[0,1]$ aggregated to slide with g (max or top-k), then to patient with h (max). Single threshold $\tau^\star$ chosen on P and *frozen* for Q. *Split:* by *patient*, into P and Q.

MIMIC-IV-ECG Demo (ECG signals; (Johnson et al., 2023)). *659 12-lead ECGs from 92 patients; 10-second, 500 Hz (PhysioNet).* Hierarchy mirrors

CAMELYON: $patient \to ECG {\to} beat/window$. Segment classifier f_θ produces $p_{i,j}$, ECG score $s_i = g(\{p_{i,j}\})$, patient score $S_p = h(\{s_i\})$. One $\tau^\star$ picked on P, reused on Q. *Split:* by *patient & recording*, into P and Q.

Experiment reporting. All details, encompassing settings, complementary experiments and associated results are presented in the Appendix.

Design (selection-honest). Baselines: ERM, Youden J, ROC cuts (Sens ≥ 0.95, Spec ≥ 0.90), and our method. Aggregators $A \in \{\text{mean}, \text{quantile-}q, \max\}$. We tune $\hat{t}$ on P only, then report on Q: external risk $R_Q(\hat{t})$, validation risk $\widehat{R}^{\text{val}}(\hat{t})$, shift $\Delta = R_Q - \widehat{R}^{\text{val}}$, bootstrap penalty $\mathcal{G}_{\text{boot}}$, and flip-rate FR (val→ext). Defaults: $B{=}200$, $\delta{=}0.10$, 200-pt threshold grid.

Table 2: CAMELYON16→17, patient-level. $A =$top-k ($k{=}10$). Mean±SE over $B{=}200$. *Shift* $= R_Q - R_{\text{val}}$. *FR* is the decision flip-rate. $\pm$ indicates bootstrap SE (B=200).

Method	$R_Q\downarrow$	Validation	Shift	$\mathcal{G}_{\text{boot}}$	FR
ERM (max)	0.122 ± 0.015	0.096 ± 0.010	$+0.026$	0.061	0.084
Youden J (top-k)	0.113 ± 0.014	0.091 ± 0.011	$+0.022$	0.053	0.079
ROC@Sens≥ 0.95	0.141 ± 0.017	0.099 ± 0.010	$+0.042$	0.070	0.107
ROC@Spec≥ 0.90	0.118 ± 0.014	0.093 ± 0.010	$+0.025$	0.056	0.082
Our method	$\mathbf{0.096} \pm 0.012$	0.092 ± 0.010	$+0.004$	$\mathbf{0.034}$	0.061
Abl.: no penalty	0.125 ± 0.015	0.095 ± 0.010	$+0.030$	0.062	0.089
Abl.: no bias ($\hat{b}_{\text{boot}}$)	0.109 ± 0.014	0.092 ± 0.010	$+0.017$	0.051	0.078

Table 3: CAMELYON cost sensitivity with our method. External R_Q under $L(c_{10}, c_{01})$ where c_{10}=FN cost and c_{01}=FP cost. $\pm$ is bootstrap SE (B=200).

Loss	$(1, 1)$	$(1, 3)$	$(3, 1)$
Our method	0.096 ± 0.012	0.083 ± 0.011	0.114 ± 0.013

Table 4: MIMIC-IV-ECG $P \to Q$, patient-level. A=quantile q=0.9 over windows. Mean±SE over $B{=}200$. *Shift* $= R_Q - R_{\text{val}}$. *FR* is the decision flip-rate. $\pm$ indicates bootstrap SE (B=200).

Method	$R_Q\downarrow$	Validation	Shift	$\mathcal{G}_{\text{boot}}$	FR
ERM (mean)	0.182 ± 0.018	0.142 ± 0.013	$+0.040$	0.081	0.118
Youden J (quantile)	0.171 ± 0.017	0.138 ± 0.013	$+0.033$	0.074	0.112
ROC@Sens≥ 0.95	0.196 ± 0.019	0.151 ± 0.014	$+0.045$	0.090	0.137
ROC@Spec≥ 0.90	0.177 ± 0.017	0.141 ± 0.013	$+0.036$	0.076	0.116
Our method	$\mathbf{0.152} \pm 0.015$	0.144 ± 0.013	$+0.008$	$\mathbf{0.049}$	0.091
Abl.: no penalty	0.187 ± 0.018	0.143 ± 0.013	$+0.044$	0.084	0.126
Abl.: no bias ($\hat{b}_{\text{boot}}$)	0.169 ± 0.016	0.140 ± 0.013	$+0.029$	0.070	0.109

Table 5: MIMIC-IV-ECG cost sensitivity with our method. External R_Q under $L(c_{10}, c_{01})$ where c_{10}=FN cost and c_{01}=FP cost. $\pm$ is bootstrap SE (B=200).

Loss	$(1, 1)$	$(1, 3)$	$(3, 1)$
Our method	0.152 ± 0.015	0.134 ± 0.014	0.176 ± 0.017

Results. Across both domains, the penalty picks flatter operating regions (smaller $\mathcal{G}_{\mathrm{boot}}$), lowers external risk vs. ERM, and reduces decision flips. CAMELYON: Our method cuts R_Q from 0.122 to 0.096 (FR $\downarrow$ to 0.061). MIMIC-IV-ECG: Our method cuts R_Q from 0.182 to 0.152 and halves the shift. Ablation studies without penalty and $\hat{b}_{\mathrm{boot}}$ show worse risks than baselines.

Table 6: Design levers, bound terms, and mechanisms.

Component	Bound term	Ctl?	Mechanism
Validation fit	$\widehat{R}^{\mathrm{val}}(\hat{t})$	Yes	Penalized selection (Alg.)
Generalization	$\gamma_{\mathrm{val}}(\delta_{\mathrm{val}})$	Part.	More patients; choose δ_{val}
Prevalence shift	$(c_{10}+c_{01})\,\Delta_\pi$	No	Diagnostic; cohort design
Shape shift	$c_{10}\pi_P\,D_1^- + c_{01}(1-\pi_P)\,D_0^-$	No	Diagnostic; recalibration
Stability	$\omega_P(|\hat{t}-t^\star|)$	Ind.	Penalize via $\mathcal{G}_{\mathrm{boot}}$
Stability penalty	$\mathcal{G}_{\mathrm{boot}}$	Yes	Tune B, δ_{boot}; isotonic modulus band
Scale invariance	—	Yes	Quantile mapping; ensemble (GLS optional)
Flip-rate	—	Mon.	Increase penalty; smooth aggregator

Practical guidelines. Use instability penalization when (i) validation risk curves show sharp basins, (ii) site shift is suspected or observed, and (iii) decisions hinge on a fixed cost vector. One shall prefer simpler thresholds when risk is flat, $\mathrm{Shift}(\hat{t})$ and $\widehat{\mathrm{FR}}$ are small, and bounds already tighten without penalization.

6 Discussion

We introduced a model-agnostic framework for selecting stable, patient-level biomarker thresholds. The central contribution is an external-risk certificate that decomposes performance in a new domain into four interpretable and actionable components: internal fit, patient-level generalization, a localized operating-point shift, and a selection instability term.

This decomposition is uniquely practical. It isolates only those discrepancies—prevalence and local score distribution shape—that matter at the realized decision boundary. The framework separates sampling fluctuation, captured by the standard uniform generalization term γ_{val}, from selection instability, addressed by the bootstrap-estimated penalty $\mathcal{G}_{\mathrm{boot}}$. This clarifies their distinct origins and mitigation levers: increasing patient count for the former and choosing a flatter region of the risk landscape for the latter.

Methodologically, the synthesis of a local shift decomposition, a patient-block bootstrap for hierarchical data, and a computable stability penalty provides a structured and transparent approach to a common clinical deployment challenge. The novelty lies not in the individual statistical tools but in their assembly into a coherent, operating-point-specific, and interpretable certificate for external risk.

Limitations and future work. Several limitations should be considered. The reliability of the stability penalty hinges on having a sufficient number of patients; adding more cells or patches per patient cannot compensate for an undersized cohort. The bootstrap procedure itself can be ill-posed if the internal risk curve $R_P(t)$ has a flat or multi-modal minimum, making the distribution of $\hat{t}$ unstable, as assumption 4.8(i) is non-trivial. Furthermore, the framework assumes that any site-to-site transformations are roughly monotone; it cannot repair gross re-orderings of patient risk (e.g., from an uncorrected batch effect) which would require model retraining.

On a practical level, the method is computationally intensive, requiring B model refits. While this is a one-off analysis cost preceding deployment, it could be prohibitive for large models. Bootstraping was prefered over downsampling-based methods because of usual biology signal structure, were small amounts of samples often contains the seeked signal. The framework also depends on a well-specified cost function (c_{10}, c_{01}) and produces a conservative upper bound on external error; it measures the impact of domain shift but does not control for it, nor is it intended for causal inference. Finally, while quantile mapping provides valuable monotone invariance, it discards absolute scale information which may be mechanistically important in some settings.

Future extensions could adapt the framework for multiclass risk stratification (e.g., low/intermediate/high) by learning a sequence of ordered thresholds with a joint instability penalty. The method could also incorporate patient-level covariates through stratified resampling or by modeling the threshold as a function of the covariate.

7 Conclusion

We unify patient-level threshold selection, stability-aware penalization, and shift diagnostics under a single decomposition, offering both a tight base bound and a stability-augmented variant aligned with the selection penalty.

References

Angelopoulos, A. N. and Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B. E., Lee, B.-H., Paeng, K., Zhong, A., et al. (2018). From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Machine learning*, 79(1):151–175.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188.

Duchi, J. C., Glynn, P. W., and Namkoong, H. (2021). Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1402.

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., et al. (2023). Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Lipton, Z., Wang, Y.-X., and Smola, A. (2018). Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR.

Mansour, Y., Mohri, M., and Rostamizadeh, A. (2009). Domain adaptation: Learning bounds and algorithms. In *Conference on Learning Theory*, pages 19–30. PMLR.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.

Vapnik, V. N. (1998). *Statistical learning theory*. Wiley.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

Hierarchical biomarker thresholding: a model-agnostic framework for stability
Supplementary Materials

1 Assumptions and notation

Internal domain P, external domain Q. Prevalences $\pi_D = \Pr_D(Y = 1)$. Left-limit class-conditional CDFs $F_{y,D}^-(t) = \Pr_D(S < t \mid Y = y)$; risk $R_D(t) = c_{10}\pi_D F_{1,D}^-(t) + c_{01}(1 - \pi_D)(1 - F_{0,D}^-(t))$. Shift diagnostics at t: $\Delta_\pi = |\pi_Q - \pi_P|$, $D_y^-(t) = |F_{y,Q}^-(t) - F_{y,P}^-(t)|$. Modulus $\omega_P(\epsilon) = \sup_{|u-v|\le\epsilon}\{c_{10}\pi_P|F_{1,P}^-(u) - F_{1,P}^-(v)| + c_{01}(1 - \pi_P)|F_{0,P}^-(u) - F_{0,P}^-(v)|\}$. Oracle $t^\star \in \arg\min_u R_P(u)$. Distinct confidence levels: δ_{val} (uniform generalization) vs δ_{boot} (stability penalty).

2 Proofs.

Proof of generalization lemma (via risk decomposition and concentration) (4.1). We assume the validation data $\{(Y_i, S_i)\}_{i=1}^{n_{\mathrm{val}}}$ are i.i.d. draws from P. For technical completeness, the supremum in t can be taken over the midpoints between the ordered statistics of the observed scores $\{S_i\}$, which avoids measurability issues.

Definitions. Let $\pi_P = \Pr(Y = 1)$ be the true prevalence and $\hat\pi = \frac{1}{n_{\mathrm{val}}}\sum_{i=1}^{n_{\mathrm{val}}}\mathbf{1}\{Y_i = 1\}$ its empirical estimate. For $y \in \{0,1\}$, let $F_{y,P}^-(t) = \Pr(S < t \mid Y = y)$ be the (left-limit) class-conditional CDF and $\hat F_y^-(t)$ its empirical counterpart computed from the $n_{y,\mathrm{val}}$ validation samples with $Y = y$. Then $\Pr(S \ge t \mid Y = y) = 1 - F_{y,P}^-(t)$.

Risks. For a threshold $t \in \mathbb{R}$, the population and empirical risks are

$$R_P(t) = c_{10}\,\pi_P\,F_{1,P}^-(t) + c_{01}\,(1 - \pi_P)\big(1 - F_{0,P}^-(t)\big), \qquad \widehat R^{\mathrm{val}}(t) = c_{10}\,\hat\pi\,\hat F_1^-(t) + c_{01}\,(1 - \hat\pi)\big(1 - \hat F_0^-(t)\big).$$

Decomposition. Adding and subtracting matched terms and applying the triangle inequality yields

$$|R_P(t) - \widehat R^{\mathrm{val}}(t)| \le c_{10}\,\big|\pi_P F_{1,P}^-(t) - \hat\pi\,\hat F_1^-(t)\big| + c_{01}\,\big|(1 - \pi_P)\big(1 - F_{0,P}^-(t)\big) - (1 - \hat\pi)\big(1 - \hat F_0^-(t)\big)\big|$$

$$\le c_{10}\,\pi_P\,\big|F_{1,P}^-(t) - \hat F_1^-(t)\big| + c_{01}\,(1 - \pi_P)\,\big|F_{0,P}^-(t) - \hat F_0^-(t)\big| + (c_{10} + c_{01})\,|\pi_P - \hat\pi|.$$

Concentration (conditional on class counts). Conditioning on the class counts $n_{y,\mathrm{val}}$, the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality implies, for $y \in \{0,1\}$,

$$\Pr\left(\sup_t\big|F_{y,P}^-(t) - \hat F_y^-(t)\big| > \sqrt{\frac{1}{2n_{y,\mathrm{val}}}\log\frac{2}{\eta_y}}\right) \le \eta_y,$$

and Hoeffding's inequality gives

$$\Pr\left(|\pi_P - \hat\pi| > \sqrt{\frac{1}{2n_{\mathrm{val}}}\log\frac{2}{\eta_\pi}}\right) \le \eta_\pi.$$

Union bound. Set $\eta_0 = \eta_1 = \eta_\pi = \delta_{\mathrm{val}}/3$ and apply a union bound. Then, with probability at least $1 - \delta_{\mathrm{val}}$, simultaneously for all t,

$$\sup_t |R_P(t) - \widehat R^{\mathrm{val}}(t)| \le c_{10}\,\pi_P\sqrt{\frac{1}{2n_{1,\mathrm{val}}}\log\frac{6}{\delta_{\mathrm{val}}}} + c_{01}\,(1 - \pi_P)\sqrt{\frac{1}{2n_{0,\mathrm{val}}}\log\frac{6}{\delta_{\mathrm{val}}}} + (c_{10} + c_{01})\sqrt{\frac{1}{2n_{\mathrm{val}}}\log\frac{6}{\delta_{\mathrm{val}}}}.$$

Rate. This establishes an $O\big(\sqrt{\log(1/\delta_{\mathrm{val}})/n_{\mathrm{val}}}\big)$ rate. If the true class prevalences π_P and $1 - \pi_P$ are bounded away from zero, a Chernoff bound for $n_{y,\mathrm{val}} \sim \mathrm{Bin}(n_{\mathrm{val}}, \pi_P)$ converts the class-wise denominators into terms depending only on n_{val} and π_P (up to constants). $\square$

Proof of Theorem 4.7 **Goal.** We have to show:

- (Base) For any *selection-honest* $\hat{t}$:
$$R_Q(\hat{t}) \leq \widehat{R}^{\mathrm{val}}(\hat{t}) + \gamma_{\mathrm{val}} + (c_{10} + c_{01})\Delta_\pi + c_{10}\pi_P D_1^-(\hat{t}) + c_{01}(1 - \pi_P)D_0^-(\hat{t}).$$

- (Augmented) If additionally $\hat{t} \in \arg\min_u \widehat{R}^{\mathrm{val}}(u)$:
$$R_Q(\hat{t}) \leq \widehat{R}^{\mathrm{val}}(\hat{t}) + \gamma_{\mathrm{val}} + (c_{10} + c_{01})\Delta_\pi + c_{10}\pi_P D_1^-(\hat{t}) + c_{01}(1 - \pi_P)D_0^-(\hat{t}) + \omega_P(|\hat{t} - t^\star|).$$

All constants and diagnostics were defined above. We proceed line-by-line in three steps.

Step 1 (Algebraic shift decomposition). Recall the form
$$R_D(t) = c_{10}\,\pi_D\,F_{1,D}^-(t) + c_{01}\,(1 - \pi_D)\,(1 - F_{0,D}^-(t)).$$

Fix the realized $\hat{t}$. Write the difference explicitly:
$$R_Q(\hat{t}) - R_P(\hat{t}) = c_{10}\big(\pi_Q F_{1,Q}^-(\hat{t}) - \pi_P F_{1,P}^-(\hat{t})\big) + c_{01}\big((1 - \pi_Q)(1 - F_{0,Q}^-(\hat{t})) - (1 - \pi_P)(1 - F_{0,P}^-(\hat{t}))\big). \quad (1)$$

Insert and subtract the mixed terms $c_{10}\pi_P F_{1,Q}^-(\hat{t})$ and $c_{01}(1 - \pi_P)(1 - F_{0,Q}^-(\hat{t}))$ to separate prevalence and shape components:

$$(??) = c_{10}(\pi_Q - \pi_P)F_{1,Q}^-(\hat{t}) + c_{10}\pi_P\big(F_{1,Q}^-(\hat{t}) - F_{1,P}^-(\hat{t})\big) \quad (2)$$

$$+ c_{01}\big((1 - \pi_Q) - (1 - \pi_P)\big)(1 - F_{0,Q}^-(\hat{t})) + c_{01}(1 - \pi_P)\big[(1 - F_{0,Q}^-(\hat{t})) - (1 - F_{0,P}^-(\hat{t}))\big] \quad (3)$$

$$= (\pi_Q - \pi_P)\big[c_{10}F_{1,Q}^-(\hat{t}) - c_{01}(1 - F_{0,Q}^-(\hat{t}))\big] \quad (4)$$

$$+ c_{10}\pi_P\big(F_{1,Q}^-(\hat{t}) - F_{1,P}^-(\hat{t})\big) + c_{01}(1 - \pi_P)\big(F_{0,P}^-(\hat{t}) - F_{0,Q}^-(\hat{t})\big). \quad (5)$$

Bounding the prevalence term. Since $0 \leq F_{1,Q}^-(\hat{t}) \leq 1$ and $0 \leq 1 - F_{0,Q}^-(\hat{t}) \leq 1$ we have
$$(\pi_Q - \pi_P)\big[c_{10}F_{1,Q}^-(\hat{t}) - c_{01}(1 - F_{0,Q}^-(\hat{t}))\big] \leq |\pi_Q - \pi_P|(c_{10}F_{1,Q}^-(\hat{t}) + c_{01}(1 - F_{0,Q}^-(\hat{t}))) \leq (c_{10} + c_{01})\Delta_\pi,$$
where $\Delta_\pi = |\pi_Q - \pi_P|$. (If $\pi_Q < \pi_P$ the same bound holds because we take absolute value.)

Bounding the shape terms. Introduce the diagnostics
$$D_1^-(t) = |F_{1,Q}^-(t) - F_{1,P}^-(t)|, \qquad D_0^-(t) = |F_{0,Q}^-(t) - F_{0,P}^-(t)|.$$

From $(??)$, using $|a| = -\min(a, -a) \geq a$ we obtain
$$R_Q(\hat{t}) - R_P(\hat{t}) \leq (c_{10} + c_{01})\Delta_\pi + c_{10}\pi_P|F_{1,Q}^-(\hat{t}) - F_{1,P}^-(\hat{t})| + c_{01}(1 - \pi_P)|F_{0,Q}^-(\hat{t}) - F_{0,P}^-(\hat{t})|$$
$$= (c_{10} + c_{01})\Delta_\pi + c_{10}\pi_P D_1^-(\hat{t}) + c_{01}(1 - \pi_P)D_0^-(\hat{t}).$$

Rearranging gives the first key inequality
$$R_Q(\hat{t}) \leq R_P(\hat{t}) + (c_{10} + c_{01})\Delta_\pi + c_{10}\pi_P D_1^-(\hat{t}) + c_{01}(1 - \pi_P)D_0^-(\hat{t}). \quad (\mathrm{S1})$$

Step 2 (Generalization insertion). Define the high-probability event
$$\mathcal{E} = \Big\{ \sup_{t \in \mathbb{R}} |R_P(t) - \widehat{R}^{\mathrm{val}}(t)| \leq \gamma_{\mathrm{val}} \Big\}, \qquad \Pr(\mathcal{E}) \geq 1 - \delta_{\mathrm{val}} \text{ (Lemma ??)}.$$

On $\mathcal{E}$, for the realized $\hat{t}$ we have
$$R_P(\hat{t}) \leq \widehat{R}^{\mathrm{val}}(\hat{t}) + \gamma_{\mathrm{val}}. \quad (6)$$

Substitute $(??)$ into $(??)$ to obtain (on $\mathcal{E}$)
$$R_Q(\hat{t}) \leq \widehat{R}^{\mathrm{val}}(\hat{t}) + \gamma_{\mathrm{val}} + (c_{10} + c_{01})\Delta_\pi + c_{10}\pi_P D_1^-(\hat{t}) + c_{01}(1 - \pi_P)D_0^-(\hat{t}). \quad (\mathrm{Base})$$

Thus the Base bound holds with probability at least $1 - \delta_{\mathrm{val}}$.

Step 3 (Stability augmentation). Assume now $\hat{t}$ is an empirical minimizer of $\widehat{R}^{\mathrm{val}}$. The modulus definition gives

$$R_Q(\hat{t}) \le R_P(t^\star) + \omega_P(|\hat{t} - t^\star|). \tag{7}$$

On $\mathcal{E}$, $R_P(t^\star) \le \widehat{R}^{\mathrm{val}}(t^\star) + \gamma_{\mathrm{val}}$; empirical optimality implies $\widehat{R}^{\mathrm{val}}(\hat{t}) \le \widehat{R}^{\mathrm{val}}(t^\star)$. Combining:

$$R_Q(\hat{t}) \le \widehat{R}^{\mathrm{val}}(\hat{t}) + \gamma_{\mathrm{val}} + \omega_P(|\hat{t} - t^\star|).$$

Insert this into (**??**) (replacing $R_P(\hat{t})$) to append $\omega_P(|\hat{t} - t^\star|)$ and obtain the Augmented bound on $\mathcal{E}$. The probability statement is unchanged. $\qquad\square$

2.1 Proof of bootstrap stability control (Proposition 4.9)

We work under the paper's main notation. In particular, for $y \in \{0,1\}$ let $F_{y,P}^-(t) = \Pr(S < t \mid Y = y)$ and recall the *internal risk modulus* in oscillation form:

$$\omega_P(\epsilon) = \sup_{t \in \mathbb{R}} \left\{ c_{10}\, \pi_P\, \mathrm{osc}_\epsilon(F_{1,P}^-; t) + c_{01}\, (1 - \pi_P)\, \mathrm{osc}_\epsilon(F_{0,P}^-; t) \right\}, \qquad \mathrm{osc}_\epsilon(F; t) := \sup_{u \in [t-\epsilon, t+\epsilon]} F(u) - \inf_{v \in [t-\epsilon, t+\epsilon]} F(v). \tag{8}$$

Let $\widehat{\omega}_P^{\uparrow}$ denote a *conservative DKW-based upper band* for ω_P constructed from the validation sample (Remark 4.4), i.e., on a grid $E \subset [0, \epsilon_{\max}]$ we have with high probability $\omega_P(\epsilon) \le \widehat{\omega}_P^{\uparrow}(\epsilon)$ for all $\epsilon \in E$.

2.1.1 Proof

Let $\hat{t}$ be selection-honest (the validation set used to evaluate $\widehat{R}_{\mathrm{val}}$ is not reused for training/selection).

Asymptotic conventions. Throughout, limits are taken as $n_{\mathrm{val}} \to \infty$ (and, when relevant, $B = B(n_{\mathrm{val}}) \to \infty$).

- **Deterministic $o(1)$.** A deterministic sequence a_n is $o(1)$ if $a_n \to 0$ as $n_{\mathrm{val}} \to \infty$. We use $o(1)$ to denote generic deterministic remainders that vanish in this limit.

- **Stochastic $o_p(1)$.** A sequence of random variables X_n is $o_p(1)$ if $X_n \xrightarrow{p} 0$, i.e., for every $\varepsilon > 0$, $\Pr(|X_n| > \varepsilon) \to 0$.

- **Stochastic $O_p(1)$ (bounded in probability).** A sequence X_n is $O_p(1)$ if for every $\varepsilon > 0$ there exists $M < \infty$ such that $\sup_n \Pr(|X_n| > M) \le \varepsilon$.

- **Bootstrap notation.** When needed, $\Pr^*(\cdot)$ and $\mathbb{E}^*[\cdot]$ denote probability and expectation under the *conditional* (patient-block) bootstrap, given the observed data.

- **Remainder symbols in Proposition ??.** The terms $\xi_n(B, \delta_{\mathrm{boot}})$ and η_n are nonnegative sequences with $\xi_n(B, \delta_{\mathrm{boot}}) \to 0$ (as $n_{\mathrm{val}} \to \infty$ and $B \to \infty$) and $\eta_n \to 0$ (from the DKW band event). Unless stated otherwise, all $o(1)$ terms can be taken deterministic after intersecting the relevant high-probability events.

Assume:

(B1) **Local well-posedness.** $R_P(t)$ has a unique minimizer $t^\star$ in a neighborhood $\mathcal{N}$ and is directionally differentiable there. The modulus ω_P in (**??**) is nondecreasing and locally Lipschitz on $[0, \epsilon_{\max}]$.

(B2) **Patient-block bootstrap consistency.** Conditionally on the data, the law of $\hat{t}^* - \hat{t}$ under the patient-block bootstrap is weakly consistent for the sampling law of $\hat{t} - t^\star$ centered at its mean. Moreover, the bootstrap bias estimate

$$\hat{b}_{\mathrm{boot}} := \frac{1}{B} \sum_{b=1}^{B} (\hat{t}^{*(b)} - \hat{t}) \quad \text{satisfies} \quad \hat{b}_{\mathrm{boot}} \xrightarrow{p} b := \mathbb{E}(\hat{t} - t^\star).$$

(B3) **Conservative DKW modulus band.** There exists a grid $E \subset [0, \epsilon_{\max}]$ and a sequence $\eta_n \downarrow 0$ such that, with probability at least $1 - \eta_n$,

$$\sup_{\epsilon \in E} \left\{ \omega_P(\epsilon) - \widehat{\omega}_P^{\uparrow}(\epsilon) \right\} \leq 0.$$

(Equivalently, $\widehat{\omega}_P^{\uparrow}$ is a uniform high-probability upper band for ω_P on E.)

Let $q^*_{1-\delta_{\text{boot}}}$ be the empirical $(1 - \delta_{\text{boot}})$ upper quantile of the *centered* bootstrap absolute deviations $\left| (\hat{t}^{*(b)} - \hat{t}) - \hat{b}_{\text{boot}} \right|$, and define the data-driven instability envelope

$$G_{\text{boot}} := \widehat{\omega}_P^{\uparrow} \left(|\hat{b}_{\text{boot}}| + q^*_{1-\delta_{\text{boot}}} \right).$$

Then there exist remainder terms $\xi_n(B, \delta_{\text{boot}}), \eta_n \to 0$ such that

$$\Pr\left\{ \omega_P(|\hat{t} - t^\star|) \leq G_{\text{boot}} + o(1) \right\} \geq 1 - \delta_{\text{boot}} - \xi_n(B, \delta_{\text{boot}}) - \eta_n,$$

where $o(1) \to 0$ as $n_{\text{val}} \to \infty$ (allowing $B = B(n_{\text{val}}) \to \infty$).

Let $b = \mathbb{E}(\hat{t} - t^\star)$ and decompose

$$|\hat{t} - t^\star| \leq \left| (\hat{t} - t^\star) - b \right| + |b|.$$

Step 1 (Bootstrap quantile coverage). By (B2), the conditional bootstrap law of $(\hat{t}^* - \hat{t}) - \hat{b}_{\text{boot}}$ approximates the sampling law of $(\hat{t} - t^\star) - b$. Standard quantile consistency for weakly convergent empirical c.d.f.s with $B \to \infty$ gives

$$q^*_{1-\delta_{\text{boot}}} = q_{1-\delta_{\text{boot}}} + o_p(1),$$

where $q_{1-\delta_{\text{boot}}}$ is the $(1 - \delta_{\text{boot}})$ quantile of $\left| (\hat{t} - t^\star) - b \right|$. Hence

$$\Pr\left(\left| (\hat{t} - t^\star) - b \right| \leq q^*_{1-\delta_{\text{boot}}} + o(1) \right) \geq 1 - \delta_{\text{boot}} - \xi_n(B, \delta_{\text{boot}}).$$

Step 2 (Bias estimation). Still under (B2), $\hat{b}_{\text{boot}} \xrightarrow{p} b$, hence $|b| = |\hat{b}_{\text{boot}}| + o_p(1)$. Combining with Step 1,

$$|\hat{t} - t^\star| \leq |\hat{b}_{\text{boot}}| + q^*_{1-\delta_{\text{boot}}} + o_p(1) \quad \text{with probability} \geq 1 - \delta_{\text{boot}} - \xi_n.$$

Step 3 (Oscillation modulus and conservative band). By monotonicity of ω_P (Definition (**??**)) and local Lipschitz continuity (B1),

$$\omega_P(|\hat{t} - t^\star|) \leq \omega_P\left(|\hat{b}_{\text{boot}}| + q^*_{1-\delta_{\text{boot}}} + o(1) \right) = \omega_P\left(|\hat{b}_{\text{boot}}| + q^*_{1-\delta_{\text{boot}}} \right) + o(1).$$

Intersect the high-probability event from (B3) (the DKW upper band) with the event from Steps 1–2. On that intersection, for all ϵ on the band grid E,

$$\omega_P(\epsilon) \leq \widehat{\omega}_P^{\uparrow}(\epsilon),$$

hence, after discretizing ϵ on E (or using continuity to pass to the limit),

$$\omega_P\left(|\hat{b}_{\text{boot}}| + q^*_{1-\delta_{\text{boot}}} \right) \leq \widehat{\omega}_P^{\uparrow}\left(|\hat{b}_{\text{boot}}| + q^*_{1-\delta_{\text{boot}}} \right) = G_{\text{boot}}.$$

Collecting terms yields

$$\omega_P(|\hat{t} - t^\star|) \leq G_{\text{boot}} + o(1)$$

with probability at least $1 - \delta_{\text{boot}} - \xi_n(B, \delta_{\text{boot}}) - \eta_n$. All $o(1)$ remainders can be taken deterministic after intersecting the high-probability events. □

Notes. The oscillation form (**??**) ties the instability penalty to the local shape of the class-conditional CDFs around the realized operating point.

The band $\widehat{\omega}_P^{\uparrow}$ is obtained by plugging DKW uniform bands for $F_{y,P}^-$ into (**??**) and propagating through the oscillation operator (Remark 4.4); its conservativeness replaces the plug-in consistency used in the earlier formulation.

2.2 Additional remarks and details regarding the nature of hierarchical

Remark Hierarchical variance layers and choice of γ_{val} scale In many biomedical settings the observed data are hierarchical. Up to four (non-exhaustive) stochastic layers can be conceptually separated for a fixed threshold t:

1. **Technical / assay noise** (e.g. instrument, batch) affecting raw measurements before any scoring model; absorbed after preprocessing into the marginal score distribution.

2. **Within-patient biological heterogeneity** (e.g. cell- or region-level variation) producing multiple raw units per patient; a deterministic aggregation (or model) maps these to a single patient-level score S_p used for thresholding.

3. **Between-patient sampling variation**: i.i.d. draws (S_p, Y_p) under internal domain P; this is the level governed by the VC=1 empirical process and generates γ_{val}.

4. **Cross-domain shift** (from P to Q): prevalence and conditional shape discrepancies; treated deterministically via $(c_{10} + c_{01})\Delta_\pi + c_{10}\pi_P D_1^- + c_{01}(1 - \pi_P)D_0^-$, not part of γ_{val}.

Additionally, **selection / optimization instability** of the empirical minimizer (threshold fluctuation) is controlled separately through the stability penalty via $\omega_P(|\hat{t} - t^\star|)$.

Law of total variance (schematic). Writing $L_p(t) = L(Y_p, \mathbf{1}\{S_p \geq t\})$, a nested variance decomposition (suppressing t) gives

$$\text{Var}(\widehat{R}^{\text{val}}) = \frac{1}{n_{\text{val}}}\Big(\underbrace{\mathbb{E}[\text{Var}(L_p \mid \text{within-patient data})]}_{\text{within-patient layer}} + \underbrace{\text{Var}(\mathbb{E}[L_p \mid \text{within-patient data}])}_{\text{between-patient layer}}\Big),$$

before introducing domain shift (which changes the target mean rather than adding stochastic variance). When each patient has m_p raw units with intra-patient correlation ρ, treating raw units as independent would underestimate variance by the design effect $\text{deff}_p \approx 1 + (\bar{m} - 1)\rho$. Patient-level aggregation circumvents this: the empirical process sees n_{val} independent sets, preserving the $\sqrt{\log(1/\delta)/n_{\text{val}}}$ rate; any multi-layer dependence is absorbed into constants.

Refined effective sample size. If one insisted on operating at the raw-unit level (size $N = \sum_p m_p$) the effective independent count satisfies by classical ICC analysis:

$$n_{\text{eff}} \approx \frac{N}{1 + (\bar{m} - 1)\rho} \quad \Rightarrow \quad \gamma_{\text{val}} \text{ inflated by } \sqrt{\frac{1 + (\bar{m} - 1)\rho}{\bar{m}}}.$$

Further sublayers (e.g. technical replicates per raw unit) multiply design effects multiplicatively or additively in first-order approximations, again only altering constants in the VC tail bound.

Why patient-level γ_{val}? External risk transfer and clinical decision units are at patient resolution; using the cluster (patient) granularity avoids pseudo-replication, keeps the VC dimension minimal, and yields a transparent decomposition: *(internal generalization) + (shift) + (stability)*. A fully expanded hierarchical γ_{val} would obscure interpretation without changing asymptotic order.

Optional reporting. One may still report a diagnostic design-effect estimate to justify constants in γ_{val}, or present a decomposed version

$$\gamma_{\text{val}} \approx (c_{10} + c_{01})B_\pi + c_{10}\pi_P B_1 + c_{01}(1 - \pi_P)B_0,$$

with B_π, B_1, B_0 the chosen prevalence and class-CDF uniform bounds (possibly adjusted by estimated design effects). We retain the compact form for readability.

3 ADDITIONAL EXPERIMENTS

3.1 Model-agnosticity

Our framework is model-agnostic by definition: it operates on the patient-level scores $S = A(\{s(X_{pi})\}_{i \in I_p})$ and only uses the empirical label prevalence and the class-conditional CDFs of $S \mid Y$ (via DKW bands) together with the oscillation-based risk modulus.

No structural assumptions on the scorer s or the aggregator A are required, and the same validity certificate holds for any learning algorithm.

The figure below is just an *illustrative example*; the identical calibration and bounding procedure applies verbatim to all models.
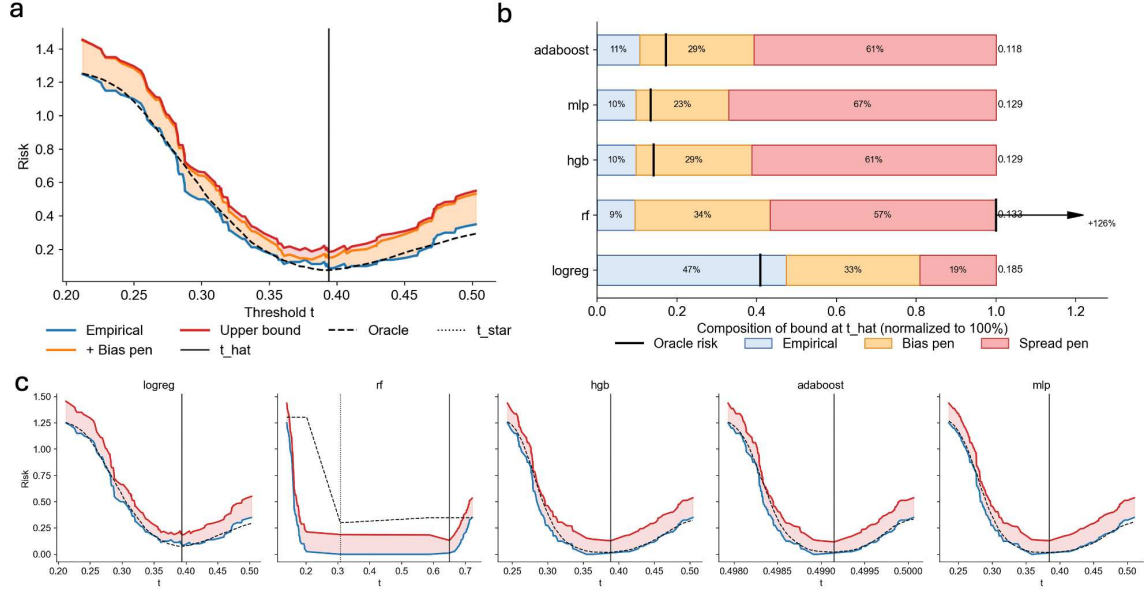

Figure 1: **Patient-domain (P-frozen) upper bound and contribution decomposition (Illustrative case).** (a) For a representative classifier, we plot the pointwise upper bound

$$U_P(t) \;=\; \widehat{R}_{\mathrm{val}}(t) \;+\; \gamma_{\mathrm{val}}(\delta_{\mathrm{val}}) \;+\; G_{\mathrm{boot}},$$

(red) together with the empirical risk $\widehat{R}_{\mathrm{val}}(t)$ (blue) and the oracle curve $R_P(t)$ (dashed black). In the P-frozen regime (no operating-point shift term), $U_P(t) \geq \widehat{R}_{\mathrm{val}}(t)$ for all t and remains above $R_P(t)$, confirming validity. The selected threshold is $\hat{t} = \arg\min_t U_P(t)$ (vertical line). (b) At $\hat{t}$, the bound decomposes into contributions that sum to 100%: empirical term $\widehat{R}_{\mathrm{val}}(\hat{t})$ (blue), the uniform patient-level generalization term $\gamma_{\mathrm{val}}(\delta_{\mathrm{val}})$ (orange; from DKW/Hoeffding), and the instability penalty G_{boot} (red; bootstrap radius passed through the oscillation modulus). The black tick marks $R_P(\hat{t})$; the number at the bar's right is $U_P(\hat{t})$. (c) Panel (a) repeated per model (logistic regression, random forest, histogram gradient boosting, AdaBoost, shallow MLP), illustrating model-agnostic application under P-frozen regime.

3.2 Notes on figures

Notes on data-generating processes and algorithmic parameters of figure 1 (illustrative case): Internal (P) scores are drawn from a two-basin mixture with a "sharp" subgroup (fraction 0.12) having class-conditional Gaussians $(\mu_0, \mu_1) = (1.9, 2.1)$ and SDs $(0.15, 0.13)$, and a "flat" subgroup $(\mu_0, \mu_1) = (2.5, 4.5)$ with SD 1.0; P includes heteroskedastic noise (amplitude 2.0, center 0.90, width 0.28). External (Q) scores are generated *independently* from their own mixture: "sharp" $(\mu_0, \mu_1) = (1.95, 2.05)$ with SDs $(0.16, 0.14)$ and "flat" $(\mu_0, \mu_1) = (2.55, 4.35)$ with SD 1.05, plus Q-specific heteroskedastic noise (amplitude 1.8, center 2.00, width 0.32). Hierarchical sampling for P uses $n_{\mathrm{patients}} = 180$ and 800 cells per patient (so with $n_{\mathrm{internal}} = 800$); Q uses $n_{\mathrm{external}} = 6000$. The instability penalty $\mathcal{G}_{\mathrm{boot}}(t)$ uses patient-level bootstrap ($B{=}200$) with moving-average smoothing (window = 5). λ is calibrated so the 0.58-quantile of the raw $\mathcal{G}_{\mathrm{boot}}(t)$ matches $1.15\times$ the empirical risk range. Selection is directional (to the right of t^{ERM}) with movement cap 0.90 and cost ratio = 1.0.

3.3 Additional information on CAMELYON16/17 processing

Data pre-processing (CAMELYON16/17; slide & patient levels). CAMELYON16 and CAMELYON17 come from different origins, and *both* were processed at two granularities: (i) slide level and (ii) patient level. For each dataset *separately*, we created two disjoint partitions P and Q by a random 50/50 split. Splits were stratified by the task label; at the patient level, all slides from the same patient were kept within the same partition to avoid leakage. Whole-slide images were read at the same fixed magnification, tissue regions were segmented to remove background, and non-overlapping tiles (256–512 px) were extracted within tissue as preprocessing (see GoogLeNet procedure). Low-quality tiles (low tissue fraction / blur / extreme brightness) were filtered. Each retained tile was scored with a PyTorch GoogLeNet (Inception v1) classifier to obtain $s(x) \in [0, 1]$. Slide scores were formed by aggregating tile scores (max); patient scores were then obtained by aggregating a patient's slide scores (quantile). All pre-/post-processing choices (magnification, tiling, QC, aggregators) were held identical in P and Q.

Post-threshold reporting (P-frozen). Thresholds were selected on P in a selection-honest manner and then *frozen* and evaluated on Q. After thresholding, we computed summary performance at both slide and patient levels (e.g., accuracy, risk under (c_{10}, c_{01}), and related metrics) within each dataset. Owing to space constraints, the main table reports *only mean values* (CAMELYON16 and CAMELYON17 reported separately).

3.4 Additional information on MIMIC-IV Demo processing

Data pre-processing (MIMIC-IV-ECG Demo; patient-level split). We used only the publicly available MIMIC-IV-ECG *Demo* subset (12-lead, 10 s, 500 Hz). To prevent leakage, we randomly split *patients* 50/50 into two disjoint partitions P and Q (all ECGs for a patient remain in the same partition). Each ECG was parsed lead-wise, detrended, and band-pass filtered; amplitudes were z-scored per lead (per record). We followed the procedure of `https://github.com/nliulab/mimic4ed-benchmark`, with taking the best performing logistic regression. The segment-level classifier produced scores $p_{i,j} \in [0, 1]$ for segment j of ECG i.

Aggregation and evaluation (P-frozen). Per-ECG scores were obtained by aggregating a record's segment scores (e.g., quantile at $q{=}0.9$): $s_i = g(\{p_{i,j}\}_j)$. Patient-level scores were formed by aggregating across that patient's ECGs (quantile): $S_p = h(\{s_i\}_{i \in \text{patient } p})$. Thresholds were tuned on P in a selection-honest manner and then *frozen* and evaluated on Q (P-frozen). In the main table, we report mean accuracy and mean risk (under (c_{10}, c_{01})) computed at the selected threshold.