Local Audit Certification of Transferred Medical Imaging Models in Resource-Constrained Settings

Sita Bissu[1], Navnit Kumar Yadav[1], Aparna Mehra[1], and Sandeep Kumar[2]

[1] Department of Mathematics, Indian Institute of Technology Delhi, New Delhi, India
maz238455@maths.iitd.ac.in, navnitydv@gmail.com, apmehra@maths.iitd.ac.in

[2] Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, India
ksandeep@iitd.ac.in

Abstract. A clinic in a resource-constrained setting often cannot retrain a medical-imaging model; it adopts a frozen pretrained or foundation model and must decide whether that model is safe on its *local* patient population. We frame this post-transfer decision as local-audit certification: from a small labelled audit and a bounded clinical loss, the site obtains a one-sided, finite-sample, distribution-free upper certificate on the deployment risk of the fixed predictor. The certificate can be compared with a local acceptable-risk threshold to decide whether to deploy, collect more labels, or reject the transfer. We instantiate the protocol with an empirical-Bernstein certificate, give a variance-adaptive audit-size statement, add a finite-sample Kullback–Leibler distributionally robust optimization (KL-DRO) margin over one-dimensional loss laws for audit–deployment mismatch, and give a two-point lower bound showing that the leading variance-dependent slack is unavoidable up to constants. On real diabetic-retinopathy transfer (EyePACS↔APTOS), the certificates are valid, become sharper with audit size, and quantify the robustness budget needed to cover a genuine cross-cohort shift.

Keywords: Transferred medical-imaging models · Risk control · Distribution shift · Diabetic retinopathy · Resource-constrained settings.

1 Introduction

The dominant way medical imaging AI reaches a resource-constrained setting (RCS) is by *transfer*: a site adopts a model trained elsewhere—a foundation model with public weights [21], or a network trained on a large high-income-country cohort—and applies it locally. The model is effectively frozen, since there is neither the data nor the compute to retrain it, and often no on-site engineer. The clinically decisive question is therefore not "can we build a better model?" but "*is this particular frozen model safe to use on our patients?*", and it must be answered under the defining RCS constraints: the local population differs from the source population (scanners, demographics, disease spectrum, image quality), and local labels are scarce and expensive.

This perspective treats transfer as a certification problem rather than a modelling problem. The goal is not to adapt the external predictor, but to upper-bound the realised risk of a fixed model on the local deployment distribution. The guarantee is finite-sample, distribution-free, and explicitly quantifies the local labelling needed for deployment. Given an acceptable risk threshold α, the certificate becomes a deployment rule: use the frozen model only if the certified upper risk is at most α; otherwise collect more labels or reject the transfer.

Our contributions are:

1. We formulate frozen-model transfer in resource-constrained medical imaging as a *local audit certification* problem: given a fixed external model and a small labelled local audit, certify an upper bound on deployment risk (Sec. 2).
2. We instantiate this protocol with a one-sided empirical-Bernstein certificate that is finite-sample, distribution-free, and valid for any bounded clinical loss (Sec. 3.1).
3. We show that the certificate is variance-adaptive: when the transferred model has low local error, the required audit size can be substantially smaller than variance-blind worst-case bounds (Sec. 3.2).
4. We extend the protocol to audit–deployment mismatch using a finite-sample KL-DRO margin over the one-dimensional loss law, avoiding density-ratio estimation in high-dimensional image space (Sec. 3.3).
5. We validate the protocol on real cross-population diabetic-retinopathy transfer between EyePACS and APTOS, showing valid coverage, shrinking certificates with audit size, and the robustness budget needed to cover a real cross-cohort shift (Sec. 4).

The empirical-Bernstein estimator itself is classical [16]; our contribution is the deployment-certification formulation, the clinic-facing audit protocol, the loss-law robustness margin, the lower-bound perspective, and the real-data calibration of the robustness budget.

Relation to prior work. Transferred and foundation medical-imaging models are increasingly reused across clinical sites, but source-domain performance alone does not provide a finite-sample certificate of target-site safety [13, 3, 21]. Conformal prediction and conformal risk control provide finite-sample, distribution-free guarantees under exchangeability [4, 1], while training-conditional perspectives clarify guarantees when the fitted model is held fixed [5]. Prediction-powered inference uses small labelled sets with unlabelled data but typically assumes a shared distribution [2]. Robust validation and covariate-shift methods address shift, often through density-ratio or feature-space assumptions [8, 6]. Our focus is different: we certify the local deployment risk of a frozen transferred imaging model from a small labelled audit, and robustify over the one-dimensional loss law rather than the high-dimensional image distribution.

2 Problem setup

Fix a measurable predictor $f : \mathcal{X} \to \mathcal{A}$ (the frozen transferred model, with prediction space $\mathcal{A}$) and a loss $\ell : \mathcal{A} \times \mathcal{Y} \to [0, 1]$, e.g. 0/1 misdiagnosis loss or

a clipped clinical cost. Let Q be the local deployment distribution on $\mathcal{X} \times \mathcal{Y}$. The object to be certified is the deployment risk $R_Q(f) = \mathbb{E}_{(X,Y)\sim Q}[\ell(f(X), Y)]$. Given confidence $(1-\delta)$ and the data available at the site, n local labelled cases $(X_i, Y_i)_{i=1}^n \overset{\text{iid}}{\sim} Q$, we seek a statistic $\hat{U}$ with

$$\mathbb{P}\big[R_Q(f) \le \hat{U}\big] \ge 1 - \delta, \tag{1}$$

where the probability is over the local sampling with f held fixed (a *training-conditional* guarantee [5]), and the bound is *distribution-free*: it holds for every Q subject only to $\ell \in [0, 1]$. Two questions follow. First, how few local labels suffice for a useful certificate? Second, what happens when the audited population is not exactly the future deployment population? We answer the latter with a distributionally robust margin over *loss laws*. This margin avoids covariate-shift modelling and density-ratio estimation in high-dimensional image space, but its guarantee is conditional on the future deployment loss law lying inside the chosen KL radius.

Audit-sampling assumption. The plain certificate requires the labelled audit cases to be sampled from the same patient stream that the clinic intends to deploy on, with the predictor f fixed before the audit labels are examined. Thus, representativeness is a sampling requirement rather than an assumption that the source and target populations are identical. Systematic audit selection—for example, preferentially labelling easier cases, higher-quality images, or patients already referred to specialist care—can invalidate the i.i.d. certificate. The KL-DRO extension in Sec. 3.3 relaxes exact audit–deployment equality at the level of the induced loss distribution, but only when the future deployment loss law lies within the specified KL radius. In practice, the audit protocol should therefore document its eligibility criteria, sampling window, exclusions, and intended deployment population.

3 Method and guarantees

Clinic-facing protocol. The workflow is as follows: fix the external predictor f before using audit labels; label n local cases; compute bounded losses $Z_i = \ell(f(X_i), Y_i)$; evaluate the empirical-Bernstein certificate $\widehat{U}_n^{\text{EB}}$; optionally replace it by the KL-DRO certificate $\widehat{U}_{n,\rho,\Lambda}^{\text{KL}}$ when audit–deployment mismatch is suspected; and deploy only if the selected certificate is no larger than the clinic's acceptable risk threshold α. Otherwise, the audit recommends collecting more labels, recalibrating, retraining, or rejecting the transfer.

Worked example. To make the decision rule concrete, suppose a clinic audits $n = 200$ local cases and obtains empirical loss $\hat{R}_n = 0.10$ and sample variance $\hat{V}_n = 0.04$. At 90% confidence ($\delta = 0.10$), Eq. (2) gives

$$\widehat{U}_n^{\text{EB}} = 0.10 + \sqrt{\frac{2(0.04)\ln(20)}{200}} + \frac{7\ln(20)}{3(199)} \approx 0.170.$$

If the clinic has pre-specified an acceptable-risk threshold $\alpha = 0.20$, this audit supports deployment under the i.i.d. audit–deployment assumption because $0.170 < 0.20$. If instead $\alpha = 0.15$, the model is not certified for deployment and the clinic should collect additional labels or reject the transfer. When a mismatch between the audited and future deployment populations is plausible, the KL-DRO certificate in Sec. 3.3 is used instead.

3.1 The deployment certificate

Write $Z_i = \ell(f(X_i), Y_i) \in [0,1]$, $\hat{R}_n = n^{-1}\sum_{i=1}^n Z_i$, and let $\hat{V}_n = (n-1)^{-1}\sum_{i=1}^n (Z_i - \hat{R}_n)^2$ be the unbiased sample variance. Throughout, $n \geq 2$.

Theorem 1. *Let $(X_i, Y_i)_{i=1}^n \overset{\text{iid}}{\sim} Q$ with f fixed. Then for every Q with $\ell \in [0,1]$, with probability at least $1-\delta$ over the audit draw,*

$$R_Q(f) \;\leq\; \hat{U}_n^{\text{EB}} := \hat{R}_n + \sqrt{\frac{2\hat{V}_n \ln(2/\delta)}{n}} + \frac{7\ln(2/\delta)}{3(n-1)}. \tag{2}$$

Proof. The variables $Z_1, \ldots, Z_n$ are i.i.d. in $[0,1]$ with mean $\mathbb{E}[Z_i] = R_Q(f)$. The one-sided empirical-Bernstein inequality of Maurer and Pontil [16] states that for i.i.d. $[0,1]$-valued variables, with probability at least $1-\delta$, $\mathbb{E}[Z] \leq \hat{R}_n + \sqrt{2\hat{V}_n \ln(2/\delta)/n} + 7\ln(2/\delta)/(3(n-1))$. Substituting $\mathbb{E}[Z] = R_Q(f)$ gives the claim. Only boundedness of ℓ is used, so the bound holds for every Q; f enters solely through the fixed map $Z_i = \ell(f(X_i), Y_i)$, giving the training-conditional reading. □

Empirical-Bernstein, rather than Hoeffding [15], is useful when the model is accurate locally: then $\hat{V}_n$ is small and the leading slack is governed by a variance term rather than the worst-case $n^{-1/2}$ range term. Proposition 1 makes this precise as an expected certificate-size statement.

3.2 How many local labels?

Proposition 1 (Expected certificate size). *Let $R := R_Q(f)$ and $L = \ln(2/\delta)$. For $Z \in [0,1]$,*

$$\mathbb{E}[\hat{U}_n^{\text{EB}}] \leq R + \sqrt{\frac{2R(1-R)L}{n}} + \frac{7L}{3(n-1)}. \tag{3}$$

Consequently, for a target $\alpha > R$, a sufficient condition for $\mathbb{E}[\hat{U}_n^{\text{EB}}] \leq \alpha$ is, up to the lower-order $O(1/n)$ term,

$$n \gtrsim \frac{2R(1-R)\ln(2/\delta)}{(\alpha - R)^2}. \tag{4}$$

By contrast, the Hoeffding certificate $\hat{R}_n + \sqrt{\ln(1/\delta)/(2n)}$ requires

$$n \gtrsim \ln(1/\delta)/(2(\alpha - R)^2)$$

at the same heuristic level.

Proof. Taking expectations in Theorem 1, using Jensen's inequality for the square-root term, and using $\mathbb{E}\hat{V}_n = \mathrm{Var}_Q(Z)$ yields

$$\mathbb{E}[\hat{U}_n^{\mathrm{EB}}] \le R + \sqrt{\frac{2\mathrm{Var}_Q(Z)L}{n}} + \frac{7L}{3(n-1)}.$$

Since $0 \le Z \le 1$ and $\mathbb{E}Z = R$, we have $\mathrm{Var}_Q(Z) \le R(1-R)$. Substitution gives Eq. (3), and solving the leading term for n gives the stated sufficient scaling. □

Thus, in the high-accuracy regime $R \ll 1$, empirical-Bernstein can require substantially fewer labels than a variance-blind certificate. This does not mean that empirical-Bernstein uniformly dominates Hoeffding; the experiments below show a crossover between the two, motivating a union-bound combination.

3.3 Robustness to audit–deployment mismatch

The certificate above assumes the audit is drawn from the deployment distribution Q. In practice the audited stream and the true deployment stream can differ: early audits may over-represent patients who are easier to label, higher-quality images, or cases already routed through specialist care. We therefore robustify over the *one-dimensional loss law*. This is deliberate: the result does not require a covariate-shift model and estimates no density ratio in image-feature space.

Let P_{a} be the audit loss law of $Z = \ell(f(X), Y)$. Suppose the true deployment loss law P_{d} satisfies $\mathrm{KL}(P_{\mathrm{d}} \| P_{\mathrm{a}}) \le \rho$. By the Donsker–Varadhan variational formula and KL-DRO duality [10, 12], the population KL-DRO identity is

$$\sup_{\mathrm{KL}(P_{\mathrm{d}}\|P_{\mathrm{a}})\le\rho} \mathbb{E}_{P_{\mathrm{d}}}[Z] = \inf_{\lambda>0} \left\{ \lambda\rho + \lambda \log \mathbb{E}_{P_{\mathrm{a}}}\left[e^{Z/\lambda}\right] \right\}. \tag{5}$$

For finite samples, the minimisation over λ must be handled uniformly. We use a finite grid $\Lambda = \{\lambda_1, \ldots, \lambda_m\} \subset (0, \infty)$. For each $\lambda \in \Lambda$, define the rescaled bounded variable

$$T_\lambda(Z) = \frac{e^{Z/\lambda} - 1}{e^{1/\lambda} - 1} \in [0, 1]. \tag{6}$$

Let $\hat{\mu}_\lambda$ and $\hat{V}_\lambda$ be the empirical mean and unbiased sample variance of

$$T_\lambda(Z_1), \ldots, T_\lambda(Z_n)$$

, and set

$$B_\lambda(\delta, m) = \hat{\mu}_\lambda + \sqrt{\frac{2\hat{V}_\lambda \ln(2m/\delta)}{n}} + \frac{7\ln(2m/\delta)}{3(n-1)}. \tag{7}$$

Finally define

$$\hat{U}^{\mathrm{KL}}_{n,\rho,\Lambda} = \min_{\lambda\in\Lambda} \left\{ \lambda\rho + \lambda \log\left(1 + (e^{1/\lambda} - 1)\min\{1, B_\lambda(\delta, m)\}\right) \right\}. \tag{8}$$

Proposition 2 (Finite-sample KL-DRO certificate). *Let* $Z_1, \dots, Z_n \overset{\text{iid}}{\sim} P_{\text{a}}$ *be audited losses in* $[0,1]$ *and let* Λ *be a fixed finite grid. With probability at least* $1-\delta$ *over the audit draw, simultaneously for every deployment loss law* P_{d} *with* $\mathrm{KL}(P_{\text{d}} \| P_{\text{a}}) \le \rho$,

$$\mathbb{E}_{P_{\text{d}}}[Z] \le \hat{U}^{\text{KL}}_{n,\rho,\Lambda}. \tag{9}$$

Proof. For a fixed λ, Theorem 1 applied to $T_\lambda(Z) \in [0,1]$ with confidence level $1-\delta/m$ gives $\mathbb{E}_{P_{\text{a}}}[T_\lambda(Z)] \le B_\lambda(\delta, m)$ with probability at least $1-\delta/m$. A union bound gives this event for all $\lambda \in \Lambda$ with probability at least $1-\delta$. On this event,

$$\mathbb{E}_{P_{\text{a}}}[e^{Z/\lambda}] = 1 + (e^{1/\lambda} - 1)\mathbb{E}_{P_{\text{a}}}[T_\lambda(Z)] \le 1 + (e^{1/\lambda} - 1)\min\{1, B_\lambda(\delta, m)\}$$

for every grid value. Combining this upper bound with the variational identity in Eq. (5) and minimising over the grid gives Eq. (8). □

The grid Λ is fixed before evaluating the audit certificate; in practice we choose it log-spaced over the range of λ used in the numerical dual. The radius ρ is a user-chosen robustness budget. The guarantee protects every deployment loss law inside the KL ball, but it is not intended to protect deployments whose loss law lies outside the chosen budget. In the experiments, cross-cohort KL estimates are used as retrospective diagnostics of the budget needed for real shifts, not as distribution-free estimates of the unknown future KL radius.

Choosing the robustness radius in practice. Proposition 2 treats ρ as a pre-specified robustness budget; it does not provide a distribution-free estimator of the future audit–deployment divergence. A clinic may choose this budget using historical audit-to-deployment shifts from the same workflow, evidence from comparable sites, or a clinically specified tolerance for population change. When a single defensible radius is unavailable, we recommend reporting $\widehat{U}^{\text{KL}}_{n,\rho,\Lambda}$ over a pre-specified range of ρ values, thereby making explicit how the deployment decision changes with the assumed shift budget. Importantly, selecting ρ adaptively from the same audit sample is not automatically covered by Proposition 2; retaining the stated finite-sample guarantee would require an independent calibration source or a separate valid upper-confidence procedure for the divergence. The real cross-cohort KL values reported below are therefore used only to interpret plausible robustness budgets retrospectively.

3.4 A corrected two-point lower bound

Theorem 2 (Unavoidable one-sided slack). *Let* $\hat{U}_n$ *be any one-sided certifier that is valid at level* $1-\delta$ *for all Bernoulli loss distributions. Fix* $p_0 \in (0, 1/2)$, *let* $\sigma^2 = p_0(1-p_0)$, *and set* $p_1 = p_0 + \varepsilon < 1$. *If*

$$n\,\mathrm{KL}(\text{Bernoulli}(p_0) \| \text{Bernoulli}(p_1)) \le \log \frac{1}{4\delta}, \tag{10}$$

then under $Z_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p_0)$,

$$\mathbb{P}_{p_0}\left(\hat{U}_n - p_0 \ge \varepsilon\right) \ge \delta. \tag{11}$$

In particular, using the standard Bernoulli KL bound

$$\mathrm{KL}(\mathrm{Bernoulli}(p_0)\|\mathrm{Bernoulli}(p_0+\varepsilon)) \lesssim \varepsilon^2/\sigma^2$$

for ε in a local neighbourhood of p_0, a slack of order

$$\sigma\sqrt{\frac{\log(1/\delta)}{n}} \tag{12}$$

is unavoidable up to constants whenever this local two-point construction is inside $[0,1]$.

Proof. Let $A = \{\hat{U}_n \geq p_1\}$. Validity under Bernoulli(p_1) gives $\mathbb{P}_{p_1}(A^c) \leq \delta$. By Bretagnolle–Huber [7],

$$\mathbb{P}_{p_0}(A) + \mathbb{P}_{p_1}(A^c) \geq \frac{1}{2}\exp\{-n\mathrm{KL}(\mathrm{Bernoulli}(p_0)\|\mathrm{Bernoulli}(p_1))\}$$

. Under Eq. (10) this lower bound is at least 2δ, so $\mathbb{P}_{p_0}(A) \geq \delta$. On A, the slack under the true mean p_0 is at least $p_1 - p_0 = \varepsilon$; the displayed rate follows from the local Bernoulli KL expansion. □

This modest lower bound rules out distribution-free one-sided certificates with a uniformly smaller leading variance scale; near the boundary, the additive $O(\log(1/\delta)/n)$ term in Theorem 1 is also needed.

4 Experiments

Setup and implementation. The frozen model f is an ImageNet-pretrained ViT-B/16 [11, 9] implemented with PyTorch and `timm` [17, 19], used as a fixed feature extractor with a logistic-regression head fit on *source* labels only using scikit-learn [18]. The backbone and head are frozen before any target audit labels are used. The task is binary referable diabetic retinopathy (International Clinical Diabetic Retinopathy (ICDR) grade ≥ 2) [20], and the certified loss is the per-example Brier score [14], $Z_i = (\hat{p}_i - Y_i)^2 \in [0,1]$. We evaluate both transfer directions: EyePACS→APTOS, using 8000 sampled EyePACS images as source and 3662 APTOS images as target, and APTOS→EyePACS, using APTOS as source and the same EyePACS sample as target [13, 3]. For coverage, the full labelled target cohort is treated as the finite deployment population, whose mean Brier loss is the ground-truth $R_Q(f)$. For each $n \in \{25, 50, 100, 200, 400, 800, 1600\}$, we draw 4000 audits with replacement from the target loss vector and report empirical coverage and mean slack at $\delta = 0.10$. KL-DRO diagnostics use exponential tilts of the audit loss law; real cross-cohort KL values are estimated from smoothed empirical loss histograms and used only as retrospective budget diagnostics.

For reproducibility, the ViT backbone is `vit_base_patch16_224` from `timm`; images are processed with the standard ImageNet evaluation transform. Source features are split stratified into a fitting set and held-out source-evaluation set, and the logistic-regression head is trained only on the source fitting split before being frozen.

Table 1. Certificate validity and slack on cross-population fundus data (δ=0.10, 4000 audits/cell). Coverage is $\mathbb{P}(\widehat{U} \geq R_Q(f))$; slack is mean $\widehat{U} - R_Q(f)$.

	EyePACS→APTOS, R_Q=0.105				APTOS→EyePACS, R_Q=0.174			
n	EB cov.	EB sl.	Hoef. cov.	Hoef. sl.	EB cov.	EB sl.	Hoef. cov.	Hoef. sl.
25	1.000	0.405	1.000	0.215	1.000	0.458	1.000	0.214
100	1.000	0.130	1.000	0.107	1.000	0.156	1.000	0.107
200	1.000	0.077	1.000	0.076	1.000	0.096	1.000	0.076
400	1.000	0.047	1.000	0.054	0.999	0.060	0.999	0.054
1600	0.998	0.019	1.000	0.027	0.998	0.026	0.999	0.027

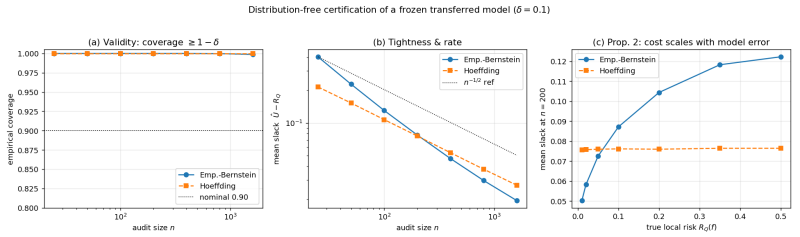

Fig. 1. Certificate behaviour. (a) Coverage(cov.) is at least $1 - \delta$ on real APTOS losses; (b) slack(sl.) contracts with audit size, with the empirical-Bernstein(EB) / Hoeffding(Hoef.) crossover near $n \approx 300$; (c) Proposition 1 illustrates variance adaptation across risk levels.

Data use and ethical scope. The experiments use only the publicly released EyePACS and APTOS challenge datasets and involve no new participant recruitment or prospective patient data collection. Our study is a retrospective methodological evaluation of risk-certification procedures for a fixed transferred model. The resulting certificate should not be interpreted as replacing prospective clinical validation, regulatory review, or site-specific clinical governance.

Population shift and certificate validity. The transferred risk changes substantially across cohorts. EyePACS→APTOS improves from held-out source Brier 0.124 to target risk 0.105, whereas APTOS→EyePACS inflates from 0.066 to 0.174 (2.64×). Table 1 shows that empirical-Bernstein covers the true target risk in at least 99.8% of audits against the nominal 90%. The bound is conservative at small n, but shrinks with audit size: on the harder target, slack falls from 0.096 at $n = 200$ to 0.026 at $n = 1600$.

From coverage to deployment readiness. Table 1 can also be read as a decision-support table: the mean certificate is approximately target risk plus mean slack. For EyePACS→APTOS, the empirical-Bernstein certificate decreases from about 0.235 at $n = 100$ to 0.124 at $n = 1600$; for the harder APTOS→EyePACS transfer, it remains higher, decreasing from about 0.330 to 0.200. A clinic can compare these values with its acceptable-risk threshold α.

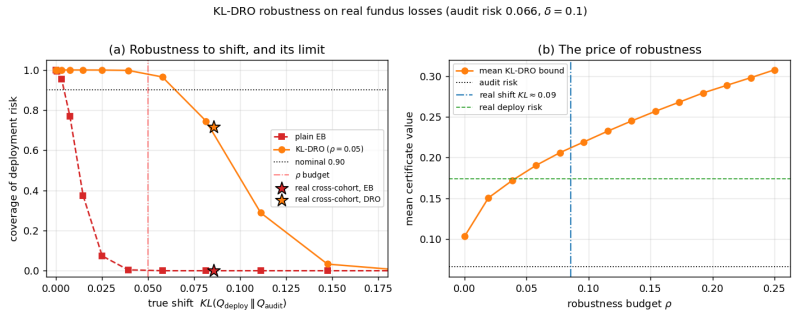

Fig. 2. KL-DRO robustness. Stars mark APTOS→EyePACS (0.066 → 0.174, KL≈0.09). Audit-only coverage fails under tilts; KL-DRO holds within budget, with larger ρ yielding looser certificates.

Empirical-Bernstein vs. Hoeffding. Empirical-Bernstein helps most when local loss variance is small. In Fig. 1c its slack is 0.66× Hoeffding's at risk 0.01, and the bounds cross near 0.05; at our observed risks, Hoeffding is sometimes marginally tighter. For a single deployed certificate we therefore use $\min(\hat{U}^{\mathrm{EB}}_{\delta/2}, \hat{U}^{\mathrm{Hoef}}_{\delta/2})$, valid by a union bound.

Shift-robustness on real losses. Figure 2 stress-tests Proposition 2 using exponential tilts of an audited APTOS loss law. This experiment asks what happens when the labelled audit stream is easier than the future deployment stream. The plain audit-only bound rapidly loses coverage under tilt, whereas KL-DRO with $\rho = 0.05$ remains valid up to its budget. The genuine APTOS→EyePACS shift has audit risk 0.066, deployment risk 0.174, and estimated KL ≈ 0.09; the audit-only certificate covers it in 0% of trials, while $\rho \approx 0.10$ restores coverage. This retrospective experiment illustrates the operational meaning of ρ: increasing the robustness budget allows the certificate to tolerate progressively larger audit–deployment shifts, at the cost of a looser upper risk bound. It should not be interpreted as a procedure for estimating the unknown future value of ρ.

What the audit tells the clinic. The audit does not improve the transferred model; it asks how large the local risk could still be, given the labelled cases observed so far. In the easier EyePACS→APTOS direction, the certificate becomes reasonably tight with more audit labels. In the harder APTOS→EyePACS direction, it remains higher, warning that additional validation, local recalibration, or replacement may be needed before deployment.

5 Discussion and RCS impact

This work formulates frozen-model deployment in resource-constrained settings as finite-sample risk certification. The method requires only a fixed transferred model, a small labelled local audit, and bounded per-example losses; it makes no

parametric assumption on the deployment distribution and can be robustified with a KL-DRO margin over loss laws when the audit stream may differ from future deployment. After the frozen model is run on the audit images, the certificate operates only on one-dimensional losses, avoiding target-domain retraining, GPU optimisation, local hyperparameter search, and density-ratio estimation in high-dimensional image space. A site can compare the certificate with its acceptable risk threshold to decide whether the model is acceptable, whether more labels are needed, or whether an audit–deployment robustness margin is necessary.

Scope and limitations. The plain certificate is only as credible as the audit sampling process: its distribution-free guarantee requires the audit to represent the intended deployment stream. Selection bias or temporal workflow changes can therefore invalidate the i.i.d. interpretation. The KL-DRO extension protects against audit–deployment mismatch only when the true deployment loss law lies inside the chosen radius, and sufficiently large values of ρ can make the certificate too conservative to support deployment.

The empirical study is intentionally scoped to one clinical task—referable diabetic-retinopathy detection—and one primary certified loss, the Brier score, evaluated across two cohorts and both transfer directions. These experiments demonstrate the certificate mechanics under a genuine cross-population shift, but they do not establish empirical generality across imaging modalities, diseases, or clinical loss functions. Future validation should therefore include additional modalities and tasks, clinically weighted losses such as false-negative-sensitive screening costs, and conventional clinical metrics including sensitivity, specificity, AUROC, and calibration error. The proposed framework is not a replacement for prospective clinical validation; rather, it provides a transparent finite-sample risk-control layer for deciding whether a frozen transferred model has sufficient local evidence to proceed to the next stage of deployment evaluation.

Acknowledgments. The first author gratefully acknowledges the University Grants Commission (UGC), India, for providing the research fellowship that supported this work. The DST-FIST grant SR/FST/MS-1/2019/45 is acknowledged for providing the computing facility in the Department of Mathematics at the Indian Institute of Technology Delhi.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Angelopoulos, A., Bates, S., Fisch, A., Lei, L., Schuster, T.: Conformal risk control. In: International conference on learning representations. vol. 2024, pp. 55198–55218 (2024)
2. Angelopoulos, A.N., Bates, S., Fannjiang, C., Jordan, M.I., Zrnic, T.: Prediction-powered inference. Science **382**(6671), 669–674 (2023)
3. Asia Pacific Tele-Ophthalmology Society: APTOS 2019 blindness detection. Kaggle competition dataset (2019), `https://www.kaggle.com/c/aptos2019-blindness-detection`

4. Bates, S., Angelopoulos, A., Lei, L., Malik, J., Jordan, M.: Distribution-free, risk-controlling prediction sets. Journal of the ACM (JACM) **68**(6), 1–34 (2021)
5. Bian, M., Barber, R.F.: Training-conditional coverage for distribution-free predictive inference. Electronic Journal of Statistics **17**(2), 2044–2066 (2023)
6. Bickel, S., Brückner, M., Scheffer, T.: Discriminative learning under covariate shift. Journal of Machine Learning Research **10**(9) (2009)
7. Bretagnolle, J., Huber, C.: Estimation des densités: risque minimax. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **47**(2), 119–137 (1979)
8. Cauchois, M., Gupta, S., Ali, A., Duchi, J.C.: Robust validation: Confident predictions even when distributions shift. Journal of the American Statistical Association **119**(548), 3033–3044 (2024)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
10. Donsker, M.D., Varadhan, S.S.: Asymptotic evaluation of certain markov process expectations for large time, i. Communications on pure and applied mathematics **28**(1), 1–47 (1975)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
12. Duchi, J.C., Namkoong, H.: Learning models with uniform performance via distributionally robust optimization. The Annals of Statistics **49**(3), 1378–1406 (2021)
13. EyePACS, California Healthcare Foundation: Diabetic retinopathy detection. Kaggle competition dataset (2015), `https://www.kaggle.com/c/diabetic-retinopathy-detection`
14. Glenn, W.B.: Verification of forecasts expressed in terms of probability. Monthly weather review **78**(1), 1–3 (1950)
15. Hoeffding, W.: Probability inequalities for sums of bounded random variables. Journal of the American statistical association **58**(301), 13–30 (1963)
16. Maurer, A., Pontil, M.: Empirical bernstein bounds and sample variance penalization. arXiv preprint arXiv:0907.3740 (2009)
17. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. the Journal of machine Learning research **12**, 2825–2830 (2011)
19. Wightman, R.: Pytorch image models. `https://github.com/huggingface/pytorch-image-models` (2019). https://doi.org/10.5281/zenodo.4414861
20. Wilkinson, C.P., Ferris, F.L., Klein, R.E., Lee, P.P., Agardh, C.D., Davis, M., Dills, D., Kampik, A., Pararajasegaram, R., Verdaguer, J.T.: Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology **110**(9), 1677–1682 (2003). https://doi.org/10.1016/S0161-6420(03)00475-5
21. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., et al.: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023)