SPATIAL RADIUS AND GROUP COVERAGE OF CONFORMAL IMAGE SETS UNDER COVARIATE SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing a high-resolution image from coarse measurements leaves uncertainty about regional averages. A prediction set can contain the true image with high probability while allowing large differences between accepted images. In this article, we study the smallest and largest averages over all accepted images that match the observed coarse measurements. These ranges give simultaneous bounds for every region whenever the set contains the true image. We derive exact ranges for nested regions and construct an accepted reconstruction with the smallest worst-case regional error. Standard calibration within observed groups preserves coverage when group proportions change and within-group distributions remain unchanged. This guarantee does not require an accurate predictor. For smoothly varying images, a fixed number of regional averages controls approximation error independently of resolution. The calibrated threshold also contributes to the image radius. In controlled chest X-ray experiments, calibrating regional averages without individual pixels reduces the largest regional range by 20% to 44% compared with calibrating both. Retinal experiments show that calibration within grades improves the smallest mean group coverage over marginal calibration.

1 INTRODUCTION

We consider reconstructing a high-resolution image from a low-resolution observation. In our main example, a 128×128 chest radiograph is observed only through the averages of its 4×4 pixel blocks, and a deep ensemble (Lakshminarayanan et al., 2017) returns several reconstructions. Image-response regression and conditional generative sampling also return such images (Zhang et al., 2023; Zhou et al., 2023; Song et al., 2026). Public retinal and chest X-ray data motivate our experiments (Yang et al., 2023; Liu et al., 2022; Wang et al., 2017; Demner-Fushman et al., 2016; Kermany et al., 2018). Clinical status is observed at prediction and defines groups whose proportions can differ between the images used to calibrate the prediction rule and the images to be predicted. A regional average, such as the mean intensity of a lung zone, remains uncertain after reconstruction, and its uncertainty must be reported for each image.

A prediction set is a collection of possible images, and coverage is the probability that it contains the true image. Conformal prediction gives marginal coverage, averaged over calibration and test observations, when reordering observations leaves their joint distribution unchanged (Vovk et al., 2005; Shafer & Vovk, 2008; Lei & Wasserman, 2014; Lei et al., 2018; Romano et al., 2019). For images, risk-control methods calibrate pixel intervals by bounding the expected fraction of uncovered pixels (Bates et al., 2021; Angelopoulos et al., 2022; 2024; Teneggi et al., 2023; Horwitz & Hoshen, 2022). Other conformal methods calibrate masks that localize unreliable regions (Kutiel et al., 2023; Adame et al., 2025; Silva-Rodríguez & Konukoglu, 2026), semantic latent factors (Sankaranarayanan et al., 2022), and scalar outputs derived from reconstructed or segmented images (Wen et al., 2025; Cheung et al., 2026a;c). Cheung et al. (2026b) analyze how additive bias changes the lengths of these scalar intervals. Principal-direction regions and spatial decompositions account for dependence in restoration and segmentation (Belhasin et al., 2024; Viti et al., 2026). Nehme et al. (2023) estimate posterior principal directions with a neural network. Segmentation, generative, functional, and multiscale sets calibrate masks, sample counts, bands, or scales (Davenport, 2024; Mossina & Friedrich, 2026; Bereska et al., 2026; Brunekreef et al., 2024; Lu et al., 2026; Shahrokhi et al., 2025; Chernozhukov et al., 2021; Diquigiovanni et al., 2022; Gray et al., 2025; Millard et al., 2026; Baheri & Amiri Shahbazi, 2025), and expected-size and volume-optimality analyses study prescribed set

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

families (Dhillon et al., 2024; Gao et al., 2025). A small expected fraction of uncovered pixels does not imply that every region of an image is covered, and separate score intervals do not impose the observed coarse averages or the constraints that regions satisfy jointly. The smallest and largest values of each regional average over all accepted images that match the observation give simultaneous statements for one image. We derive exact ranges for nested regional averages under these observation and score constraints.

Weighted conformal prediction adapts marginal coverage to covariate shift through a density ratio (Tibshirani et al., 2019), following density-ratio methods for transfer (Shimodaira, 2000; Sugiyama et al., 2007; Kanamori et al., 2009; Sugiyama et al., 2012). Estimated weights, doubly robust calibration, divergence-based protection, and data-dependent shift analysis extend the settings with target coverage (Lei & Candès, 2021; Yang et al., 2024; Cauchois et al., 2024; Fannjiang et al., 2022), and methods for nonexchangeable data, online shifts, and unknown shifts control coverage under changing distributions (Barber et al., 2023; Gibbs & Candès, 2024; Qiu et al., 2023). Group-wise validity, localization, and conditional histograms address conditional calibration (Vovk, 2013; Guan, 2023; Sesia & Romano, 2021). Linear conditional calibration, multivalid prediction, and Kandinsky conformal prediction control coverage over specified function classes (Gibbs et al., 2025; Jung et al., 2023; Bairaktari et al., 2025), and Min et al. (2026) give a nonasymptotic analysis that covers structured data. When only group proportions change, calibration within groups keeps coverage in every group, and Bhattacharyya & Barber (2026) study weighting for this shift. We use this standard group calibration and study the size of the calibrated set.

In this article, we study the exact regional ranges of conformal image sets. The predictor returns K images for a query, and the score is the largest standardized deviation of a candidate image from their mean over a class of nested regional averages. Thresholds are calibrated within observed groups. For every nested region, we compute the smallest and largest average over all images that pass the calibrated threshold and match the observed coarse image. If the set contains the true image, these ranges contain all its regional averages at once, and half the widest range is the exact radius of the set.

This article makes three contributions. First, we derive the exact regional ranges by one upward and one downward pass over the partition tree, in time linear in the number of regions. We also construct an accepted reconstruction that matches the observation and has the smallest worst-case regional error, at most $3/2$ times the exact radius when a single level of four subregions is unknown and at most twice the radius in general. Second, we relate the radius to the number of linear measurements. For smoothly varying images, a number of regional averages independent of resolution controls the omitted detail. A small normalized image radius for unstructured images requires a number of linear measurements proportional to the pixel count when the accepted set contains an image with intensities away from zero and one. For sets containing all regular images with accepted linear measurements, we obtain the same approximation exponent under strict intensity and regularity bounds. These results concern linear approximation (Pinkus, 1985). Third, we evaluate the sets with a deep ensemble for fourfold super-resolution of chest radiographs. The group-calibrated sets contain 0.901 of target images across all regional averages down to 2×2 regions, with coverage 0.905 and 0.900 in the two diagnostic groups. The uncalibrated ensemble band contains no target image, and regional intervals from conformal risk control are narrower but cover every region of no target image. Among the rules with simultaneous coverage, the sets with a constant scale have the smallest radius, 0.140, and those standardized by the ensemble spread have the smallest mean regional half-width, 0.104. Calibrating on regional averages alone, without single pixels, reduces the largest regional range by 20% to 44% on the same replications. The accepted reconstruction matches the observation, its mean absolute regional error 0.0082 is close to 0.0081 for the ensemble mean, and with a constant scale its worst-case regional error over the set is 0.143. In five controlled cohorts of retinal and chest X-ray images, calibration within groups keeps mean group coverage near the nominal level, and a public masked autoencoder (He et al., 2022) illustrates an externally fitted predictor.

The regional ranges assume that the observed coarse averages are exact. The coverage guarantee concerns changes in group proportions. A change in the within-group image distribution enters through a separate error term, and it lowered group coverage in official source and test splits. The remainder of the article is organized as follows. Section 2 defines the procedures, Section 3 gives the theoretical results, Section 4 reports the experiments, and Section 5 discusses their scope. The Appendix gives the proofs, the complete formulation, supporting theory, and numerical details.

2 PROBLEM, SPATIAL TARGET, AND CALIBRATION

We formulate the image reconstruction problem with observed covariates. Let $X \in \mathcal{X} \subset \mathbb{R}^p$ be a vector of p observed covariates and $Y \in \mathcal{Y} \subset \mathbb{R}^V$ the image with V pixels. The source and target joint distributions $P_{\mathrm{src}}, P_{\mathrm{tar}}$ have covariate marginals $P_{\mathrm{src},X}, P_{\mathrm{tar},X}$. We observe n labeled source pairs $\mathcal{D}_{\mathrm{src}} = \{(X_i, Y_i)\}_{i=1}^n$. Specifically, covariate shift leaves the conditional image distribution $P_{Y|X=x}$ unchanged:

$$P_{\mathrm{src}}(dx, dy) = P_{\mathrm{src},X}(dx)P_{Y|X=x}(dy), \qquad P_{\mathrm{tar}}(dx, dy) = P_{\mathrm{tar},X}(dx)P_{Y|X=x}(dy). \tag{1}$$

Here, a nonnegative function f with $\mathbb{E}_{P_{\mathrm{src},X}} f(X) = 1$ defines the shifted population $dP_f(x, y) = f(x)dP_{\mathrm{src}}(x, y)$. We write $\mathbb{E}_f$ for expectation under P_f. Thus, f changes covariate frequencies, and $\mathcal{F}$ contains the shifts for which coverage is required.

In this framework, a spatial functional, such as the average intensity in a region, assigns a number to each candidate image. For a class $\mathcal{H}$ of spatial functionals and a nonempty set $C \subseteq \mathcal{Y}$, define

$$\mathrm{rad}_{\mathcal{H}}(C) = \inf_{y_0 \in \mathcal{Y}} \sup_{y \in C} \sup_{h \in \mathcal{H}} |h(y) - h(y_0)|. \tag{2}$$

We define $\mathrm{rad}_{\mathcal{H}}(\emptyset) = 0$. Note that the radius is the largest deviation from the best center, measured through $\mathcal{H}$; regional averages localize this uncertainty, and normalized linear projections measure it in a full-image norm.

At a query x, the predictor returns K independent images from its conditional distribution Q_x^{T} of reconstructions. For each spatial functional h, the sample mean and standard deviation of h over the K returned images give $\widehat{\mu}_h(x)$ and $\widehat{\sigma}_h(x)$, with divisor K for the variance. We add a fixed $\lambda > 0$ to prevent division by zero. For a threshold $q \geq 0$, define

$$R_n(x, y) = \sup_{h \in \mathcal{H}} \frac{|h(y) - \widehat{\mu}_h(x)|}{\widehat{\sigma}_h(x) + \lambda}, \qquad \mathcal{C}_q(x) = \{y \in \mathcal{Y} : R_n(x, y) \leq q\}. \tag{3}$$

The score in (3) is the largest standardized deviation across the functionals. Thus $\mathcal{C}_q(x)$ imposes $|h(y) - \widehat{\mu}_h(x)| \leq q\{\widehat{\sigma}_h(x) + \lambda\}$ for every $h \in \mathcal{H}$, with wider intervals for greater predictor spread. Predictor fitting, spatial summaries, and preprocessing are fixed using training data before calibration responses are used. Predictor randomization is independent of the calibration sample and identical for calibration and test queries.

In the medical reconstruction experiments, $Y \in [0, 1]^V$ is the fine image. In the five cohorts of Section 4.2, the covariate vector X consists of the 4×4 array $Z = (z_1, \ldots, z_{16})$ of observed coarse block averages and the clinical status G. The class $\mathcal{H}$ contains the 84 averages over the 2×2, 4×4, and 8×8 dyadic partitions, where the 64 finest are not determined by Z. With $\bar{h}_b \in \mathcal{H}$ denoting the average over coarse block b, the set at threshold q restricted to images matching the observation is

$$C_Z(x) = \Big\{ y \in [0, 1]^V : \bar{h}_b(y) = z_b \ (b = 1, \ldots, 16), \ \ |h(y) - \widehat{\mu}_h(x)| \leq q\{\widehat{\sigma}_h(x) + \lambda\} \ (h \in \mathcal{H}) \Big\}.$$

The true image satisfies these equalities, and the restriction leaves the coverage event unchanged. If $C_Z(x)$ is nonempty, each fine-region average has a closed-form range over $C_Z(x)$, which Proposition 2 turns into the exact radius. The neural study of Section 4.1 uses the same construction with 32×32 observed blocks and the 5,461 nested dyadic averages down to 2×2 regions.

Let $0 < \alpha < 1$ be a desired exclusion probability, called the miscoverage level. We split the labeled source data into training data $\mathcal{D}_{\mathrm{tr}}$ and calibration data $\mathcal{D}_{\mathrm{cal}}$ indexed by $\mathcal{I}_{\mathrm{cal}}$ with $n_{\mathrm{cal}} = |\mathcal{I}_{\mathrm{cal}}|$, and a test observation (X_*, Y_*) has score $R_* = R_n(X_*, Y_*)$. Let $G(X) \in \{1, \ldots, L\}$ be an observed group label fixed before calibration. In group g, we order the n_g calibration scores as $R_{g,(1)} \leq \cdots \leq R_{g,(n_g)}$ and define $R_{g,(n_g+1)} = +\infty$. We compute

$$k_g = \lceil (n_g + 1)(1 - \alpha) \rceil, \qquad \widehat{q}_g = R_{g,(k_g)}, \qquad \widehat{\mathcal{C}}_G(x) = \mathcal{C}_{\widehat{q}_{G(x)}}(x). \tag{4}$$

Equation (4) accepts an image when its score does not exceed the k_gth smallest calibration score of its own group. Each group has one deterministic threshold, with no target responses used. Groups with fewer than $1/\alpha - 1$ calibration observations receive an infinite threshold.

Algorithm 1 Group calibration of image prediction sets

1: Fix observed groups, spatial functionals, and preprocessing using training data.
2: Query the predictor using only information available at prediction.
3: Compute the calibration scores in (3).
4: Within each group, compute the corrected order statistic in (4).
5: Return $\widehat{\mathcal{C}}_G(x)$ for a query in group $G(x)$.

For a broader shift class and tolerance $\tau_n \geq 0$, let $R_i = R_n(X_i, Y_i)$. With $\mathbf{1}\{\cdot\}$ denoting an indicator, define

$$\Gamma_n^+(q; \mathcal{F}) = \sup_{f \in \mathcal{F}} \frac{1}{n_{\text{cal}}} \sum_{i \in \mathcal{I}_{\text{cal}}} f(X_i)\{\mathbf{1}\{R_i > q\} - \alpha\}, \qquad \widehat{q} = \inf\{q \geq 0 : \Gamma_n^+(q; \mathcal{F}) \leq \tau_n\}. \quad (5)$$

If the calibration scores are finite, the set $\widehat{\mathcal{C}}(x) = \mathcal{C}_{\widehat{q}}(x)$ uses the smallest q for which the sample average of $f(X_i)\{\mathbf{1}\{R_i > q\} - \alpha\}$ is at most τ_n for every $f \in \mathcal{F}$. For group indicators, $\max_g \widehat{q}_g$ provides a common conservative threshold, and each group threshold is at most this common value.

The medical experiments use Algorithm 1, and the resolution experiment uses the same group rule with the projected-norm score in Section 3. Appendix S6.1 describes two additional calibrators. The first is a candidate-augmented rule with coverage exactly $1 - \alpha$ for finite-dimensional shift classes. It has the target of the randomized augmented quantile regression of Gibbs et al. (2025) and balances the same moments with a mixture of thresholds. This exact coverage is averaged over calibration samples, the target draw, and a uniform random variable U_0 on $(0, 1)$, independent of all data. The second is a static mixture that draws one calibrated threshold. Exact coverage alone does not control size. For example, returning $\mathcal{Y}$ with probability $1 - \alpha$ and $\emptyset$ with probability α gives exact coverage for every distribution, but its nonempty set has radius $\text{rad}_{\mathcal{H}}(\mathcal{Y})$. Recall that the task is conditional reconstruction given clinical status, not diagnosis. A prevalence change is a shift in X that does not by itself establish the common conditional distribution in (1). Proposition 3 quantifies its failure.

3 Theoretical Results

Coverage depends on the calibration population and the protected groups, whereas image-set size also depends on the information in the spatial summaries. Within-group exchangeability means that reordering scores within a group leaves their joint distribution unchanged. Calibration within groups is the Mondrian construction of Vovk et al. (2005), and Bhattacharyya & Barber (2026) note that a group-conditional guarantee implies coverage under every change of group proportions. The following theorem restates both facts for the image score. It requires no condition on predictor accuracy and holds conditionally on the training sample and the predictor state fixed before calibration. The query randomness is part of the score distribution.

Theorem 1. *Suppose calibration and test scores are exchangeable within every observed group conditional on $\mathcal{D}_{\text{tr}}$ and the calibration group labels. Then, for every g with $P_{\text{src}}(G = g) > 0$,*

$$\mathbb{P}\{Y_* \in \widehat{\mathcal{C}}_G(X_*) \mid G(X_*) = g, \mathcal{D}_{\text{tr}}\} \geq 1 - \alpha. \quad (6)$$

For every shift $f(x) = a_{G(x)}$ with $a_g \geq 0$ and $\mathbb{E}_{\text{src}} f(X) = 1$, it holds that $P_f\{Y_* \in \widehat{\mathcal{C}}_G(X_*) \mid \mathcal{D}_{\text{tr}}\} \geq 1 - \alpha$.

Theorem 1 gives a coverage guarantee for every mixture with an unchanged within-group joint distribution of images and covariates. This guarantee does not require an accurate predictor. Coverage in (6) averages over calibration samples, and a given calibration sample can fall short. The following proposition uses conditionally independent and identically distributed observations, abbreviated as i.i.d., to bound the shortfall for a fixed calibration sample, using a lower bound $\pi_0 > 0$ on all group probabilities. Write $t_+ = \max\{t, 0\}$ and fix $0 < \delta < 1$.

Proposition 1. *Suppose the calibration observations are conditionally i.i.d., $P_{\text{src}}(G = g) \geq \pi_0 > 0$ for every g, and $n_{\text{cal}}\pi_0 \geq 8 \log(2L/\delta)$. Then, with probability at least $1 - \delta$,*

$$\max_g \left[(1 - \alpha) - P_{\text{src}}\{Y \in \widehat{\mathcal{C}}_G(X) \mid G = g, \mathcal{D}_{\text{cal}}, \mathcal{D}_{\text{tr}}\}\right]_+ \leq \sqrt{\frac{\log(4L/\delta)}{n_{\text{cal}}\pi_0}}.$$

Proposition 1 shows that, for a fixed calibration sample, every group's coverage shortfall is of order $(n_{\mathrm{cal}}\pi_0)^{-1/2}$ up to a logarithmic factor. Its proof also gives a bound based on observed group counts under conditional i.i.d. sampling alone.

For the generic rule, take an i.i.d. source calibration sample and write $\Delta^-_{\mathcal{F}}(C) = \sup_{f\in\mathcal{F}}[(1-\alpha) - P_f\{Y \in C(X)\}]_+$ for the largest coverage shortfall over the shift class. For $B_F > 0$, $A_F \geq 2$, $v_F \geq 0$, and $0 < \delta < 1$, set $\psi_n(\delta, \mathcal{F}) = CB_F\{[1 + (v_F + 2)\log(A_F n_{\mathrm{cal}}) + \log(1/\delta)]/n_{\mathrm{cal}}\}^{1/2}$, where C is a universal constant. The following theorem bounds the largest undercoverage over $\mathcal{F}$. For a probability distribution Q on $\mathcal{X}$, the $L_2(Q)$ covering number is the smallest number of functions needed to approximate every shift to a given root mean squared accuracy under Q. A function class is pointwise measurable if it has a countable pointwise dense subclass.

Theorem 2. *Suppose $\mathcal{F}$ is pointwise measurable, $0 \leq f \leq B_F$ for every $f \in \mathcal{F}$, and its $L_2(Q)$ covering number at radius ηB_F is at most $(A_F/\eta)^{v_F}$, uniformly over probability measures Q and $0 < \eta \leq 1$. If the score is finite, measurable, and fixed independently of the calibration observations, then, with probability at least $1-\delta$,*

$$\Delta^-_{\mathcal{F}}(\widehat{\mathcal{C}}) \leq \tau_n + \psi_n(\delta, \mathcal{F}). \tag{7}$$

Theorem 2 bounds the coverage shortfall by the tolerance plus the calibration error $\psi_n(\delta, \mathcal{F})$, which is of order $n_{\mathrm{cal}}^{-1/2}$ up to a logarithmic factor and an unspecified universal constant; a larger tolerance increases the shortfall bound and can reduce set size. Appendix Theorem S14 gives the uniform empirical-process bound used in the proof; Appendix Theorem S8 and Corollary S1 localize the threshold near the worst-case population quantile under a one-sided quantile condition and control overcoverage under sufficient conditions.

3.1 Predictor error and set size

We bound changes in radius due to predictor bias and finitely many returned images. The coverage guarantee needs neither. Let $A(x)$ bound the change in radius per unit change of threshold for two set families. If their scores differ by at most $e \geq 0$ everywhere, their finite thresholds on the same calibration sample differ by at most e and their radii by at most $2A(x)e$, by Appendix Proposition S2. When returned images preserve the known constraints, Appendix Theorem S2 bounds the difference between expected radii based on K returned images and on source conditional moments by predictor bias and a query term of order $K^{-1/2}$ up to a logarithmic factor, under any covariate distribution. Appendix S1.9 enforces the coarse observation in the returned images. The rate $K^{-1/2}$ cannot be improved for the threshold in a two-point family (Appendix Theorem S20 and Proposition S33), and Appendix Theorem S11 and Proposition S20 give finite-support and uniform-in-x versions.

To compare resolutions, we denote the Euclidean norm by $\|\cdot\|_2$. We use the root mean squared intensity norm $\|y\|_V = V^{-1/2}\|y\|_2$ and its radius rad_V, which equals (2) for $\mathcal{H} = \{y \mapsto V^{-1/2}u^\top y : u \in \mathbb{R}^V, \|u\|_2 \leq 1\}$. On a $2^J \times 2^J$ grid with integer $J \geq 0$, $V = 4^J$ and P_j replaces pixels by averages over $M = 4^j$ dyadic blocks for $j = 0, \ldots, J$. With I denoting the identity map, we define the regular class $\mathcal{Y}^{(V)}_{s,R}$ as the set of images in $[0,1]^V$ satisfying $\|(I - P_j)y\|_V \leq R2^{-sj}$ for every j, where $s, R > 0$ remain fixed across resolutions. This bounds the detail omitted by the level-j block averages and holds for regular-grid samples of Hölder functions of order $0 < s \leq 1$, for a suitable R, by Appendix Lemma S1.

For a center $\widehat{\mu}(x) \in \mathcal{Y}^{(V)}_{s,R}$, let $\widehat{\mathcal{C}}_j(x)$ be the set of candidates $y \in \mathcal{Y}^{(V)}_{s,R}$ with $\|P_j\{y - \widehat{\mu}(x)\}\|_V$ at most its calibrated threshold $\widehat{q}_j(x)$. For arbitrary M, let $A : \mathbb{R}^V \to \mathbb{R}^M$ be a linear map, and let $\mathcal{B}(x) \subseteq \mathbb{R}^M$ be an acceptance region. We define $C^{(M)}(x) = \{y \in [0,1]^V : Ay \in \mathcal{B}(x)\}$. Let $0 < \rho \leq 1/2$ be a fixed constant. The following theorem bounds the projected set from above and every linear reduction on the full box from below.

Theorem 3. *For $\widehat{\mathcal{C}}_j$ defined above,*

$$\mathrm{rad}_V\{\widehat{\mathcal{C}}_j(x)\} \leq \widehat{q}_j(x) + 2R2^{-sj}. \tag{8}$$

For $0 < \epsilon \leq 2R$, a remainder at most ϵ is attained with

$$M \leq \min\{V, 4(2R/\epsilon)^{2/s}\}. \tag{9}$$

If $M < V$ and $C^{(M)}(x) \cap [\rho, 1-\rho]^V \neq \emptyset$, then

$$\mathrm{rad}_V\{C^{(M)}(x)\} \geq \rho\sqrt{1 - M/V}. \tag{10}$$

Theorem 3 distinguishes regular and unstructured images. For regular images, the class bounds the omitted detail $(I - P_j)y$, and the budget in (9) is independent of resolution. For unstructured images, a radius $r < \rho$ requires $M \geq V\{1 - (r/\rho)^2\}$ by (10), including for fitted linear maps. Without the interior condition, coverage $1 - \alpha$ and uniform pixel marginals give an expected radius of at least $(1 - \alpha)^2(1 - M/V)^{3/2}/4$ by Appendix Theorem S4. This lower bound requires $M/V \to 1$ for convergence to zero. For any summary, including nonlinear ones, Appendix Theorem S6 gives half the largest spatial difference among images sharing an accepted summary as a lower bound, which can equal zero. Orthogonal projection sharpens (8) to $\{\widehat{q}_j(x)^2 + R^2 2^{-2sj}\}^{1/2}$ in Appendix S1.11.2, and Appendix Theorem S12 provides a Haar version. Under intensity and regularity margins, Appendix S1.12 gives the same approximation exponent from below within the regular class, preserving the coarse observation. Appendix Proposition S6 covers Hölder images. Note that the calibrated threshold remains in the upper bound.

3.2 Exact regional radius under the observed coarse image

We use the maximum-score set $C_Z(x)$ of Section 2 in the medical analysis. When the set is nonempty, its constraints determine the exact regional ranges and the best accepted image.

Let m_j be the predictor mean of fine region j and a_j its standard deviation plus λ. Let $\mathcal{B}_b$ index the four fine regions in coarse block b with observed average z_b. For $j \in \mathcal{B}_b$ and $q \geq 0$, we define

$$l_j = \max\{0, m_j - qa_j\}, \qquad u_j = \min\{1, m_j + qa_j\},$$
$$L_j = \max\Big\{l_j, 4z_b - \sum_{k \in \mathcal{B}_b \setminus \{j\}} u_k\Big\}, \qquad U_j = \min\Big\{u_j, 4z_b - \sum_{k \in \mathcal{B}_b \setminus \{j\}} l_k\Big\}.$$

The first pair is the Section 2 interval constraint clipped to $[0, 1]$. The second pair tightens this constraint because the four fine-region averages in block b sum to $4z_b$. Furthermore, the score constraints of the 20 coarser averages are constant on $C_Z(x)$. They either exclude every candidate or impose no restriction. In contrast, the comparison L_2 score uses the Euclidean norm of all 84 standardized regional residuals. The following proposition gives the exact radius.

Proposition 2. *If $C_Z(x) \neq \emptyset$, then*

$$\mathrm{rad}_{\mathcal{H}}\{C_Z(x)\} = \frac{1}{2} \max_{1 \leq j \leq 64} (U_j - L_j). \tag{11}$$

For the corresponding L_2 score set, the right side is an upper bound.

Proposition 2 gives the radius reported in Table 2, and the group-L_2 comparison uses an upper bound. The center of this set need not match the observation. Appendix Proposition S1 constructs an accepted center with optimal regional error, at most $3/2$ times this radius. Appendix Theorem S1 gives exact ranges for nested regions of unequal sizes and for several unknown levels, as in the initial neural design of Appendix S7, by one upward and one downward pass with time linear in the number of regions. Appendix Proposition S4 bounds the radius difference between two nonempty intersections. This difference is at most three times the largest sum of absolute changes in m_j and qa_j across fine regions. This result holds without any lower bound on the conditional standard deviation. At the boundary of feasibility, the radius can jump from zero to $1/2$, and empty intersections are hence counted separately. A radius of zero alone does not imply an accurate predictor.

The shift class need not contain the true target population. When $P_{\mathrm{tar},X} \ll P_{\mathrm{src},X}$, we define $r = dP_{\mathrm{tar},X}/dP_{\mathrm{src},X}$ and $\kappa(r, \mathcal{F}) = \frac{1}{2} \inf_{f \in \mathcal{F}} \mathbb{E}_{P_{\mathrm{src},X}} |r(X) - f(X)|$. The following theorem gives the coverage cost of approximating the target density ratio by the protected shift class.

Theorem 4. *Suppose the target distribution satisfies (1) and $P_{\mathrm{tar},X} \ll P_{\mathrm{src},X}$. Then,*

$$P_{\mathrm{tar}}\{Y \in \widehat{\mathcal{C}}(X)\} \geq 1 - \alpha - \Delta^-_{\mathcal{F}}(\widehat{\mathcal{C}}) - \kappa(r, \mathcal{F}). \tag{12}$$

Table 1: Neural $4\times$ super-resolution: coverage and exact regional ranges.

Method	Coverage	Normal	Pneumonia	Radius	2×2 half-width
Proposed, standardized scale	0.901	0.905	0.900	0.348	0.104
Proposed, constant scale	0.899	0.905	0.897	0.140	0.139
One pooled threshold	0.909	0.895	0.913	0.352	0.105
Uncalibrated ensemble	0.000	0.000	0.000	0.078	0.022
Regionwise risk control	0.000	0.000	0.000	0.064	0.018

Theorem 4 bounds how far the target coverage can fall below the nominal level. This shortfall is at most the calibrated shortfall plus the total-variation distance from the true covariate distribution to the protected shift class. When (1) fails, a separate term is needed. We define

$$\zeta = \int \|P_{\mathrm{tar},Y|X=x} - P_{\mathrm{src},Y|X=x}\|_{\mathrm{TV}}\, P_{\mathrm{tar},X}(dx),$$

where $P_{\mathrm{tar},Y|X=x}$ and $P_{\mathrm{src},Y|X=x}$ are the target and source conditional image distributions. Recall that total variation is the largest absolute probability difference over measurable response sets. The following proposition separates conditional-distribution error from density-ratio approximation.

Proposition 3. *Suppose $P_{\mathrm{tar},X} \ll P_{\mathrm{src},X}$, and set $r = dP_{\mathrm{tar},X}/dP_{\mathrm{src},X}$. For every measurable prediction set C and shift class $\mathcal{F}$,*

$$P_{\mathrm{tar}}\{Y \in C(X)\} \geq 1 - \alpha - \Delta^{-}_{\mathcal{F}}(C) - \kappa(r, \mathcal{F}) - \zeta.$$

Proposition 3 separates a change in the covariate distribution from a change in the conditional image distribution. The former enters through $\kappa(r, \mathcal{F})$, whereas the latter enters through ζ.

4 Experiments

We ask three questions. How informative are the exact ranges for a neural super-resolution predictor, and what do other guarantees deliver on the same predictor? What does calibration within observed groups change when group proportions shift? How many regional summaries does a regular image class need as resolution grows? Every reported experiment uses 100 replications, and coverage is the realized frequency with which the prediction set contains the true image.

4.1 Neural super-resolution of chest radiographs

We use PneumoniaMNIST at resolution 224 (Yang et al., 2023; Kermany et al., 2018), resized to 128×128. The observation is the 32×32 array of block averages, a fourfold reduction in each direction, and the diagnosis, normal or pneumonia, defines the group. The predictor is a deep ensemble (Lakshminarayanan et al., 2017) of $K = 5$ U-Nets (Ronneberger et al., 2015) with independent initializations. It is fitted on 3,000 images disjoint from all calibration and target images, and its members are the returned images. The class $\mathcal{H}$ contains the 5,461 dyadic averages over squares with sides from 2 to 128 pixels. The 4,096 averages over 2×2 regions are unknown, and coarser averages are fixed by the observation. An initial design also certified the 16,384 single pixels. Their residuals set the threshold, and its radius 0.488 was close to the largest possible value $1/2$. After these results, we specified the present design using only a separate validation pool and evaluated it once on the same replications. On that pool, the ridge $\lambda = 0.0025$ minimizes the mean range, by the rule of the initial design. As λ grows, the set approaches that of the constant-scale score $\max_{h \in \mathcal{H}} |h(y) - \widehat{\mu}_h(x)|$, which minimizes the radius on the same pool, and we report both scales. Each replication draws 240 normal and 160 pneumonia calibration images and 100 normal and 300 pneumonia target images, which raises the pneumonia proportion from 40% to 75%. We compare the proposed sets with the same sets reported through separate per-functional intervals, one pooled threshold, an uncalibrated ensemble band, and 2×2 intervals from conformal risk control (Angelopoulos et al., 2024). Appendix S7 gives the architecture, training, comparisons, and both designs.

Table 1 reports means over 100 replications. With the standardized and constant scales, the proposed sets contain 0.901 and 0.899 of target images, and in these images every reported range contains

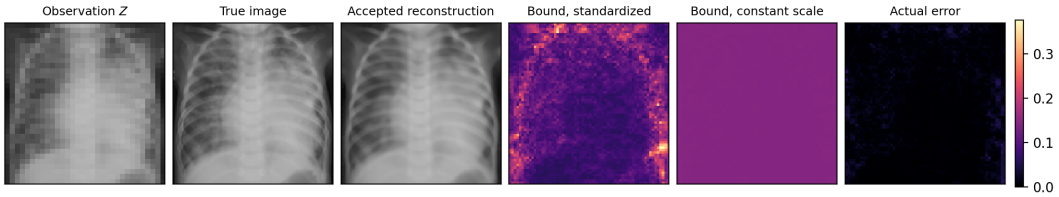


Figure 1: A target radiograph in the neural study. Each bound is the largest deviation of any accepted image from the accepted reconstruction for that scale in each 2×2 region, and the reconstruction and error use the standardized scale.

Table 2: Controlled prevalence changes: coverage and exact regional radius.

Cohort	Proposed cov.	Proposed rad.	Common cov.	Common rad.	Kandinsky cov.	Kandinsky rad.
Retina 64×64	0.906	0.128	0.967	0.160	0.891	0.122
Retina 128×128	0.906	0.124	0.967	0.156	0.891	0.118
ChestMNIST	0.902	0.274	0.909	0.278	0.899	0.272
NLM report cohort	0.903	0.261	0.912	0.266	0.895	0.244
ChestMNIST MAE	0.898	0.447	0.926	0.454	0.890	0.437

the true regional average. The standardized sets cover 0.905 of normal and 0.900 of pneumonia images, and one pooled threshold covers 0.895 and 0.913. The uncalibrated band at the normal maximum quantile contains no target image, because the ensemble spread understates the error. Regionwise risk control gives 2×2 intervals of mean half-width 0.018 and leaves 9.5% of regions uncovered, as its guarantee allows, but no target image has all regions covered. Among the rules with simultaneous coverage, the standardized scale gives the smallest mean half-width, 0.104, and the constant scale gives the smallest radius, 0.140. With single pixels also certified, as in the initial design, the largest 2×2 half-range rises to 0.437 and 0.249 for the two scales on the same replications, and excluding pixels lowers it in every replication. Separate conformal intervals for each region with a Bonferroni correction equal $[0, 1]$ unless a group has at least $40{,}959$ calibration images. The accepted reconstruction matches the observed block averages exactly, unlike the ensemble mean. Its certified error is 1.01 and 1.02 times the radius on average for the two scales and bounds the error of every region in every covered image. With the constant scale, the certified error averages 0.143 against a largest realized error of 0.087, and the mean absolute regional error 0.0081 matches that of the ensemble mean. Figure 1 shows one target image.

4.2 Group calibration in five controlled cohorts

We next compare calibration rules in five cohorts: RetinaMNIST at two resolutions, ChestMNIST, the National Library of Medicine (NLM) report cohort, and ChestMNIST with a public masked autoencoder (MAE). The observation is a 4×4 array of block averages, and $\mathcal{H}$ contains the 84 averages over the 2×2, 4×4, and 8×8 dyadic partitions. A ridge image predictor fitted on the training fold returns six images with independent pixel noise of standard deviation 0.04, and the MAE reconstructs masked patches of the upsampled coarse observation. Source and target images are drawn from one finite benchmark, and only the group composition changes, as in Theorem 1. Appendix S8 gives the counts, preprocessing, and predictors.

Table 2 reports coverage and the exact regional radius of Proposition 2. The mean coverage of the proposed rule within individual groups ranges from 0.895 to 0.909. With five retinal grades, calibrating within grades raises the smallest mean group coverage from 0.803 for marginal calibration, 0.810 for ratio weighting, and 0.818 for kernel weighting to 0.903 at resolution 64, and from 0.799, 0.810, and 0.812 to 0.901 at resolution 128. The common threshold is the largest group threshold, which keeps the guarantee of Theorem 1 in every group with one threshold. Its radius is at least that of the proposed rule by construction, and the proposed rule reduces the mean radius by 19.9% and 20.4% in the two retinal analyses and by at most 1.8% with two chest groups. The reported linear variant of Kandinsky calibration (Bairaktari et al., 2025) omits the finite-sample correction. Its mean coverage ranges from 0.890 to 0.899, below 0.9 in all five settings, and its mean radius

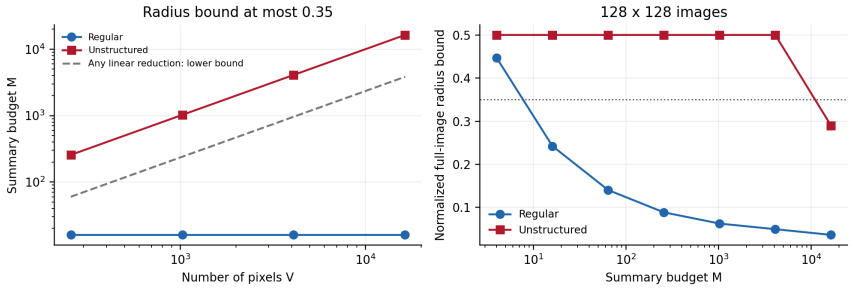


Figure 2: Summary budget and normalized image radius over 100 replications.

is slightly smaller. The uncalibrated plug-in rule covers 0.698, 0.631, and 0.027 of target images in ChestMNIST, the NLM cohort, and ChestMNIST MAE, against 0.902, 0.903, and 0.898 after calibration. The largest mean empty-set fraction of the proposed rule is 0.002, and requiring an accepted center increases the mean radius by at most 1.30%. Appendices S8.5, S8.6, and S8.7 report paired comparisons, randomized deployment, and separate source and test pools.

4.3 Summary budget across resolutions

The resolution study evaluates spatially regular random functions and independent uniform pixels on grids of side 16 to 128 (Appendix S9). For regular images, the calibrated bound (8) falls below 0.35 with 16 summaries at every resolution, and the orthogonal bound of Appendix S1.11.2 needs 4. For unstructured images, the reported radius bound stays at $1/2$ until the full image is summarized, and Theorem 3 requires at least $15V/64$ linear summaries for radius 0.35 whenever the set contains an image in $[0.4, 0.6]^V$. Mean empirical coverage ranges from 0.898 to 0.909 across levels and resolutions. In the query study of Appendix S10, increasing the number of returned images reduces the mean discrepancy from the radius computed with exact predictor moments, as Appendix Theorem S2 predicts.

5 Discussion

We study the regional uncertainty of conformal image sets. Exact ranges describe every accepted image at once, cost time linear in the number of nested regions, and give an accepted reconstruction with a certified worst-case regional error. In the neural study, the certified ranges are much wider than regional intervals from risk control, which is the price of statements that hold for every region of an image at once. Adding functionals to the certified class cannot lower the calibrated threshold, and excluding single pixels reduced the largest regional range by 20% to 44%. The class should hence contain only the averages that a question requires. Standardizing by the ensemble spread gave the narrowest typical ranges. The constant scale gave the smallest radius and a certified error within a factor of two of the largest realized error, and it suits questions in which one bound must hold for every region. The exact calculation changed the reported ranges by less than 1% there, because the ensemble mean nearly matched the observed block averages, and it guarantees that the reported ranges are the smallest implied by the set. Calibration within groups keeps coverage when group proportions change and within-group distributions remain unchanged. This guarantee averages over calibration samples. For a fixed regularity class, a fixed number of regional averages controls the omitted detail at every resolution, and the calibrated threshold remains part of the radius.

The regional calculation assumes exact coarse measurements, and the experiments quantify reconstruction uncertainty given clinical status. They do not establish diagnostic performance. When source and target pools differ within groups, coverage can fall below the nominal level. In official source and test splits, the smallest mean group coverage fell to between 0.848 and 0.876 in four of the five earlier cohorts and to 0.874 for the constant scale of the neural study, whose standardized scale kept at least 0.917 in both groups (Appendices S8.7 and S7). Proposition 3 bounds the effect of a new target distribution through the density-ratio approximation error $\kappa(r, \mathcal{F})$ and the conditional-distribution error ζ.

AI USE STATEMENT

We used large language models throughout this project. They drafted and revised manuscript text, including the abstract and Introduction. The authors checked every retained statement, and take full responsibility for the content.

ETHICS STATEMENT

This study analyzes public benchmark image cohorts and does not collect new human-subject data. The reported prediction sets quantify uncertainty under prespecified deployment shifts and are not clinical decisions. The analyses are subject to the limitations of the source cohorts, including their label definitions and acquisition settings.

REPRODUCIBILITY STATEMENT

Appendix S7 gives both neural super-resolution designs, the pools, the architecture, the training, and the rule for λ. Appendix S8 gives the medical design, implementation details, and direct comparisons. Appendix S9 specifies the simulated image distributions. Appendix S1 proves the main results. Appendices S11 and S12 give the supporting proofs. The medical analyses use MedMNIST v2 (Yang et al., 2023), including PneumoniaMNIST (Kermany et al., 2018), the NIH chest X-ray benchmark through ChestMNIST (Yang et al., 2023; Wang et al., 2017), and the NLM Indiana University Chest X-ray collection through Open-i (Demner-Fushman et al., 2016). The reproducibility archive records the entry scripts, seeds, the registered designs of the neural study, checksums of the trained ensemble members, generated table and figure sources, software dependencies, and data-access instructions.

COMPETING INTERESTS

The authors report no competing interests.

REFERENCES

Eduardo Adame, Daniel Csillag, and Guilherme Tegoni Goedert. Image super-resolution with guarantees via conformalized generative models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 134523–134546. Curran Associates, Inc., 2025. doi: 10.52202/085713-4488. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/c37b07a300a2b7acb1e6d59363617a41-Abstract-Conference.html.

Anastasios N. Angelopoulos, Amit Pal Kohli, Stephen Bates, Michael I. Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 717–730, 2022.

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

Ali Baheri and Marzieh Amiri Shahbazi. Conformal prediction across scales: Finite-sample coverage with hierarchical efficiency. *Results in Applied Mathematics*, 26:100589, 2025. doi: 10.1016/j.rinam.2025.100589.

Konstantina Bairaktari, Jiayun Wu, and Steven Wu. Kandinsky conformal prediction: Beyond class- and covariate-conditional coverage. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 2581–2602, 2025. URL https://proceedings.mlr.press/v267/bairaktari25a.html.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

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

Omer Belhasin, Yaniv Romano, Daniel Freedman, Ehud Rivlin, and Michael Elad. Principal uncertainty quantification with spatial correlation for image restoration problems. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 46(5):3321–3333, 2024. doi: 10.1109/TPAMI.2023.3343031.

Jacqueline Isabel Bereska, Hamed Karimi, and Reza Samavi. SACP: Spatially-adaptive conformal prediction in uncertainty quantification of medical image segmentation. In *Proceedings of The 8th International Conference on Medical Imaging with Deep Learning*, volume 301 of *Proceedings of Machine Learning Research*, pp. 127–152. PMLR, 2026. URL https://proceedings.mlr.press/v301/bereska26a.html.

Aabesh Bhattacharyya and Rina Foygel Barber. Group-weighted conformal prediction. *Electronic Journal of Statistics*, 20(1):1171–1199, 2026. doi: 10.1214/26-EJS2506.

Joren Brunekreef, Eric Marcus, Ray Sheombarsing, Jan-Jakob Sonke, and Jonas Teuwen. Kandinsky conformal prediction: Efficient calibration of image segmentation algorithms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4135–4143, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Brunekreef_Kandinsky_Conformal_Prediction_Efficient_Calibration_of_Image_Segmentation_Algorithms_CVPR_2024_paper.html.

Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C. Duchi. Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association*, 119 (548):3033–3044, 2024. doi: 10.1080/01621459.2023.2298037.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021.

Matt Y. Cheung, Tucker J. Netherton, Laurence E. Court, Ashok Veeraraghavan, and Guha Balakrishnan. Metric-guided conformal bounds for probabilistic image reconstruction. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, volume 16166 of *Lecture Notes in Computer Science*, pp. 169–179. Springer, 2026a. doi: 10.1007/978-3-032-06593-3_16.

Matt Y. Cheung, Tucker J. Netherton, Laurence E. Court, Ashok Veeraraghavan, and Guha Balakrishnan. Bias-aware conformal prediction for metric-based imaging pipelines. In *2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI)*, pp. 01–05. IEEE, 2026b. doi: 10.1109/ISBI61048.2026.11515847. URL https://doi.org/10.1109/ISBI61048.2026.11515847.

Matt Y. Cheung, Ashok Veeraraghavan, and Guha Balakrishnan. COMPASS: Robust feature conformal prediction for medical segmentation metrics. In *International Conference on Learning Representations*, 2026c. URL https://openreview.net/forum?id=uBy4TCgGiT.

Samuel Davenport. Conformal confidence sets for biomedical image segmentation. *arXiv preprint arXiv:2410.03406*, 2024. URL https://arxiv.org/abs/2410.03406.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. doi: 10.1093/jamia/ocv080.

Guneet S. Dhillon, George Deligiannidis, and Tom Rainforth. On the expected size of conformal prediction sets. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1549–1557. PMLR, 2024. URL https://proceedings.mlr.press/v238/dhillon24a.html.

Jacopo Diquigiovanni, Matteo Fontana, and Simone Vantini. Conformal prediction bands for multivariate functional data. *Journal of Multivariate Analysis*, 189:104879, 2022.

Clara Fannjiang, Stephen Bates, Anastasios N. Angelopoulos, Jennifer Listgarten, and Michael I. Jordan. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, 119(43):e2204569119, 2022.

Chao Gao, Liren Shan, Vaidehi Srinivas, and Aravindan Vijayaraghavan. Volume optimality in conformal prediction with structured prediction sets. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18495–18527. PMLR, 2025. URL https://proceedings.mlr.press/v267/gao25k.html.

Isaac Gibbs and Emmanuel J. Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

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

Ander Gray, Vignesh Gopakumar, Sylvain Rousseau, and Sebastien Destercke. Guaranteed prediction sets for functional surrogate models. In *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 1569–1585, 2025. URL https://proceedings.mlr.press/v286/gray25a.html.

Leying Guan. Localized conformal prediction: a generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Eliahu Horwitz and Yedid Hoshen. Conffusion: Confidence intervals for diffusion models. *arXiv preprint arXiv:2211.09795*, 2022. URL https://arxiv.org/abs/2211.09795.

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

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C. S. Valentim, Huiying Liang, Sally L. Baxter, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, 2018. doi: 10.1016/j.cell.2018.02.010.

Gilad Kutiel, Regev Cohen, Michael Elad, Daniel Freedman, and Ehud Rivlin. Conformal prediction masks: Visualizing uncertainty in medical imaging. In *Trustworthy Machine Learning for Healthcare*, volume 13932 of *Lecture Notes in Computer Science*, pp. 163–176. Springer Nature Switzerland, 2023. doi: 10.1007/978-3-031-39539-0_14.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6402–6413, 2017.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Lihua Lei and Emmanuel J. Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):911–938, 2021.

Ruhan Liu, Xiangning Wang, Qiang Wu, Ling Dai, Xi Fang, Tao Yan, Jaemin Son, Shiqi Tang, Jiang Li, Zijian Gao, Adrian Galdran, J. M. Poorneshwaran, Hao Liu, Jie Wang, Yerui Chen, Prasanna Porwal, Gavin Siew Wei Tan, Xiaokang Yang, Chao Dai, Haitao Song, Mingang Chen, Huating Li, Weiping Jia, Dinggang Shen, Bin Sheng, and Ping Zhang. DeepDRiD: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns*, 3(6):100512, 2022.

Kerri Lu, Dan M. Kluger, Stephen Bates, and Sherrie Wang. Conformal prediction sets for instance segmentation. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 3999–4026, 2026. URL https://proceedings.mlr.press/v337/lu26b.html.

David Millard, Lars Lindemann, and Ali Baheri. Split conformal prediction in the function space via neural operator learning. In *ICLR 2026 Workshop on AI and Partial Differential Equations*, 2026. URL https://openreview.net/forum?id=0twOHJg60V.

Yinjie Min, Liuhua Peng, and Changliang Zou. A unified theory of conditional coverage in conformal prediction with applications. *arXiv preprint arXiv:2605.11602*, 2026. URL https://arxiv.org/abs/2605.11602.

Luca Mossina and Corentin Friedrich. Conformal prediction for image segmentation using morphological prediction sets. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15963 of *Lecture Notes in Computer Science*, pp. 78–88. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-04965-0_8.

Elias Nehme, Omer Yair, and Tomer Michaeli. Uncertainty quantification via neural posterior principal components. In *Advances in Neural Information Processing Systems*, volume 36, pp. 37128–37141, 2023. doi: 10.52202/075280-1613.

Allan Pinkus. *n-Widths in Approximation Theory*. Springer, 1985. doi: 10.1007/978-3-642-69894-1.

Hongxiang Qiu, Edgar Dobriban, and Eric Tchetgen Tchetgen. Prediction sets adaptive to unknown covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5): 1680–1705, 2023. doi: 10.1093/jrsssb/qkad069.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI*

2015, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Swami Sankaranarayanan, Anastasios Angelopoulos, Stephen Bates, Yaniv Romano, and Phillip Isola. Semantic uncertainty intervals for disentangled latent spaces. In *Advances in Neural Information Processing Systems*, volume 35, pp. 6250–6263, 2022. doi: 10.52202/068431-0452.

Matteo Sesia and Yaniv Romano. Conformal prediction using conditional histograms. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6304–6315, 2021.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Hooman Shahrokhi, Devjeet Raj Roy, Yan Yan, Venera Arnaoudova, and Jana Doppa. Conformal prediction sets for deep generative models via reduction to conformal regression. In *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 3718–3748, 2025. URL https://proceedings.mlr.press/v286/shahrokhi25a.html.

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

Julio Silva-Rodríguez and Ender Konukoglu. Trustworthy endoscopic super-resolution. *arXiv preprint arXiv:2604.18001*, 2026. URL https://arxiv.org/abs/2604.18001.

Shanshan Song, Tong Wang, Guohao Shen, Yuanyuan Lin, and Jian Huang. Wasserstein generative regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(1): 330–351, 2026. doi: 10.1093/jrsssb/qkaf053.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

Jacopo Teneggi, Matthew Tivnan, Web Stayman, and Jeremias Sulam. How to trust your diffusion model: A convex optimization approach to conformal risk control. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33940–33960. PMLR, 2023. URL https://proceedings.mlr.press/v202/teneggi23a.html.

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

Bruno Viti, Elias Karabelas, and Martin Holler. CONSIGN: Conformal segmentation informed by spatial groupings via decomposition. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10007344.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Machine Learning*, 92(2–3): 349–376, 2013.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3462–3471, 2017. doi: 10.1109/CVPR.2017.369.

Jeffrey Wen, Rizwan Ahmad, and Philip Schniter. Task-driven uncertainty quantification in inverse problems via conformal prediction. In *Computer Vision – ECCV 2024*, volume 15118 of *Lecture Notes in Computer Science*, pp. 182–199. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-73027-6_11.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2 – a large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Yachong Yang, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Doubly robust calibration of prediction sets under covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(4):943–965, 2024. doi: 10.1093/jrsssb/qkae009.

Daiwei Zhang, Lexin Li, Chandra Sripada, and Jian Kang. Image response regression via deep neural networks. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5): 1589–1614, 2023. doi: 10.1093/jrsssb/qkad073.

Xingyu Zhou, Yuling Jiao, Jin Liu, and Jian Huang. A deep generative approach to conditional sampling. *Journal of the American Statistical Association*, 118(543):1837–1848, 2023. doi:

10.1080/01621459.2021.2016424.

Supplementary Material for

Spatial Radius and Group Coverage of Conformal Image Sets under Covariate Shift

The first appendix proves the group-calibration guarantee, the exact regional-radius result, and the normalized image-radius result used in the main manuscript. The subsequent appendices give the complete formulation, supporting results and extensions, numerical analyses, and remaining proofs. Appendix S8 specifies the medical reconstruction experiments, and Appendix S9 gives the controlled image distributions. Table 3 summarizes notation used in the supporting theory. The supporting theory calls the queried predictor the teacher, and teacher samples are the returned images of the main manuscript. Appendix S7 gives the neural super-resolution study.

S1 Proofs for the deployable procedure and normalized image radius

The implemented calibration rule maintains coverage for observed groups without randomizing the deployed threshold. The first proof gives its finite-sample guarantee. The remaining results quantify conditional-distribution changes, compare calibrated predictors, and show how spatial approximation controls the normalized image radius.

S1.1 Group calibration and prevalence changes

We fix $G : \mathcal{X} \to \{1, \dots, L\}$ before calibration. Conditional on the training sample and any common predictor state, let F_g be the source score distribution within group g, including independently generated query outputs. Let n_g be the number of calibration observations in that group. The convention $R_{g,(n_g+1)} = +\infty$ includes groups with insufficient calibration data and does not replace their threshold by a marginal threshold.

Proof of Theorem 1. We condition on the training sample, the common predictor state, and calibration group labels $G_i = G(X_i)$. Within group g, the n_g calibration scores and the test score are exchangeable. We define $k_g = \lceil (n_g + 1)(1 - \alpha) \rceil$. Random ordering of ties gives

$$\mathbb{P}\{R_* \le R_{g,(k_g)} \mid G(X_*) = g, \mathcal{D}_{\mathrm{tr}}, (G_i)_i\} \ge \frac{k_g}{n_g + 1} \ge 1 - \alpha.$$

The inequality also holds when $k_g = n_g + 1$, because the threshold is infinite. Integrating over the calibration labels proves the group guarantee. For $f(x) = a_{G(x)}$ with $a_g \ge 0$ and $\sum_g a_g P_{\mathrm{src}}(G = g) = 1$,

$$P_f\{Y \in \widehat{\mathcal{C}}_G(X) \mid \mathcal{D}_{\mathrm{tr}}\} = \sum_g a_g P_{\mathrm{src}}(G = g) P_{\mathrm{src}}\{Y \in \widehat{\mathcal{C}}_G(X) \mid G = g, \mathcal{D}_{\mathrm{tr}}\} \ge 1 - \alpha.$$

This completes the proof. □

The proof uses exchangeability within each observed group, not an accurate image predictor. In a finite benchmark, the same rank argument applies when calibration and test observations are selected uniformly without replacement from one group-specific pool after training observations are removed. Separate institutional or official test pools need not satisfy that condition.

The next proof controls the random group counts and does not assume that the realized groups are large.

Proof of Proposition 1. The multiplicative binomial bound gives

$$\mathbb{P}\{\min_g n_g < n_{\mathrm{cal}} \pi_0 / 2\} \le L \exp(-n_{\mathrm{cal}} \pi_0 / 8) \le \delta / 2.$$

Conditional on group labels satisfying $\min_g n_g \geq n_{\mathrm{cal}}\pi_0/2$, the empirical score distribution $\widehat{F}_g$ obeys the Dvoretzky–Kiefer–Wolfowitz inequality with Massart's constant (Massart, 1990). Hence, except on a set of probability $\delta/2$,

$$\max_g \sup_t |\widehat{F}_g(t) - F_g(t)| \leq \sqrt{\frac{\log(4L/\delta)}{2\min_g n_g}} \leq \sqrt{\frac{\log(4L/\delta)}{n_{\mathrm{cal}}\pi_0}}.$$

Both bounds hold with probability at least $1-\delta$ by a union bound. At every finite group threshold, $\widehat{F}_g(R_{g,(k_g)}) \geq k_g/n_g \geq 1-\alpha$. An infinite threshold has coverage one.

Under conditional i.i.d. sampling alone, the same Dvoretzky–Kiefer–Wolfowitz inequality gives a simultaneous group-wise coverage shortfall bound $\{\log(2L/\delta)/(2n_g)\}^{1/2}$ for every group with $n_g > 0$, with probability at least $1-\delta$. Conditional on the group labels, each failure probability is at most δ/L. A union bound and the preceding quantile inequality give this conclusion. A group of positive source probability with $n_g = 0$ has threshold $+\infty$ and coverage one. This completes the proof. □

S1.2 Generic shift-class calibration

The generic-class guarantee uses a uniform bound over thresholds, because the selected threshold depends on calibration scores. Bounds on the weights and on the number of functions needed to approximate the shift class give this uniform control.

Proof of Theorem 2. Theorem S14 gives, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}} \sup_q \left| P_f\{R_n(X,Y) > q\} - \alpha - \frac{1}{n_{\mathrm{cal}}} \sum_i f(X_i)\{\mathbf{1}\{R_i > q\} - \alpha\} \right| \leq \psi_n(\delta, \mathcal{F}).$$

Proposition S7 shows that the infimum in (5) is attained. Evaluating the preceding inequality at $\widehat{q}$, taking the supremum over f, and using $\Gamma_n^+(\widehat{q};\mathcal{F}) \leq \tau_n$ proves (7). This completes the proof. □

S1.3 Proof under conditional-distribution changes

Proposition 3 uses the density ratio r, its approximation error $\kappa(r,\mathcal{F})$, and the conditional-distribution error ζ defined before the proposition. Its proof compares the target conditional distribution with the source conditional distribution before approximating the target covariate ratio by $\mathcal{F}$.

Proof of Proposition 3. If C is data-dependent, we condition on the data and randomization used to construct it. We define $p_C(x) = P_{\mathrm{src},Y|X=x}\{Y \in C(x)\}$. For every $f \in \mathcal{F}$, both r and f integrate to one, and hence

$$\begin{aligned} P_{\mathrm{tar}}\{Y \in C(X)\} &\geq P_f\{Y \in C(X)\} + \mathbb{E}_{P_{\mathrm{src},X}}[(r(X) - f(X))\{p_C(X) - 1/2\}] - \zeta \\ &\geq P_f\{Y \in C(X)\} - \frac{1}{2}\mathbb{E}_{P_{\mathrm{src},X}}|r(X) - f(X)| - \zeta \\ &\geq 1 - \alpha - \Delta_{\mathcal{F}}^-(C) - \frac{1}{2}\mathbb{E}_{P_{\mathrm{src},X}}|r(X) - f(X)| - \zeta. \end{aligned}$$

Taking the infimum over f proves the claim. Setting $\zeta = 0$ also proves Theorem 4. This completes the proof. □

S1.4 Proof of the exact regional-radius proposition

The proof identifies the feasible interval of each fine-region mean under the block-average equalities and verifies that the midpoint construction attains the radius in Proposition 2.

Proof of Proposition 2. We denote the mean of a candidate image in fine region j by t_j. Within a coarse block, $l_j \leq t_j \leq u_j$ and $\sum_{j\in\mathcal{B}_b} t_j = 4z_b$ are feasible exactly when $\sum l_j \leq 4z_b \leq \sum u_j$. The other three intervals attain every intermediate sum, giving coordinate range $[L_j, U_j]$. Every feasible vector is realized by an image constant on each fine region.

The piecewise-constant center with $t_j^0 = (L_j + U_j)/2$ belongs to $\mathcal{Y} = [0,1]^V$. It need not satisfy the coarse observation. Its fine-mean errors are at most $\max_j(U_j - L_j)/2$, as are its coarser-mean errors by averaging. Two feasible images attain the widest interval's endpoints L_j and U_j, giving the matching diameter lower bound. The L_2 score set is contained in this coordinate-constrained set and satisfies the same upper bound. This completes the proof. □

The L_2 comparison requires a separate nonemptiness check. Let $E_0(x)$ be the sum of squared standardized residuals for the 20 observed coarser functionals. In coarse block b, the smallest contribution from the fine regions is attained at

$$t_j(\nu_b) = \min\{1, \max\{0, m_j + \nu_b a_j^2\}\}, \qquad \sum_{j \in \mathcal{B}_b} t_j(\nu_b) = 4z_b.$$

The multiplier ν_b is found by monotone bisection. The convex quadratic minimization conditions give this formula, including coordinates at zero or one. Every feasible vector of means corresponds to an image, hence the L_2 set is nonempty exactly when

$$E_0(x) + \sum_b \sum_{j \in \mathcal{B}_b} \{t_j(\nu_b) - m_j\}^2 / a_j^2 \le q^2, \qquad q \ge 0.$$

The numerical implementation uses this criterion for the reported empty-set frequency. It does not treat feasibility of the enclosing coordinate intervals as sufficient for the L_2 set.

S1.5 An optimal center satisfying the observation

The center in (11) need not match the observed coarse image. An image that must also satisfy the observation and the score constraints requires a different minimization. For nonempty $C_Z(x)$, we define

$$r_Z(x) = \inf_{y_0 \in C_Z(x)} \sup_{y \in C_Z(x)} \sup_{h \in \mathcal{H}} |h(y) - h(y_0)|.$$

The next proposition bounds the additional error from requiring an accepted center.

Proposition S1. *If $C_Z(x) \neq \emptyset$, then*

$$\mathrm{rad}_{\mathcal{H}}\{C_Z(x)\} \le r_Z(x) \le \frac{3}{2}\mathrm{rad}_{\mathcal{H}}\{C_Z(x)\}. \tag{S.1}$$

The infimum defining $r_Z(x)$ is attained by an image constant on each fine region. Both constants are sharp.

Proposition S1 converts the calibrated set into one compatible image with the smallest worst-case regional error. Appendix S1.5 gives its exact radius and a finite construction using only the regional ranges. On the coverage event, this image has error at most $r_Z(x)$ in every regional average. An empty set is reported as such. No accepted image is asserted in that case. Equation (S.1) does not change the radius reported in Table 2, whose center is unrestricted within $[0,1]^V$.

For a nonempty maximum-score set, we define $w_j = U_j - L_j$. In coarse block b, we define $D_b = 4z_b - \sum_{j \in \mathcal{B}_b} L_j$ and $T_b = \sum_{j \in \mathcal{B}_b} w_j$. We order the four widths as $w_{b,(1)} \ge \cdots \ge w_{b,(4)}$ and define $W_{b,k} = \sum_{i=1}^{k} w_{b,(i)}$. We define

$$\begin{aligned} r_b &= \max\left\{ \frac{w_{b,(1)}}{2}, \max_{1 \le k \le 4} \frac{W_{b,k} - \min\{D_b, T_b - D_b\}}{k} \right\}, \\ r_* &= \max_{1 \le b \le 16} r_b. \end{aligned} \tag{S.2}$$

Only four widths need to be ordered per block. No optimization over individual pixels is required.

Proof of Proposition S1. The exact coordinate ranges from Proposition 2 give the same block polytope as the original constraints:

$$\left\{ t : l_j \le t_j \le u_j, \sum_j t_j = 4z_b \right\} = \left\{ t : L_j \le t_j \le U_j, \sum_j t_j = 4z_b \right\}.$$

Here and below, sums in a block are over $\mathcal{B}_b$. One inclusion follows from the coordinate extrema. The other follows from $L_j \geq l_j$ and $U_j \leq u_j$.

For a proposed center with means c_j, the largest fine-region error is

$$\max_j \max\{U_j - c_j, c_j - L_j\}.$$

Every coordinate extreme is attained by a feasible image. The coarse averages of any accepted center equal those of every accepted image, and the coarser functionals add no error. A radius at most r is hence attainable in block b precisely when

$$\underline{c}_j(r) := \max\{L_j, U_j - r\} \leq c_j \leq \overline{c}_j(r) := \min\{U_j, L_j + r\}, \qquad \sum_j c_j = 4z_b.$$

The coordinate intervals are nonempty if and only if $r \geq \max_j w_j/2$. Their lower and upper sums are

$$\sum_j \underline{c}_j(r) = \sum_j L_j + \sum_j (w_j - r)_+,$$
$$\sum_j \overline{c}_j(r) = \sum_j U_j - \sum_j (w_j - r)_+.$$

Thus their sums bracket $4z_b$ if and only if

$$\sum_j (w_j - r)_+ \leq \min\{D_b, T_b - D_b\}.$$

For decreasing widths, the identity

$$\sum_j (w_j - r)_+ = \max_{0 \leq k \leq 4} \{W_{b,k} - kr\}, \qquad W_{b,0} = 0,$$

shows that the smallest feasible r is r_b in (S.2). Nonempty blocks can be filled independently. Hence $r_Z(x) = r_*$.

A center attaining this value can be constructed without a limiting argument. In each block, we define $c_j = \underline{c}_j(r_*)$ and $\eta_b = 4z_b - \sum_j c_j$. In a fixed coordinate order, we add $\min\{\eta_b, \overline{c}_j(r_*) - c_j\}$ to c_j and subtract the same amount from η_b. The bracketing inequalities imply that the final remainder is zero. The image with constant intensity c_j on fine region j is in $[0,1]^V$, has the observed coarse averages, and satisfies every score constraint. This proves attainment and the exact formula, including zero-width intervals. All operations are measurable functions of the original ranges. Sorting ties do not affect the ordered sums.

To bound the extra radius, we define $w_b = \max_{j \in \mathcal{B}_b} w_j$. Attainment of U_j with all other coordinates at least their lower extrema gives $D_b \geq w_j$. Attainment of L_j gives $T_b - D_b \geq w_j$. Thus

$$\min\{D_b, T_b - D_b\} \geq w_b, \qquad \frac{W_{b,k} - \min\{D_b, T_b - D_b\}}{k} \leq \frac{k-1}{k} w_b \leq \frac{3}{4} w_b.$$

It follows that $w_b/2 \leq r_b \leq 3w_b/4$. Taking the largest value over blocks and using (11) proves (S.1).

For sharpness, we let $L_j = 0$ and $U_j = 1$ in every block. With $4z_b = 1$, the block is the simplex $\{t \in [0,1]^4 : \sum_j t_j = 1\}$, which gives $D_b = 1$, $T_b = 4$, $r_b = 3/4$, and the unrestricted radius is $1/2$. With $4z_b = 2$, both radii equal $1/2$. These are maximum-score sets. We use deterministic predictor images equal to $1/2$, regularized scales λ, and threshold $1/(2\lambda)$ for every regional functional. The coarser constraints hold for the stated observations. Thus both constants are attained within the stated image model. This completes the proof. □

The same calculation applies to $m \geq 2$ equal-sized fine regions per coarse block. Replacing four by m in (S.2) yields

$$\frac{1}{2} \max_j w_j \leq r_* \leq \frac{m-1}{m} \max_j w_j.$$

Indeed, the block-sum feasibility argument is unchanged and $(k-1)/k \leq (m-1)/m$. The simplex with sum one attains the upper constant. Thus the additional regional error from an accepted center is exactly determined by the regional ranges and observed sums, and does not depend on the pixel resolution.

S1.6 EXACT RANGES FOR NESTED REGIONAL PARTITIONS

Exact range calculation does not require a complete observation at one fixed dyadic level. Nested regions can have unequal sizes, and observed averages can be specified at different levels. The dependence between the regional means is then determined by the partition tree.

We consider a finite rooted tree $\mathcal{T}$ of nonempty pixel regions. The children of an internal region v partition that region, and the leaves partition the image. We denote its children by $\mathrm{ch}(v)$, the leaves by $\mathcal{L}$, and its average intensity by $h_v(y)$. For a child u of v, we define $\omega_{vu} = |u|/|v|$. Thus $\omega_{vu} > 0$, $\sum_{u\in\mathrm{ch}(v)} \omega_{vu} = 1$, and $h_v(y) = \sum_{u\in\mathrm{ch}(v)} \omega_{vu} h_u(y)$. The score and observation give intervals $[l_v, u_v] \subseteq [0,1]$. An exact observation z_v is represented by $l_v = u_v = z_v$. Intersecting all restrictions at the same region gives

$$C = \{y \in [0,1]^V : l_v \le h_v(y) \le u_v \text{ for every } v \in \mathcal{T}\}, \qquad \mathcal{H} = \{h_v : v \in \mathcal{T}\}.$$

An interval with lower bound greater than its upper bound makes C empty.

The first recursion processes children before their parent. For a leaf, we define $a_v = l_v$ and $b_v = u_v$. At an internal node, we define

$$\begin{aligned} a_v &= \max\left\{ l_v, \sum_{u\in\mathrm{ch}(v)} \omega_{vu} a_u \right\}, \\ b_v &= \min\left\{ u_v, \sum_{u\in\mathrm{ch}(v)} \omega_{vu} b_u \right\}. \end{aligned} \tag{S.3}$$

If any $a_v > b_v$, the set is empty. If every interval is nonempty, we define $L_o = a_o$ and $U_o = b_o$ at the root o. We process each parent before its children and define

$$\begin{aligned} L_u &= \max\left\{ a_u, \frac{L_v - \sum_{w\in\mathrm{ch}(v)\setminus\{u\}} \omega_{vw} b_w}{\omega_{vu}} \right\}, \\ U_u &= \min\left\{ b_u, \frac{U_v - \sum_{w\in\mathrm{ch}(v)\setminus\{u\}} \omega_{vw} a_w}{\omega_{vu}} \right\}. \end{aligned} \tag{S.4}$$

The next theorem shows that these are ranges over the whole accepted image set, not bounds obtained by treating regional intervals as independent.

Theorem S1. *Recursions (S.3)–(S.4) detect whether C is empty. If it is nonempty, then*

$$L_v = \min_{y\in C} h_v(y), \quad U_v = \max_{y\in C} h_v(y) \quad (v \in \mathcal{T}), \qquad \mathrm{rad}_{\mathcal{H}}(C) = \frac{1}{2} \max_{v\in\mathcal{L}} (U_v - L_v).$$

After the original intervals are computed, the recursions use $O(|\mathcal{T}|)$ arithmetic operations and comparisons and $O(|\mathcal{T}|)$ storage.

Theorem S1 applies to irregular nested partitions and partial multiresolution observations without changing the calibration rule. Non-nested overlapping regions are outside this tree construction. A center represented by its leaf intensities has the same storage bound. Writing all of its pixels requires an additional $O(V)$ operations.

Proof of Theorem S1. Within a subtree rooted at v, we consider all images satisfying the interval restrictions in that subtree. By induction, the attainable mean at each child u ranges over $[a_u, b_u]$, with every value attainable. The child regions are disjoint, and their images can be chosen independently. Their weighted average hence ranges over the full interval

$$\left[\sum_u \omega_{vu} a_u, \sum_u \omega_{vu} b_u \right].$$

Intersecting this interval with $[l_v, u_v]$ gives (S.3). An empty subtree excludes every image. If all the computed intervals are nonempty, the root has an attainable value and recursively choosing child means realizes a full image in C. This proves the emptiness criterion and the root ranges.

Suppose now that $[L_v, U_v]$ is the global attainable range at a parent. Let a child's value be $t \in [a_u, b_u]$. It is compatible with the parent exactly when

$$[\omega_{vu}t + \sum_{w \neq u} \omega_{vw}a_w,\ \omega_{vu}t + \sum_{w \neq u} \omega_{vw}b_w] \cap [L_v, U_v] \neq \emptyset.$$

Solving the two inequalities for t gives (S.4). Each compatible parent value has a realization outside this child's subtree, and each chosen child mean has a realization inside that subtree. The other child means fill the required weighted sum. Thus every value in $[L_u, U_u]$ is globally attainable. Induction from the root proves all claimed extrema.

We define the center intensity on leaf v as $(L_v + U_v)/2$. This image belongs to $[0, 1]^V$. Its error in every leaf mean is at most $\max_{v \in \mathcal{L}}(U_v - L_v)/2$. Every other regional mean is a convex combination of leaf means, and it has the same error bound. For a leaf of maximum width, two accepted images attain its lower and upper extremes. For every center, at least one of those images has error at least half this width in that functional. This proves the radius equality.

Each child contributes once to each weighted sum in (S.3). We store the two total sums at every parent. The sums excluding a child in (S.4) are then obtained by subtracting its single term, and each edge requires a fixed number of arithmetic operations and comparisons. A finite tree has $|\mathcal{T}| - 1$ edges. Storing the intervals and the parent totals proves the stated bounds. These counts do not include predictor queries or the initial computation of the score intervals. This completes the proof. □

If C is nonempty, an accepted image can also be computed to any prescribed radius accuracy. Let $D = \max_{v \in \mathcal{L}}(U_v - L_v)$ and let r_C be the minimum regional error when the center is restricted to C. At a trial radius $r \geq D/2$, we replace the leaf intervals by

$$[\max\{L_v, U_v - r\},\ \min\{U_v, L_v + r\}], \qquad v \in \mathcal{L},$$

keep the original internal-region intervals, and apply (S.3). This test is feasible if and only if an accepted center has worst-case regional error at most r. Indeed, the leaf inequalities are necessary because each coordinate extreme is attained. They are sufficient because all other region errors are weighted averages of leaf errors.

For $D = 0$, any feasible center is optimal. For $D > 0$, $D/2 \leq r_C \leq D$. We test $D/2$ first. If it is infeasible, we bisect the interval $[D/2, D]$, keeping its upper end feasible, until its length is at most a specified $\epsilon > 0$. The returned upper value is at most $r_C + \epsilon$. A feasible center at that value is constructed by choosing any attainable root mean and distributing each parent's remaining weighted sum among its children within their computed intervals. Starting at the child lower bounds, we increase a child by the smaller of its available interval length and the remaining sum divided by its positive weight. The upper-sum inequality ensures that the remaining sum is zero after the final child. Repeating this at every node gives a feasible image. The procedure uses $O\{|\mathcal{T}|[1 + \log(1 + D/\epsilon)]\}$ exact-real operations and comparisons, followed by $O(V)$ operations if every pixel is written. The stopping rule itself uses only arithmetic and comparisons. This is an approximation guarantee for the optimal radius, not a floating-point error guarantee.

S1.7 Comparing two scores on the same calibration sample

Comparing two calibrated predictors does not require comparing either threshold with a population quantile. The ordering of their calibration scores gives a direct comparison, including when the score distribution has atoms. Equation (19) in the proof of Theorem 3.2 of the preprint version of Lei et al. (2018), arXiv:1604.04173v2, gives this comparison for scalar regression residuals. We give the weighted comparison and its consequences for image-set radius.

Let $S, T : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ be finite measurable scores. Let $\widehat{q}_S, \widehat{q}_T$ be the thresholds from (5) on the same calibration observations, with the same nonnegative shift weights and tolerance $\tau \geq 0$. Write $\mathcal{C}_t^S(x) = \{y \in \mathcal{Y} : S(x, y) \leq t\}$ and define $\mathcal{C}_t^T$ similarly. For the radius conclusion, suppose both families satisfy the radial Lipschitz property of Proposition S12 with a common measurable bound A, where $\sup_{f \in \mathcal{F}} \mathbb{E}_f A(X) \leq A_{\mathcal{F}} < \infty$, and their expected radii are finite. These radius properties follow from that proposition for the stated linear score family. The following proposition compares thresholds and radii directly.

Proposition S2. *Suppose* $e \geq 0$ *and* $\sup_{x,y} |S(x,y) - T(x,y)| \leq e$. *Then*

$$|\widehat{q}_S - \widehat{q}_T| \leq e. \tag{S.5}$$

Under the stated radius conditions,

$$\begin{aligned} |\mathrm{rad}_{\mathcal{H}}\{\mathcal{C}^S_{\widehat{q}_S}(x)\} - \mathrm{rad}_{\mathcal{H}}\{\mathcal{C}^T_{\widehat{q}_T}(x)\}| &\leq 2A(x)e, \\ \sup_{f\in\mathcal{F}} \mathbb{E}_f \left|\mathrm{rad}_{\mathcal{H}}\{\mathcal{C}^S_{\widehat{q}_S}(X)\} - \mathrm{rad}_{\mathcal{H}}\{\mathcal{C}^T_{\widehat{q}_T}(X)\}\right| &\leq 2A_{\mathcal{F}}e. \end{aligned} \tag{S.6}$$

Proposition S2 bounds the radius difference directly through the score difference. No additional calibration-error term appears. It needs neither a density near the oracle quantile nor a small-error localization condition. The identical calibration sample, weights, and tolerance are essential. The proposition does not assert coverage for scores fitted on that sample.

Proof. We define Γ_S and Γ_T as the discrepancies in (S.30). Nonnegative weights and the score bound give, for $q \geq 0$,

$$\mathbf{1}\{S(X_i,Y_i) > q + e\} \leq \mathbf{1}\{T(X_i,Y_i) > q\}, \qquad \Gamma_S(q+e) \leq \Gamma_T(q).$$

The finite-score argument of Proposition S7 gives attained thresholds. Thus $\Gamma_S(\widehat{q}_T + e) \leq \tau$ and $\widehat{q}_S \leq \widehat{q}_T + e$. Interchanging S, T proves (S.5). For every x,

$$\mathcal{C}^S_{\widehat{q}_S}(x) \subseteq \mathcal{C}^T_{\widehat{q}_T+2e}(x), \qquad \mathcal{C}^T_{\widehat{q}_T}(x) \subseteq \mathcal{C}^S_{\widehat{q}_S+2e}(x).$$

Radius monotonicity and the two radial Lipschitz bounds give the first inequality in (S.6). Taking expectations with $\sup_f \mathbb{E}_f A(X) \leq A_{\mathcal{F}}$ gives the second and bounds the difference of the worst-case expected radii. This completes the proof. □

The same argument applies group by group to (4). If two vectors of group scores differ by at most e, their order statistics with the same index differ by at most e. At least k values of the second vector are at most the first vector's kth order statistic plus e, and the reverse comparison follows by interchanging the two vectors. If the corrected index equals $n_g + 1$, both thresholds are infinite and both score sets before the coarse-image restriction equal $\mathcal{Y}$. The finite-threshold radius comparison thus also applies to the implemented group rule. An observed-coarse-image intersection requires the separate argument below.

S1.8 Expected radius with new predictor queries

Finite covariate support is not needed when predictor queries are generated for each new image and set size is averaged over those queries. We define $\mathcal{Y} = [0,1]^V$. We assume $\mathcal{H}_q$ is nonempty, finite, and linear with $|h(y)| \leq B_{\mathcal{H}}$. We also assume $\mathcal{Y}_x \subseteq \mathcal{Y}$ is the known nonempty closed convex candidate class at x. The choice $\mathcal{Y}_x = \mathcal{Y}$ imposes no additional restriction. Known coarse averages define another such class. We suppose the source conditional distribution and Q^{T}_x are supported on $\mathcal{Y}_x$. We assume the uniform population mean and second-moment biases are at most $b_{\mu,n}$ and $b_{2,n}$, respectively, for $h \in \mathcal{H}_q$. These quantities refer to the source conditional moments, not fitted residuals.

We use K conditionally i.i.d. predictor images at each calibration observation and at prediction. Given the calibration covariates and the trained predictor, all calibration query randomness is independent of the calibration responses, and the batch U_i at X_i has distribution $(Q^{\mathrm{T}}_{X_i})^K$. Let U denote the prediction query batch, which is independent of the calibration observations and their query batches conditionally on X and the trained predictor, and define $\widehat{\mathcal{C}}_K(x;U)$ and $\widehat{\mathcal{C}}^\circ(x)$ for the query-based and source-moment score sets restricted to $\mathcal{Y}_x$. Their respective thresholds $\widehat{q}_K$ and $\widehat{q}^\circ$ use the same calibration observations and either the same rule (5) with a fixed tolerance $\tau_n \geq 0$ or the same group rule (4). The next theorem uses no regularity condition on the covariate distribution. Radius expectations use the measurability convention in Appendix S11.

Theorem S2. *Under the preceding query conditions, let* $\mathcal{F}$ *be pointwise measurable with* $f \geq 0$ *and* $\mathbb{E}_{\mathrm{src}} f(X) = 1$. *For* $n_{\mathrm{cal}}, K \geq 1$ *and* $0 < \delta < 1$, *with probability at least* $1 - \delta$,

$$\begin{aligned} &\sup_{f\in\mathcal{F}} \left|\mathbb{E}_f \mathbb{E}_{U|X} \mathrm{rad}_{\mathcal{H}_q}\{\widehat{\mathcal{C}}_K(X;U)\} - \mathbb{E}_f \mathrm{rad}_{\mathcal{H}_q}\{\widehat{\mathcal{C}}^\circ(X)\}\right| \\ &\quad \leq C\Big\{b_{\mu,n} + \sqrt{b_{2,n} + 2B_{\mathcal{H}} b_{\mu,n}} + \sqrt{\frac{1 + \log(n_{\mathrm{cal}}|\mathcal{H}_q|/\delta)}{K}}\Big\}, \end{aligned} \tag{S.7}$$

where C depends only on $B_{\mathcal{H}}$ and λ.

Theorem S2 replaces uniform estimation over covariate values by concentration at the calibration observations and averaging over independent prediction queries. It requires no covariate smoothness, conditional variance lower bound, or entropy bound for $\mathcal{F}$. In (S.7), the calibration data and their query batches are fixed. The radius is averaged over both the target covariates and the new prediction queries. For a coarse-image intersection, its support condition requires the returned predictor images to preserve the known coarse averages. Without this support condition, Proposition S4 gives the applicable nonempty-set comparison. Proposition S20 gives a uniform score comparison when the predictor is represented by random maps with controlled entropy.

Proof of Theorem S2. We condition on the trained predictor distribution and the source calibration observations. For one query batch at x, we define $D(x,U) = \sup_{y\in\mathcal{Y}} |R_K(x,y;U) - R_q^\circ(x,y)|$, where R_K is the empirical-moment score and R_q° is its source-moment counterpart. We define

$$b = b_{\mu,n} + \sqrt{b_{2,n} + 2B_{\mathcal{H}} b_{\mu,n}}.$$

The second-moment identity gives $\sup_{x,h} |\widetilde{\sigma}_h(x) - \sigma_h^\circ(x)| \le \sqrt{b_{2,n} + 2B_{\mathcal{H}} b_{\mu,n}}$, where $\widetilde{\sigma}_h(x)$ and $\sigma_h^\circ(x)$ are the predictor and source conditional standard deviations of $h(Y)$. Lemma S6, a union bound over $\mathcal{H}_q$, and Proposition S10 give, uniformly in x,

$$\mathbb{P}\{D(x,U) > C[b + \sqrt{\{1 + \log(|\mathcal{H}_q|) + t\}/K}] \mid x\} \le e^{-t}, \qquad t > 0.$$

Integrating this tail gives $\sup_x \mathbb{E}_{U|x} D(x,U) \le C[b + \sqrt{\{1+\log(|\mathcal{H}_q|)\}/K}]$. A union bound at the n_{cal} calibration observations gives, with conditional probability at least $1-\delta$,

$$d_{\mathrm{cal}} := \max_{i\in\mathcal{I}_{\mathrm{cal}}} D(X_i,U_i) \le C\Big[b + \sqrt{\frac{1+\log(n_{\mathrm{cal}}|\mathcal{H}_q|/\delta)}{K}}\Big].$$

This union bound does not require independence across calibration batches. Nonnegative weights and a common tolerance give $|\widehat{q}_K - \widehat{q}^\circ| \le d_{\mathrm{cal}}$ by Proposition S2. Scores are bounded by $2B_{\mathcal{H}}/\lambda$, hence generic thresholds are finite and attained. Group thresholds satisfy the same bound when finite. At index $n_g + 1$, both sets equal $\mathcal{Y}_x$ and contribute zero radius difference.

Convexity and the support condition place both mean images in $\mathcal{Y}_x$. Proposition S12's contraction argument applies within $\mathcal{Y}_x$, allowing radius centers in $\mathcal{Y}$ and taking $A = B_{\mathcal{H}} + \lambda$. For finite thresholds, the score inclusions give, at each new x, U,

$$\left|\mathrm{rad}_{\mathcal{H}_q}\{\widehat{\mathcal{C}}_K(x;U)\} - \mathrm{rad}_{\mathcal{H}_q}\{\widehat{\mathcal{C}}^\circ(x)\}\right| \le (B_{\mathcal{H}} + \lambda)\{d_{\mathrm{cal}} + D(x,U)\}.$$

We integrate over the independent query batch and then under P_f. The identity $\mathbb{E}_{\mathrm{src}} f(X) = 1$ makes the bound uniform in f, proving (S.7). Integrating the conditional failure probability over the calibration observations gives the unconditional probability bound. This completes the proof. □

S1.9 Predictor queries that preserve the observed image

The support condition in Theorem S2 can be enforced before calibration. Let $\ell \in \{0,\dots,J\}$ be the observed partition level. This gives a procedure for predictors whose returned images do not preserve the observed coarse averages. Here $z(x)$ denotes the blockwise-constant observed image and $z_b(x)$ its intensity in coarse block b. In this subsection, $\mathcal{Y}_x = \{y \in [0,1]^V : P_\ell y = z(x)\}$ is the known coarse-image class, and $P_x = P_{Y|X=x}$ is supported on $\mathcal{Y}_x$. Let Π_x be the metric projection onto $\mathcal{Y}_x$ for $\|\cdot\|_V$, with inner product $\langle u,v\rangle_V = V^{-1}u^\top v$. We project every returned image separately and then compute the same sample means, standard deviations, and conformal scores. The procedure uses the observed coarse image, not a response from the calibration or target sample.

For distributions P, Q on $[0,1]^V$, let $W_{2,V}(P,Q) = \inf\{\mathbb{E}\|U-W\|_V^2\}^{1/2}$, where the infimum is over pairs with marginals P and Q. Write $\Pi_x Q$ for the distribution of $\Pi_x U$ when $U \sim Q$. For the finite class $\mathcal{H}_q$ of linear functionals, suppose $|h(y)| \le B_{\mathcal{H}}$ on the image box and $|h(y) - h(y')| \le L_{\mathcal{H}}\|y - y'\|_V$. Put $\mu_h(P) = \mathbb{E}_P h(Y)$ and $\sigma_h(P) = \{\mathrm{Var}_P h(Y)\}^{1/2}$. The following proposition gives an explicit projection and controls the two score moments without a positive variance assumption.

Proposition S3. *For a coarse block b with pixel set I_b, $\Pi_x u$ is given by*

$$(\Pi_x u)_v = \min\{1, \max\{0, u_v + \nu_b\}\}, \quad v \in I_b, \qquad \sum_{v \in I_b} (\Pi_x u)_v = |I_b| z_b(x).$$

For every Q on $[0,1]^V$,

$$\begin{aligned} &W_{2,V}(P_x, \Pi_x Q) \le W_{2,V}(P_x, Q), \\ &\max_{h \in \mathcal{H}_q} \max \left\{ |\mu_h(P_x) - \mu_h(\Pi_x Q)|, |\sigma_h(P_x) - \sigma_h(\Pi_x Q)| \right\} \le L_{\mathcal{H}} W_{2,V}(P_x, Q). \end{aligned} \tag{S.8}$$

This block projection is the capped-simplex problem studied by Wang & Lu (2015). The scalar equation for each ν_b is monotone. Sorting the breakpoints $-u_v$ and $1 - u_v$ identifies a linear interval on which to solve it exactly; bisection is a numerical alternative. The projected image is unique even when the scalar solution is not. At $z_b = 0$ or $z_b = 1$, the whole block is fixed. Equation (S.8) is linear in the quadratic transport distance, including at zero variance. It is not a linear bound in the 1-Wasserstein distance W_1 computed with the same image norm. Proposition S27 gives the separate bound for W_1.

Assume $W_{2,V}(P_x, Q_x^{\mathrm{T}}) \le \epsilon$ for almost every source covariate x, conditionally on the fitted predictor. Apply Π_x to K conditionally i.i.d. queries at each calibration observation and at prediction, with the independence conditions of Theorem S2. Write $\widehat{\mathcal{C}}_K^{\Pi}(x; U)$ for the resulting set restricted to $\mathcal{Y}_x$. Its source-moment comparison set is $\widehat{\mathcal{C}}^{\circ}(x)$ from that theorem. Both procedures use the same calibration observations and the same rule and tolerance. The following bound averages over new predictor queries and does not require their errors to be small simultaneously at every covariate value.

Theorem S3. *Under the preceding conditions, for every pointwise measurable class $\mathcal{F}$ with $f \ge 0$ and $\mathbb{E}_{\mathrm{src}} f(X) = 1$, with probability at least $1 - \delta$,*

$$\begin{aligned} &\sup_{f \in \mathcal{F}} \left| \mathbb{E}_f \mathbb{E}_{U|X} \mathrm{rad}_{\mathcal{H}_q} \{\widehat{\mathcal{C}}_K^{\Pi}(X; U)\} - \mathbb{E}_f \mathrm{rad}_{\mathcal{H}_q} \{\widehat{\mathcal{C}}^{\circ}(X)\} \right| \\ &\qquad \le C \left\{ L_{\mathcal{H}} \epsilon + \sqrt{\frac{1 + \log(n_{\mathrm{cal}} |\mathcal{H}_q| / \delta)}{K}} \right\}, \end{aligned} \tag{S.9}$$

where C depends only on $B_{\mathcal{H}}$ and λ.

Equation (S.9) requires no covariate smoothness, entropy bound for the shift class, or lower bound on conditional variances. The predictor discrepancy ϵ is a population quantity used in the analysis; it is not used by the projection or calibration procedure. Applying the same projection at calibration and prediction also preserves the exchangeability needed by Theorem 1. This is an optional procedure. The existing medical tables concern the predictor outputs specified in Appendix S8, not this projected procedure, and do not establish a radius improvement from projection.

Proof of Proposition S3. For $0 < z_b < 1$, the continuous nondecreasing function

$$F_b(\nu) = \sum_{v \in I_b} \min\{1, \max\{0, u_v + \nu\}\}$$

has limits 0 and $|I_b|$ at the two ends of the real line. Thus a solution exists. Let $t_v = \min\{1, \max\{0, u_v + \nu_b\}\}$. Projection onto $[0,1]$ gives $(u_v + \nu_b - t_v)(y_v - t_v) \le 0$ for every $y_v \in [0,1]$. For any feasible block y, summing and using $\sum_v (y_v - t_v) = 0$ gives

$$\sum_{v \in I_b} (u_v - t_v)(y_v - t_v) \le 0.$$

This is the variational inequality for the unique minimizer of $\sum_v (y_v - u_v)^2$ subject to the box and block-sum constraints. When $z_b = 0$ or $z_b = 1$, the constraints fix every entry in the block to z_b. The formula also gives a measurable projection as a function of u and z: for interior z_b, one may select $\nu_b = \inf\{r \in \mathbb{Q} : F_b(r) \ge |I_b| z_b\}$ and use continuity; at $z_b = 0$ and $z_b = 1$ use the fixed block.

For a coupling (Y, U) of (P_x, Q), $\Pi_x Y = Y$. The projection inequality for a closed convex set implies $\|\Pi_x u - \Pi_x v\|_V \le \|u - v\|_V$. To verify it, apply the two variational inequalities at u, v, add them, and obtain

$$\|\Pi_x u - \Pi_x v\|_V^2 \le \langle \Pi_x u - \Pi_x v, u - v \rangle_V.$$

Cauchy–Schwarz proves the assertion. Hence

$$\mathbb{E}\|Y - \Pi_x U\|_V^2 \le \mathbb{E}\|Y - U\|_V^2.$$

Taking the infimum over couplings proves the first part of (S.8). Let $Z = h(Y)$ and $W = h(\Pi_x U)$. We write $\|Z\|_{L_2} = (\mathbb{E}Z^2)^{1/2}$ and $\mathrm{sd}(Z) = \mathrm{Var}(Z)^{1/2}$. Then

$$|\mathbb{E}Z - \mathbb{E}W| \le \|Z - W\|_{L_2}, \qquad |\,\mathrm{sd}(Z) - \mathrm{sd}(W)| \le \|(Z - \mathbb{E}Z) - (W - \mathbb{E}W)\|_{L_2} \le \|Z - W\|_{L_2}.$$

The Lipschitz bound on h and the preceding coupling inequality prove the moment bound after taking the infimum. This completes the proof. □

Proof of Theorem S3. Condition on the fitted predictor and the source calibration observations. For one projected query batch, write $D(x, U) = \sup_{y \in [0,1]^V} |R_K^{\Pi}(x, y; U) - R_q^{\circ}(x, y)|$, where R_K^{Π} uses the empirical projected-query moments and R_q° uses the source conditional moments. If e_μ and e_σ are the largest mean and standard-deviation errors over $\mathcal{H}_q$, respectively, the positive ridge and $|h| \le B_{\mathcal{H}}$ give

$$D(x, U) \le \lambda^{-1} e_\mu + 2B_{\mathcal{H}} \lambda^{-2} e_\sigma.$$

Proposition S3 bounds the two population biases by $L_{\mathcal{H}}\epsilon$. Lemma S6 applies to the bounded projected queries, including constant functionals. A union bound over $\mathcal{H}_q$ gives, uniformly for almost every x and for $t > 0$,

$$\mathbb{P}\left\{ D(x, U) > C\left[L_{\mathcal{H}}\epsilon + \sqrt{\{1 + \log|\mathcal{H}_q| + t\}/K} \right] \,\middle|\, x \right\} \le e^{-t}.$$

Hence, with probability at least $1 - \delta$,

$$d_{\mathrm{cal}} := \max_{i \in \mathcal{I}_{\mathrm{cal}}} D(X_i, U_i) \le C\left\{ L_{\mathcal{H}}\epsilon + \sqrt{\frac{1 + \log(n_{\mathrm{cal}}|\mathcal{H}_q|/\delta)}{K}} \right\}.$$

Integration of the same tail gives $\sup_x \mathbb{E}_{U|x} D(x, U) \le C\{L_{\mathcal{H}}\epsilon + \sqrt{(1 + \log|\mathcal{H}_q|)/K}\}$, where the supremum is essential under the source covariate distribution. Nonnegative weights and the same tolerance imply $|\widehat{q}_K^{\Pi} - \widehat{q}^{\circ}| \le d_{\mathrm{cal}}$ by Proposition S2. The order-statistic argument gives the same conclusion for finite group thresholds. When a group threshold is infinite, both candidate sets equal $\mathcal{Y}_x$.

The empirical and source mean images belong to the convex set $\mathcal{Y}_x$, and every standard deviation plus ridge is at most $B_{\mathcal{H}} + \lambda$. Proposition S12, applied within $\mathcal{Y}_x$, and the score inclusions hence give

$$\left| \mathrm{rad}_{\mathcal{H}_q}\{\widehat{\mathcal{C}}_K^{\Pi}(x; U)\} - \mathrm{rad}_{\mathcal{H}_q}\{\widehat{\mathcal{C}}^{\circ}(x)\} \right| \le (B_{\mathcal{H}} + \lambda)\{d_{\mathrm{cal}} + D(x, U)\}$$

for finite thresholds; the difference is zero for the simultaneous infinite thresholds. Integrating first over the independent prediction batch and then under P_f, and using $\mathbb{E}_{\mathrm{src}} f(X) = 1$, proves (S.9). Averaging over the conditioning observations proves the stated probability bound. Finally, the measurable, deterministic map $u \mapsto \Pi_x u$ is the same at calibration and prediction. Independent query batches passed through this map leave the within-group score vector exchangeable, and the rank proof of Theorem 1 is unchanged. This completes the proof. □

S1.10 Stability of the radius under the observed coarse image

The coarse observation can exclude the predictor center. Moving candidate images toward that center may violate the known coarse averages, but the exact interval calculation gives a direct comparison of two nonempty regional sets.

Fix x and its observed coarse averages. For $v \in \{S, T\}$, let $C_Z^v(x)$ be a maximum-score set as in Proposition 2, with fine-region centers m_j^v, positive scales a_j^v, and finite threshold $q_v \ge 0$. Define $D = \max_j\{|m_j^S - m_j^T| + |q_S a_j^S - q_T a_j^T|\}$. Also put $d_m = \max_j |m_j^S - m_j^T|$, $d_a = \max_j |a_j^S - a_j^T|$, and $a_S = \max_j a_j^S$. The following proposition applies to the same regional sets used in the medical comparisons.

Proposition S4. *If $C_Z^S(x) \neq \emptyset$ and $C_Z^T(x) \neq \emptyset$, then*

$$|\mathrm{rad}_{\mathcal{H}}\{C_Z^S(x)\} - \mathrm{rad}_{\mathcal{H}}\{C_Z^T(x)\}| \leq 3D. \tag{S.10}$$

If $|q_S - q_T| \leq e$, then $D \leq d_m + q_T d_a + a_S e$.

Proposition S4 links errors in the predictor moments and the calibrated threshold to the reported exact regional radius. It requires no positive lower bound on the conditional standard deviation. The nonemptiness condition concerns the two sets being compared, not a population assumption used to prove coverage.

Proof of Proposition S4. We use superscripts S, T for the interval endpoints in Proposition 2. Clipping at zero or one is 1-Lipschitz, which gives $|l_j^S - l_j^T| \leq D$ and $|u_j^S - u_j^T| \leq D$. Each block has four fine regions. Hence

$$|L_j^S - L_j^T| \leq 3D, \qquad |U_j^S - U_j^T| \leq 3D, \qquad |(U_j^S - L_j^S) - (U_j^T - L_j^T)| \leq 6D.$$

The identity (11) gives (S.10). Finally, $|q_S a_j^S - q_T a_j^T| \leq a_j^S |q_S - q_T| + q_T |a_j^S - a_j^T|$ gives the bound for D. This completes the proof. □

The restriction to two nonempty sets is necessary. We consider an observed coarse image with every block average equal to $1/2$, all scales equal to one, and a piecewise-constant predictor center equal to zero on one observed block and $1/2$ elsewhere. For $q < 1/2$, the zero-centered block cannot have average $1/2$. Hence the intersection is empty. At $q = 1/2$, all four fine means in that block must equal $1/2$, but in another block a fine mean can attain both zero and one while the block average remains $1/2$. The coarser score constraints are satisfied. Hence

$$\mathrm{rad}_{\mathcal{H}}\{C_Z(q)\} = 0 \quad (q < 1/2), \qquad \mathrm{rad}_{\mathcal{H}}\{C_Z(1/2)\} = 1/2.$$

Thus radius need not be continuous at the boundary of feasibility. Reporting empty-set frequency separately is necessary. A radius equal to zero alone does not demonstrate an accurate predictor.

S1.11 A NORMALIZED RADIUS AND ARBITRARY LINEAR SUMMARIES

The normalization $\|y\|_V = V^{-1/2}\|y\|_2$ measures root mean squared image intensity and is comparable across pixel resolutions. The radius rad_V uses this norm in (2). Let P_j be the orthogonal projection onto images constant on each cell of a $2^j \times 2^j$ partition. Its rank is 4^j. The response class $\mathcal{Y}_{s,R}^{(V)}$ consists of images in $[0,1]^V$ satisfying

$$\|(I - P_j)y\|_V \leq R2^{-sj}, \qquad 0 \leq j \leq J, \quad V = 4^J. \tag{S.11}$$

This is a population support restriction. It is not an empirical property assumed after observing a calibration sample.

For a predictor center $\widehat{\mu}(x) \in \mathcal{Y}_{s,R}^{(V)}$, we calibrate the score $\|P_j\{y - \widehat{\mu}(x)\}\|_V$ and define

$$\widehat{\mathcal{C}}_j(x) = \{y \in \mathcal{Y}_{s,R}^{(V)} : \|P_j\{y - \widehat{\mu}(x)\}\|_V \leq \widehat{q}_j(x)\}. \tag{S.12}$$

The threshold can be group-specific as in Theorem 1, or it can use the generic shift-class rule. Both calibrate the membership score of the set in (S.12).

Proof of Theorem 3. For every $y \in \widehat{\mathcal{C}}_j(x)$, orthogonal decomposition and (S.11) give

$$\|y - \widehat{\mu}(x)\|_V \leq \|P_j\{y - \widehat{\mu}(x)\}\|_V + \|(I - P_j)y\|_V + \|(I - P_j)\widehat{\mu}(x)\|_V \leq \widehat{q}_j(x) + 2R2^{-sj}.$$

The admissible center gives the radius upper bound. The smallest $j \leq J$ with $2R2^{-sj} \leq \epsilon$ satisfies $4^j \leq 4(2R/\epsilon)^{2/s}$. If no such $j < J$ exists, we define $j = J$, where the remainder is zero.

For the lower bound, we fix a realized linear map A with rank at most M and an accepted image $y^\dagger \in [\rho, 1-\rho]^V$. The polytope

$$K = \ker(A) \cap [-\rho, \rho]^V$$

has dimension $d = V - \operatorname{rank}(A) \geq V - M$. A vertex v has at least d independent active coordinate constraints. Fewer would permit $v + tu$ and $v - tu$ for a nonzero u and small $t > 0$. Hence $\|v\|_2^2 \geq d\rho^2$. Both $y^\dagger + v$ and $y^\dagger - v$ are in the box and accepted at summary $Ay^\dagger$. For every center y_0,

$$\max\{\|y^\dagger + v - y_0\|_V, \|y^\dagger - v - y_0\|_V\} \geq \|v\|_V \geq \rho\sqrt{1 - M/V}.$$

Taking the infimum over centers proves the lower bound, pathwise for the realized map and acceptance set. This includes data-dependent linear summaries. This completes the proof. □

The upper bound in Theorem 3 controls a prediction set through a spatial approximation error. The lower bound applies to every linear summary map, including a fitted principal-component map. It does not extend to arbitrary nonlinear maps. For nonlinear maps, the spatial differences among images with the same accepted summary still give a radius bound. That bound can equal zero and must be checked for the particular representation class.

S1.11.1 An expected lower bound from pixel marginals

Coverage and the marginal pixel distributions give a lower bound on the expected radius without requiring an image with every pixel in the interior. We fix a target distribution P_f and define $p_\rho = V^{-1}\sum_{v=1}^{V} P_f\{\rho \leq Y_v \leq 1 - \rho\}$ for $0 \leq \rho \leq 1/2$. All probabilities and expectations below include the randomness used to construct the set. The linear map and acceptance region may be data-dependent.

Theorem S4. *Suppose $0 \leq M < V$ and $P_f\{Y \in C^{(M)}(X)\} \geq 1 - \alpha$. Then*

$$\mathbb{E}_f\left[\operatorname{rad}_V\{C^{(M)}(X)\}\right] \geq \frac{1}{\sqrt{1 - M/V}} \int_0^{1/2} \left[p_\rho - M/V - \alpha(1 - M/V)\right]_+ d\rho. \tag{S.13}$$

If $Y_v \sim \operatorname{Unif}(0,1)$ for every v, then

$$\mathbb{E}_f\left[\operatorname{rad}_V\{C^{(M)}(X)\}\right] \geq \frac{(1-\alpha)^2}{4}(1 - M/V)^{3/2}. \tag{S.14}$$

Equation (S.13) requires no independence between pixels. For uniform pixel marginals and fixed $\alpha < 1$, (S.14) shows that an expected radius converging to zero requires $M/V \to 1$. This conclusion concerns the full set defined by the linear summaries. An additional spatial support restriction, such as $\mathcal{Y}_{s,R}^{(V)}$, changes that set and permits the upper bound in Theorem 3.

Proof of Theorem S4. We define $d = V - M$. For $y \in [0,1]^V$, we let $r_v = \min(y_v, 1 - y_v)$ and order these margins as $r_{(1)} \leq \cdots \leq r_{(V)}$. We fix a realized map A with rank at most M, and consider

$$K_y = \{u \in \ker(A) : |u_v| \leq r_v,\ 1 \leq v \leq V\}.$$

If N margins are positive, their coordinate subspace contains a polytope of dimension at least $(N - M)_+$. A vertex has that many independent active bounds. Fewer would permit a displacement in both signs. Including the $V - N$ zero-coordinate bounds gives at least d coordinates with $|u_v| = r_v$. For any d of these, indexed by A_0, Cauchy–Schwarz gives

$$\|u\|_2 \geq d^{-1/2} \sum_{v \in A_0} r_v \geq d^{-1/2} \sum_{j=1}^{d} r_{(j)}.$$

The box contains $y + u$ and $y - u$, both with summary Ay. If $y \in C^{(M)}(x)$, both are accepted, giving radius at least $\|u\|_2/\sqrt{V}$. We define $N_y(\rho) = \sum_{v=1}^{V} \mathbf{1}\{r_v \geq \rho\}$. Integration gives

$$\sum_{j=1}^{d} r_{(j)} = \int_0^{1/2} \{N_y(\rho) - M\}_+ \, d\rho.$$

On the coverage event,

$$\operatorname{rad}_V\{C^{(M)}(X)\} \geq \frac{\mathbf{1}\{Y \in C^{(M)}(X)\}}{\sqrt{Vd}} \int_0^{1/2} \{N_Y(\rho) - M\}_+ \, d\rho.$$

For each ρ, we define $D_\rho = \{N_Y(\rho) - M\}_+$. Because $0 \le D_\rho \le d$ and $\mathbb{E}_f N_Y(\rho) = V p_\rho$,

$$\mathbb{E}_f[D_\rho \mathbf{1}\{Y \in C^{(M)}(X)\}] \ge \left[\mathbb{E}_f D_\rho - d P_f\{Y \notin C^{(M)}(X)\}\right]_+ \ge \left[V p_\rho - M - \alpha d\right]_+.$$

Tonelli's theorem proves (S.13). Because p_ρ is nonincreasing, this bound is also at least $\rho[p_\rho - M/V - \alpha(1 - M/V)]_+ / \sqrt{1 - M/V}$ for each $\rho \in [0, 1/2]$. For uniform marginals, $p_\rho = 1 - 2\rho$. With $a = (1 - \alpha)(1 - M/V)$,

$$\int_0^{1/2} [a - 2\rho]_+ \, d\rho = \int_0^{a/2} (a - 2\rho) \, d\rho = \frac{a^2}{4},$$

which proves (S.14). This completes the proof. □

The expected lower bound also applies after fitting the summary map, provided that the candidate set includes every image in $[0, 1]^V$ with an accepted summary value. It does not give a lower bound for a decoder image or for a set restricted by an additional regularity model.

S1.11.2 A SHARPER RADIUS FROM ORTHOGONAL PROJECTION

Projecting the center removes its fine-scale variation and gives a sharper radius bound for the same prediction set. For the class and set in (S.11)–(S.12), and a finite threshold $\widehat{q}_j(x) \ge 0$,

$$\mathrm{rad}_V\{\widehat{\mathcal{C}}_j(x)\} \le \{\widehat{q}_j(x)^2 + R^2 2^{-2sj}\}^{1/2}, \qquad j < J. \tag{S.15}$$

At $j = J$, the bound is $\widehat{q}_J(x)$. The change is in the radius bound, not in the calibrated set or its coverage.

For (S.15), pixel averaging gives $P_j\widehat{\mu}(x) \in [0, 1]^V$. Nested projections give $\|(I - P_k)P_j\widehat{\mu}(x)\|_V \le \|(I - P_k)\widehat{\mu}(x)\|_V$ for $k < j$ and zero for $k \ge j$. Thus $P_j\widehat{\mu}(x) \in \mathcal{Y}^{(V)}_{s,R}$. For $y \in \widehat{\mathcal{C}}_j(x)$,

$$\begin{aligned}\|y - P_j\widehat{\mu}(x)\|_V^2 &= \|P_j\{y - \widehat{\mu}(x)\}\|_V^2 + \|(I - P_j)y\|_V^2 \\ &\le \widehat{q}_j(x)^2 + R^2 2^{-2sj}.\end{aligned}$$

Taking the supremum over candidates and then the infimum over centers proves the bound. At full resolution $I - P_J = 0$.

To attain both terms in (S.15), we fix $j < J$, define $a = R2^{-sj}$, and let $q = \widehat{q}_j(x) \ge 0$. We suppose $q + a \le 1/2$. Let $\mathbf{1}$ be the constant-one image. We let w be constant on level-$(j+1)$ blocks, with two $+1$ and two -1 children in each level-j block. Then $P_j w = 0$ and $\|w\|_V = 1$. At center $\widehat{\mu} = \mathbf{1}/2$, both $y_\pm = \mathbf{1}/2 \pm (q\mathbf{1} + aw)$ are in the box with projected score q. Their tails equal a for $k \le j$ and zero for $k \ge j + 1$, placing them in $\mathcal{Y}^{(V)}_{s,R}$. Their distance $2\sqrt{q^2 + a^2}$ gives

$$\mathrm{rad}_V\{\widehat{\mathcal{C}}_j\} = \sqrt{q^2 + R^2 2^{-2sj}}. \tag{S.16}$$

Equation (S.16) shows that the sharper bound is exact for this regular bounded-image construction. It does not assert equality for every calibrated set. Table 29 compares this bound with (8) on the same calibrated image sets.

S1.11.3 REGULAR FUNCTIONS ON PIXEL GRIDS

The next lemma verifies (S.11) for discretizations of a spatially regular function. It makes the resolution-independent constant explicit.

Lemma S1. *Suppose y consists of values of a function $g : [0, 1]^2 \to [0, 1]$ on the regular pixel-center grid, and $|g(u) - g(v)| \le L\|u - v\|_2^s$ for $0 < s \le 1$. Then*

$$\|(I - P_j)y\|_V \le L 2^{s/2} 2^{-sj}, \qquad 0 \le j \le J.$$

If $s = 1$, the sharper bound $L2^{-j}/\sqrt{6}$ holds.

Lemma S1 verifies the support restriction with a constant independent of the pixel count for discretized Hölder functions. The stronger statement requires spatial regularity. It does not follow from bounded pixel intensities alone.

Proof. Within a cell of side 2^{-j}, pairwise distances are at most $\sqrt{2}\,2^{-j}$, and each value differs from its average by at most $L2^{s/2}2^{-sj}$. Squaring and averaging gives the first bound. For $s = 1$, let U, U' be independent uniform grid points in the cell. Then

$$\mathrm{Var}\{g(U)\} = \tfrac{1}{2}\mathbb{E}\{g(U) - g(U')\}^2 \le \tfrac{L^2}{2}\mathbb{E}\|U - U'\|_2^2 \le L^2 2^{-2j}/6.$$

Averaging over cells proves the second bound. This completes the proof. □

S1.12 Linear summaries within the regular image class

The preceding upper and lower bounds concern different candidate classes. We now compare linear reductions within $\mathcal{Y}_{s,R}^{(V)}$ and keep the observed coarse image fixed. Represent the observed level-ℓ coarse averages by their blockwise-constant image $z(X)$, which implies that $P_\ell Y = z(X)$. Write $\mathcal{Y}_{s,R}^{(V)}(x) = \{y \in \mathcal{Y}_{s,R}^{(V)} : P_\ell y = z(x)\}$ and assume it is nonempty. For $\ell \le j \le J$, let $\widehat{\mu}(x) \in \mathbb{R}^V$ be arbitrary, let $q \ge 0$, and define

$$C_{j,z}(x) = \{y \in \mathcal{Y}_{s,R}^{(V)}(x) : \|P_j\{y - \widehat{\mu}(x)\}\|_V \le q\}, \qquad D_j(x) = \mathcal{Y}_{s,R}^{(V)}(x) \cap \mathrm{range}(P_j).$$

The distance $\mathrm{dist}_V(u, D) = \inf_{v \in D}\|u - v\|_V$ measures whether the fitted center is compatible with the regular candidate class and the observation. The next bound does not assume that the predictor mean itself belongs to that class.

Proposition S5. *If $C_{j,z}(x) \neq \emptyset$ and $q < \infty$, then*

$$\mathrm{rad}_V\{C_{j,z}(x)\}^2 \le q^2 - \mathrm{dist}_V\{P_j\widehat{\mu}(x), D_j(x)\}^2 + R^2 2^{-2sj}\mathbf{1}\{j < J\}. \tag{S.17}$$

In (S.17), the calibrated uncertainty and the omitted detail contribute in orthogonal directions. A center outside the feasible class uses part of the threshold to reach that class; it does not require an additional radius term. This result extends the orthogonal bound in Appendix S1.11.2 to an arbitrary fitted center and an observed coarse image. It changes neither the accepted set nor its coverage event.

For the converse, let $A_x : \mathbb{R}^V \to \mathbb{R}^M$ and $\mathcal{B}(x) \subseteq \mathbb{R}^M$ be fitted from the available data, and put $C_{A,z}(x) = \{y \in \mathcal{Y}_{s,R}^{(V)}(x) : A_x y \in \mathcal{B}(x)\}$. Thus every candidate with the same accepted linear measurements and known coarse averages is included. The rank bound below counts the calibrated measurements; the fixed coarse observation is available separately. The population support condition leaves both a pixel margin ρ and half the regularity allowance unused. Put $k = \lceil \log_4(2M) \rceil$ and $d = 4^k$.

Theorem S5. *Suppose $1 \le M \le V/8$, $4^\ell \le 2M$, $0 < \rho \le 1/2$, and $\mathrm{rank}(A_x) \le M$. If*

$$\mathbb{P}\{Y \in \mathcal{Y}_{s,R/2}^{(V)} \cap [\rho, 1-\rho]^V,\ P_\ell Y = z(X)\} = 1, \qquad \mathbb{P}\{Y \in C_{A,z}(X)\} \ge 1 - \alpha,$$

then

$$\mathbb{E}\mathrm{rad}_V\{C_{A,z}(X)\} \ge (1-\alpha)\min\{\rho, (R/2)2^{-sk}\}\sqrt{1 - M/d} \ge \frac{(1-\alpha)8^{-s/2}}{2\sqrt{2}}\min\{\rho, RM^{-s/2}\}. \tag{S.18}$$

Theorem S5 permits random, data-dependent linear measurements and does not require independent pixels. Its support condition includes nonconstant regular images with a fixed intensity margin. Proposition S5 and (S.18) give the same exponent $M^{-s/2}$ for the detail left by a linear reduction, at scales finer than the known observation and before the pixel dimension is reached. The lower bound concerns the stated full inverse-image sets, not nonlinear encoders or arbitrary prediction procedures. The upper bound also includes the calibrated threshold. Hence this comparison does not claim joint minimax optimality in calibration sample size, query count, and total radius. The approximation exponent is classical; the lower bound connects the unresolved directions to prediction-set coverage while preserving the observed image.

Proof of Proposition S5. Fix x and write $K = \mathcal{Y}_{s,R}^{(V)}(x)$ and $u = P_j\widehat{\mu}(x)$. The set K is compact and convex. For $y \in K$, block averaging preserves $[0,1]^V$ and $P_\ell P_j y = P_\ell y$. If $r \le j$, then

$$\|(I - P_r)P_j y\|_V = \|P_j(I - P_r)y\|_V \le \|(I - P_r)y\|_V \le R2^{-sr}.$$

If $r > j$, then $(I - P_r)P_j y = 0$. Hence, $P_j K \subseteq K$ and $D_j(x) = P_j K$ is a nonempty compact convex set. Let c be the metric projection of u onto $D_j(x)$ in the inner product $\langle \cdot, \cdot \rangle_V$. For $t \in D_j(x)$, the projection inequality gives

$$\langle u - c, t - c \rangle_V \le 0, \qquad \|t - c\|_V^2 \le \|t - u\|_V^2 - \|u - c\|_V^2.$$

For $y \in C_{j,z}(x)$, take $t = P_j y$. Because $c \in \mathrm{range}(P_j)$,

$$\|y - c\|_V^2 = \|P_j y - c\|_V^2 + \|(I - P_j)y\|_V^2 \le q^2 - \|u - c\|_V^2 + R^2 2^{-2sj} \mathbf{1}\{j < J\}.$$

Nonemptiness implies $\|u - c\|_V \le q$. The point c is an admissible image center, and $\|u - c\|_V = \mathrm{dist}_V\{u, D_j(x)\}$. Taking the supremum over y proves (S.17). This completes the proof. □

Proof of Theorem S5. The assumptions on M and ℓ imply $\ell \le k \le J - 1$ and $2M \le d < 8M$. In each of the d level-k blocks choose a vector w_b that equals 1 on two of its four level-$(k+1)$ children, -1 on the other two, and zero outside that block. Their supports are disjoint, their block averages equal zero, and

$$P_r w_b = 0 \quad (r \le k), \qquad P_r w_b = w_b \quad (r \ge k+1), \qquad \Big\| \sum_{b=1}^{d} u_b w_b \Big\|_V^2 = d^{-1} \sum_{b=1}^{d} u_b^2.$$

Fix a realization of the fitted procedure and a response $y \in \mathcal{Y}^{(V)}_{s,R/2} \cap [\rho, 1-\rho]^V$ with $P_\ell y = z(x)$. Put $a = \min\{\rho, (R/2)2^{-sk}\}$ and $v(u) = \sum_b u_b w_b$. For vectors, we use $\|u\|_\infty = \max_b |u_b|$. For $\|u\|_\infty \le a$, each coordinate of $v(u)$ has magnitude at most a. Thus $y \pm v(u) \in [0,1]^V$ and $P_\ell\{y \pm v(u)\} = z(x)$. For $r \le k$,

$$\|(I - P_r)\{y \pm v(u)\}\|_V \le (R/2)2^{-sr} + \|v(u)\|_V \le (R/2)2^{-sr} + a \le R2^{-sr}.$$

For $r \ge k+1$, the perturbation has no omitted detail, and the same bound follows from $y \in \mathcal{Y}^{(V)}_{s,R/2}$. The whole translated cube is hence contained in $\mathcal{Y}^{(V)}_{s,R}(x)$.

Let $T : \mathbb{R}^d \to \mathbb{R}^M$ be $Tu = A_x v(u)$ and let $L = \ker T$, of dimension $r_0 \ge d - M$. The compact polytope $L \cap [-a, a]^d$ has nonempty relative interior and a vertex u^*. At least r_0 coordinates of u^* have magnitude a. Indeed, if fewer were active, the corresponding coordinate functionals would have a nonzero common null vector $h \in L$. For sufficiently small $t > 0$, both $u^* + th$ and $u^* - th$ would lie in the polytope, contradicting that u^* is a vertex. It follows that

$$A_x v(u^*) = 0, \qquad \|v(u^*)\|_V^2 = d^{-1}\|u^*\|_2^2 \ge a^2(d - M)/d.$$

On $\{y \in C_{A,z}(x)\}$, both $y + v(u^*)$ and $y - v(u^*)$ belong to $C_{A,z}(x)$. For every image center c, the triangle inequality yields

$$2\|v(u^*)\|_V \le \|y + v(u^*) - c\|_V + \|y - v(u^*) - c\|_V.$$

Thus $\mathrm{rad}_V\{C_{A,z}(x)\} \ge a\sqrt{1 - M/d}$ on the coverage event. This is a pointwise statement and requires no measurable selection of the vertex. Taking expectations gives the first inequality in (S.18). Finally, $d < 8M$, $d \ge 2M$, and $2^{-sk} = d^{-s/2}$ imply

$$a\sqrt{1 - M/d} \ge \frac{1}{\sqrt{2}} \min\{\rho, (R/2)(8M)^{-s/2}\} \ge \frac{8^{-s/2}}{2\sqrt{2}} \min\{\rho, RM^{-s/2}\}.$$

This proves the second inequality. This completes the proof. □

The support condition in Theorem S5 is not restricted to constant images. For example, we fix a level $\ell \le r \le J - 1$, we choose one of the zero-average block vectors w from the proof at that level, and we define $Y = \frac{1}{2}\mathbf{1} + Tw$, where $|T| \le \min\{1/2 - \rho, (R/2)2^{-sr}\}$. For every coarser level $j \le r$, $\|(I - P_j)Y\|_V \le |T| \le (R/2)2^{-sj}$. For $j \ge r + 1$, the omitted detail is zero. Thus any distribution of T on this interval satisfies the stated support condition and has a fixed observed coarse image. More generally, any distribution supported on the indicated smaller regular class is allowed. The argument concerns the information lost by the accepted summary, not the quality of a particular predictor.

S1.13 THE SAME COMPARISON FOR HÖLDER IMAGES

The block-approximation class permits some images with dyadic discontinuities. We next obtain the same exponent on a Hölder class, using perturbations that equal zero at block boundaries. This separates spatial smoothness from alignment with the summary partition.

Let $0 < s \le 1$ and $L > 0$, let t_v be the pixel centers, and let $\mathcal{K}^{(V)}_{s,L}$ be the arrays $y \in [0,1]^V$ satisfying $|y_v - y_w| \le L\|t_v - t_w\|_2^s$ for every v, w. These are samples of Hölder functions. Clipping $\inf_v\{y_v + L\|t - t_v\|_2^s\}$ to $[0,1]$ gives an interpolating extension with the same constant. We define $\mathcal{K}^{(V)}_{s,L}(x) = \{y \in \mathcal{K}^{(V)}_{s,L} : P_\ell y = z(x)\}$. For a projected-score set, we define $C^{\mathrm{H}}_{j,z}(x)$ as $C_{j,z}(x)$ with this candidate class, and define $D^{\mathrm{H}}_j(x) = P_j\mathcal{K}^{(V)}_{s,L}(x)$. For a general linear reduction, we define $C^{\mathrm{H}}_{A,z}(x) = \{y \in \mathcal{K}^{(V)}_{s,L}(x) : A_x y \in \mathcal{B}(x)\}$. All distances and radii use $\|\cdot\|_V$. We define $B_s = 2^{1-s}(\sqrt{5}\pi)^s$.

Proposition S6. *For $\ell \le j \le J$, $C^{\mathrm{H}}_{j,z}(x) \ne \emptyset$, and $q < \infty$,*

$$\mathrm{rad}_V\{C^{\mathrm{H}}_{j,z}(x)\}^2 \le q^2 - \mathrm{dist}_V\{P_j\widehat{\mu}(x), D^{\mathrm{H}}_j(x)\}^2 + L^2 2^s 2^{-2sj}\mathbf{1}\{j < J\}. \tag{S.19}$$

Suppose $1 \le M \le V/8$, $4^\ell \le 2M$, $0 < \rho \le 1/2$, $\mathrm{rank}(A_x) \le M$, and

$$\mathbb{P}\{Y \in \mathcal{K}^{(V)}_{s,L/2} \cap [\rho, 1-\rho]^V,\ P_\ell Y = z(X)\} = 1, \qquad \mathbb{P}\{Y \in C^{\mathrm{H}}_{A,z}(X)\} \ge 1 - \alpha.$$

Then

$$\mathbb{E}\mathrm{rad}_V\{C^{\mathrm{H}}_{A,z}(X)\} \ge \frac{(1-\alpha)8^{-s/2}}{4\sqrt{2}B_s}\min\{\rho, LM^{-s/2}\}. \tag{S.20}$$

Equations (S.19) and (S.20) compare the same Hölder candidate class in resolution-normalized units. The known coarse image, fitted linear measurements, and population coverage all appear in the lower bound. As before, the comparison concerns the unresolved-detail exponent, not the part of the radius due to the calibrated threshold.

Proof. The class $\mathcal{K}^{(V)}_{s,L}(x)$ is a compact convex polytope. Its projection $D^{\mathrm{H}}_j(x)$ is nonempty and compact when $C^{\mathrm{H}}_{j,z}(x)$ is nonempty. Every projected image is in $[0,1]^V$ and matches the observed level-ℓ averages. Let c be the projection of $P_j\widehat{\mu}(x)$ onto $D^{\mathrm{H}}_j(x)$. The projection and orthogonality inequalities in the proof of Proposition S5, with the within-block bound of Lemma S1, give the upper bound in (S.19) for $\sup_{y \in C^{\mathrm{H}}_{j,z}(x)}\|y - c\|_V^2$. Block averaging need not preserve the Hölder class. Let c_1 be the projection of c onto the nonempty closed convex class $\mathcal{K}^{(V)}_{s,L}(x)$. For every candidate y, the projection inequality gives $\|y - c_1\|_V^2 \le \|y - c\|_V^2 - \|c - c_1\|_V^2$. Thus a center in that class satisfies the same bound. The within-block argument uses only the stated pairwise grid inequalities.

For the lower bound, we define $k = \lceil \log_4(2M) \rceil$ and $d = 4^k$, which implies that $\ell \le k \le J - 1$ and $2M \le d < 8M$. On the unit square, we define $\phi(t_1, t_2) = \sin(2\pi t_1)\sin(\pi t_2)$. In each level-$k$ block b, we let w_b be the samples of the rescaled ϕ, and define it to be zero outside that block. The function is zero on the block boundary. The number $N = 2^{J-k}$ of grid points in each block direction is an even integer. The identities

$$\frac{1}{N}\sum_{i=0}^{N-1}\sin\{2\pi(i+1/2)/N\} = 0, \qquad \frac{1}{N}\sum_{i=0}^{N-1}\sin^2\{\pi(i+1/2)/N\} = \frac{1}{2}$$

and

$$\frac{1}{N}\sum_{i=0}^{N-1}\sin^2\{2\pi(i+1/2)/N\} = \begin{cases} 1, & N = 2, \\ 1/2, & N \ge 4 \end{cases}$$

follow by summing the geometric series for the corresponding complex exponentials. Thus each w_b has zero average on its block and

$$P_\ell w_b = 0, \qquad \Big\|\sum_{b=1}^{d} u_b w_b\Big\|_V^2 \ge \frac{1}{4d}\sum_{b=1}^{d} u_b^2.$$

If $|u_b| \le a$, the piecewise function with value $u_b\phi$ on the rescaled block has absolute value at most a. Its gradient norm inside every block is at most $\sqrt{5}\pi a 2^k$. It is continuous across boundaries, and integration along a line segment through the finitely many blocks gives the same global Lipschitz bound. For points t, t', its increment is hence at most

$$\min\{2a, \sqrt{5}\pi a 2^k \|t - t'\|_2\} \le B_s a 2^{sk} \|t - t'\|_2^s.$$

We define $a = \min\{\rho, L2^{-sk}/(2B_s)\}$. If $y \in \mathcal{K}^{(V)}_{s,L/2} \cap [\rho, 1-\rho]^V$, then $y \pm \sum_b u_b w_b \in \mathcal{K}^{(V)}_{s,L}(x)$ for every coefficient vector in $[-a, a]^d$ whenever $P_\ell y = z(x)$. This uses the triangle inequality for the Hölder increments and the zero coarse averages of the perturbation.

We apply the vertex argument in the proof of Theorem S5 to $u \mapsto A_x \sum_b u_b w_b$. It gives a kernel vector u^* with at least $d - M$ coordinates of magnitude a. Both perturbed images are accepted on the coverage event, and their half-distance is at least $(a/2)\sqrt{1 - M/d}$. Hence

$$\mathbb{E}\mathrm{rad}_V\{C^{\mathrm{H}}_{A,z}(X)\} \ge \frac{1-\alpha}{2} \min\{\rho, Ld^{-s/2}/(2B_s)\}\sqrt{1 - M/d}.$$

Because $d \ge 2M$, $d < 8M$, and $B_s \ge 2$, this proves (S.20). This completes the proof. □

S1.14 Relation to the coefficientwise Haar result

Theorem S12 uses orthonormal Haar coefficients without the $V^{-1/2}$ normalization and the explicit coefficient-decay class in (S.24). Its constant R must remain fixed in that normalization. A discretization of one continuous image can have Haar coefficients of order $\sqrt{V}$. The fixed-R condition is not automatic under image resampling. Theorem 3 and Lemma S1 give the resolution comparison directly in normalized image units.

The coefficient condition in (S.24) is not a characterization of the full bounded-variation class. Bounded variation has wavelet descriptions involving weak sequence bounds (Cohen et al., 2003). The present statements use their explicit approximation or coefficient conditions. They do not assert that every bounded-variation image satisfies the same pointwise decay with the same constants.

S2 Setup

The complete formulation distinguishes the populations whose coverage is protected from the spatial features whose uncertainty is measured. The shift weights specify the protected target populations. The image functionals specify which spatial differences contribute to set size.

S2.1 Data structure and covariate shift

Covariates may change across populations while the conditional image distribution remains the same.

Let $X \in \mathcal{X} \subset \mathbb{R}^p$ denote the covariates, and let $Y \in \mathcal{Y} \subset \mathbb{R}^V$ denote an image-valued response observed on a grid of size V. We consider a source population P_{src} and a target population P_{tar} on (X, Y). The labeled source sample is $\mathcal{D}_{\mathrm{src}} = \{(X_i, Y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{\mathrm{src}}$, and the unlabeled target covariate sample is $\mathcal{D}^X_{\mathrm{tar}} = \{X_{n+j}\}_{j=1}^m \overset{\text{i.i.d.}}{\sim} P_{\mathrm{tar},X}$. We assume a common conditional response distribution under covariate shift. There exists a Markov kernel $P_{Y|X=x}$ on $\mathcal{Y}$ such that

$$P_{\mathrm{src}}(dx, dy) = P_{\mathrm{src},X}(dx)P_{Y|X=x}(dy), \qquad P_{\mathrm{tar}}(dx, dy) = P_{\mathrm{tar},X}(dx)P_{Y|X=x}(dy). \quad \text{(S.21)}$$

Definition S1. Let $\mathcal{F}$ be a nonempty collection of nonnegative measurable functions $f : \mathcal{X} \to \mathbb{R}_+$ such that $\mathbb{E}_{P_{\mathrm{src},X}}\{f(X)\} = 1$. Each $f \in \mathcal{F}$ induces a shifted distribution $dP_f(x, y) = f(x)\, dP_{\mathrm{src}}(x, y)$.

The weight f changes the covariate distribution while preserving the conditional image distribution in (S.21).

We consider a finite-dimensional linear shift class $\mathcal{F}_d$. Let $\phi : \mathcal{X} \to \mathbb{R}^d$ be a measurable feature map, define $B(x) = (1, \phi_1(x), \ldots, \phi_d(x))^\top$, let $\Theta_{\mathrm{aff}} \subseteq \mathbb{R}^{d+1}$ be a fixed parameter set, and define

$$\mathcal{F}_d = \left\{ f_\theta(x) = \theta^\top B(x) : \theta \in \Theta_{\mathrm{aff}},\ f_\theta(x) \ge 0\ \forall x,\ \mathbb{E}_{P_{\mathrm{src},X}}\{f_\theta(X)\} = 1 \right\}. \quad \text{(S.22)}$$

Table 3: Summary of notation.

Symbol	Meaning
$X \in \mathcal{X} \subset \mathbb{R}^p$	covariates
$Y \in \mathcal{Y} \subset \mathbb{R}^V$	image-valued response on a grid of size V
$d_{\mathcal{Y}}$	reference metric on $\mathcal{Y}$ used in Wasserstein distances W_1
$P_{\mathrm{src}}, P_{\mathrm{tar}}$	source and target joint distributions on (X, Y)
$\mathcal{D}_{\mathrm{src}} = \{(X_i, Y_i)\}_{i=1}^n$	labeled source sample
$\mathcal{D}_{\mathrm{tar}}^X = \{X_{n+j}\}_{j=1}^m$	unlabeled target covariates
$\mathcal{D}_{\mathrm{tr}}, \mathcal{D}_{\mathrm{cal}}$	training and calibration folds, $n_{\mathrm{cal}} = \lvert\mathcal{I}_{\mathrm{cal}}\rvert$
$\mathcal{F}$	class of shift weights $f : \mathcal{X} \to \mathbb{R}_+$ with $\mathbb{E}_{P_{\mathrm{src},X}}\{f(X)\} = 1$
$\mathcal{F}_d$	finite-dimensional linear shift class (S.22)
$\Theta_{\mathrm{aff}}, \Theta_{\mathrm{tilt}}$	linear shift and ratio estimation parameter sets
$\mathcal{H}$	class of scalar image functionals $h : \mathcal{Y} \to \mathbb{R}$
$\mathcal{X}_q, \mathcal{H}_q$	finite-query set and screened functional subclass
$B_{\mathcal{H}}, L_{\mathcal{H}}$	image-functional boundedness and Lipschitz constants
$\varepsilon_W, \varepsilon_K$	screened W_1/Kullback–Leibler teacher distance bounds
$\widetilde{\mu}_h(x), \widetilde{\sigma}_h(x)$	teacher conditional mean and std of $h(Y)$
C_{sc}	screened score perturbation constant
B_F, A_F, v_F	shift class boundedness and metric-entropy constants
$A_{\mathcal{F}}$	target average bound for the radial Lipschitz coefficient
$\mu_h^\circ(x), \sigma_h^\circ(x)$	oracle conditional mean and std of $h(Y) \mid X = x$
$\widehat{\mu}_h(x), \widehat{\sigma}_h(x)$	teacher-sample plug-in estimates
$R^\circ(x, y)$	oracle sup-norm score (S.25)
$R_n(x, y)$	predictor-informed sup-norm score (S.26)
$\mathcal{C}_q(x)$	score ball $\{y : R_n(x, y) \le q\}$
$\mathrm{rad}_{\mathcal{H}}(C)$	$\mathcal{H}$-radius of a set $C \subseteq \mathcal{Y}$
$B(z), B_i$	basis map and calibration basis vector
$V_{x,y}(q), \bar{B}_x$	candidate-augmented empirical moment vectors (S.27)
$\pi_{x,y}$	candidate-augmented mixture weights, exact finite-dimensional case
$\mathcal{Q}_{\mathrm{cal}}, \widehat{\pi}$	calibration threshold-representative grid and static threshold mixture (S.46)
$\widehat{a}_{\mathrm{fd}}^+(x, y)$	candidate-wise randomized exclusion probability
$\widehat{C}_{\mathrm{fd}}^+(x; U_0)$	transductive exact finite-dimensional prediction set
$\widehat{q}$	calibrated threshold, generic discrepancy case (S.30)
$q_{\alpha,\mathcal{F}}^\circ$	worst-case oracle $(1-\alpha)$-quantile
$q_{\alpha,\mathcal{F},q}^\circ$	screened worst-case oracle quantile of R_q°
$r_{0,q}, c_{-,q}$	screened local-quantile radius and lower density constant
$\psi_n(\delta, \mathcal{F})$	calibration empirical-process bound (S.34)
$e_n(\delta)$	predictor-informed score perturbation rate (S.33)
$\kappa(r, \mathcal{F})$	total-variation approximation error of true ratio r by $\mathcal{F}$
α, λ, K	miscoverage, ridge parameter, teacher sample size

For a real-valued function g, we write $\|g\|_\infty = \sup_z |g(z)|$ over its domain. When $\mathcal{F}_d$ is used with the calibration empirical-process bound, we choose Θ_{aff} such that $\sup_{f \in \mathcal{F}_d} \|f\|_\infty \le B_d < \infty$. This holds when Θ_{aff} is compact and B is uniformly bounded on $\mathcal{X}$. Note that the exact-validity result in Theorem S7 does not require this uniform boundedness.

We also consider a generic nonparametric shift class $\mathcal{F}$, a uniformly bounded class of shift weights with controllable empirical complexity; examples include Vapnik–Chervonenkis (VC) subgraph classes, finite subgroup families, bounded neural-network classes, and reproducing kernel Hilbert space (RKHS) balls intersected with positivity constraints. In the medical reconstruction analyses, observed clinical status is part of X. The protected class contains normalized indicators of these observed groups and their nonnegative mixtures. Equation (4) defines the implemented rule; the generic and linear moment formulations below describe the additional procedures.

S2.2 Image functionals and sizes

Regional averages and linear contrasts specify which spatial features the prediction set should describe. We allow a general measurable class and impose moment conditions only when its conditional moments are used.

Definition S2. Let $\mathcal{H}$ be a nonempty class of measurable maps $h : \mathcal{Y} \to \mathbb{R}$. When the oracle-normalized score R° is used, each $h \in \mathcal{H}$ is assumed to be conditionally square-integrable under the source distribution.

For the empty set, we define $\mathrm{rad}_{\mathcal{H}}(\emptyset) = 0$. For a nonempty set $C \subseteq \mathcal{Y}$, we define

$$\mathrm{rad}_{\mathcal{H}}(C) = \inf_{y_0 \in \mathcal{Y}} \sup_{y \in C} \sup_{h \in \mathcal{H}} |h(y) - h(y_0)| . \tag{S.23}$$

The $\mathcal{H}$-radius measures the size of the prediction set along the directions of the summaries in $\mathcal{H}$. By centering the summaries around a reference image y_0, this radius reports how far the images in the set can spread along the spatial features of interest.

Expectations of radii of data-dependent prediction sets are read as outer expectations when the radius map is not measurable. Appendix S11 states this measurability convention in full.

The choice of $\mathcal{H}$ fixes which spatial features the reported size describes, and it sets the constants that enter the efficiency bounds. Pixel projections report the largest pixelwise deviation. Regional averages report uncertainty at the scale of prespecified regions. Sparse contrasts on disjoint regions report differences such as local background, left against right, or foreground against background. Wavelet or multiscale coefficients report scale-localized features. A larger class describes more features, but it raises the cost of controlling the reported size through its boundedness, Lipschitz, and cardinality constants. This trade-off between the features reported and the cost of controlling them is made precise by the efficiency bounds in Appendix S4.

The classes above are generic, and the radius definition does not use the spatial arrangement of the V pixels. These definitions also apply to an arbitrary vector response in $\mathbb{R}^V$. We make the two-dimensional grid explicit and identify the image regularity that the calibrated set exploits. Let $\mathcal{Y} \subseteq [0,1]^V$ consist of images on a $2^J \times 2^J$ grid with $V = 4^J$ pixels, and let $\{\psi_{j,k}\}$ be an orthonormal two-dimensional Haar system on the grid. This system has a coarse scaling functional and, at each detail level $j \in \{1, \dots, J\}$, the $3 \cdot 4^{j-1}$ detail functionals indexed by orientation and location k, with $j = 1$ the coarsest level and $j = J$ the finest. Each functional $h_{j,k}(y) = \langle \psi_{j,k}, y \rangle$ is an element of $\mathcal{H}$, with the normalization $\|\psi_{j,k}\|_2 = 1$. A level-j detail functional is supported on 4^{J-j+1} pixels, which implies that $\|\psi_{j,k}\|_\infty = 2^{j-1}/\sqrt{V}$ and $|h_{j,k}(y)| \le \|y\|_2$ for every $y \in \mathcal{Y}$.

Definition S3. The multiresolution class $\mathcal{H}^{\mathrm{MR}}$ consists of the coarse scaling functional and the Haar detail functionals $h_{j,k}$ at every detail level $1 \le j \le J$. For $0 \le J' \le J$, the level-J' truncation $\mathcal{H}^{\mathrm{MR}}_{\le J'}$ consists of the scaling functional and the detail functionals at levels $1 \le j \le J'$. A class $\mathcal{H}$ is scale complete if, for every detail level $1 \le j \le J$, it contains at least one detail functional $h_{j,k}$.

The truncation contains exactly $\left|\mathcal{H}^{\mathrm{MR}}_{\le J'}\right| = 4^{J'}$ functionals, with $\mathcal{H}^{\mathrm{MR}}_{\le 0}$ equal to the scaling functional alone. This count is the summary budget in Theorem S12.

Definition S4. Fix $s > 0$ and $0 < R < \infty$. A model belongs to the image-regularity class $\mathcal{R}_V(s, R)$ if the response space satisfies $\mathcal{Y} = \mathcal{Y}_R$, where $\mathcal{Y}_R$ is the image set with the specified Haar coefficient decay

$$\mathcal{Y}_R = \left\{ y \in [0,1]^V : \max_k |h_{j,k}(y)| \le R\, 2^{-sj} \text{ for every detail level } 1 \le j \le J \right\}. \tag{S.24}$$

It belongs to the unstructured class $\mathcal{V}_V(\underline{c})$ if $\mathcal{Y} = [0,1]^V$ and, for $P_{\mathrm{src},X}$-almost every x and every detail level $1 \le j \le J$, there is a location k such that

$$\operatorname*{ess\,sup}_{P_{Y|X=x}} |h_{j,k}(Y) - \mathbb{E}\{h_{j,k}(Y) \mid X = x\}| \ge \underline{c} > 0.$$

The condition specifies a pointwise decay rate for every candidate image. It is not a characterization of the full bounded-variation class. The scaling coefficient is unconstrained. The normalized spatial-approximation class in (S.11) has a separate guarantee in Theorem 3. The constant R is held fixed as the resolution V grows. Because $\mathcal{Y} = \mathcal{Y}_R$ under $\mathcal{R}_V(s, R)$, every prediction set is contained in $\mathcal{Y}_R$ and its candidate images share the decay. The unstructured class is the scale-free case in which detail energy persists at every level.

S2.3 Oracle and predictor-informed scores

We use a score based on the true conditional means and standard deviations as a theoretical benchmark, called the oracle score. The implemented score uses the empirical moments of the returned predictor images.

For $h \in \mathcal{H}$, let $\mu_h^\circ(x) = \mathbb{E}\{h(Y) \mid X = x\}$ and $\sigma_h^\circ(x) = \mathrm{Var}\{h(Y) \mid X = x\}^{1/2}$ denote the oracle conditional location and scale. When these conditional moments enter pointwise suprema over $x \in \mathcal{X}$, we work with fixed measurable versions for which the stated suprema and bounds hold. We fix a miscoverage level $\alpha \in (0, 1)$ and a ridge parameter $\lambda > 0$. The oracle conformity score is defined as

$$R^\circ(x, y) = \sup_{h \in \mathcal{H}} \frac{|h(y) - \mu_h^\circ(x)|}{\sigma_h^\circ(x) + \lambda}. \tag{S.25}$$

The score is the largest standardized deviation of the candidate image from its conditional center over the summaries in $\mathcal{H}$, and a threshold on this single scalar simultaneously constrains every summary in the class. The ridge parameter λ keeps the standardization bounded when the conditional scale is small.

The external predictor, which we call the teacher, is available only through queries. At each queried covariate value x, it returns conditional image samples $\widetilde{Y}^{(1)}(x), \ldots, \widetilde{Y}^{(K)}(x)$ drawn independently from a query distribution Q_x^{T} on $\mathcal{Y}$ that approximates the conditional distribution $P_{Y|X=x}$. From these outputs, we construct estimators $\widehat{\mu}_h$ and $\widehat{\sigma}_h$. When teacher samples are available, these estimators are the plug-in moments:

$$\widehat{\mu}_h(x) = K^{-1} \sum_{k=1}^{K} h\{\widetilde{Y}^{(k)}(x)\}, \qquad \widehat{\sigma}_h^2(x) = K^{-1} \sum_{k=1}^{K} \left[h\{\widetilde{Y}^{(k)}(x)\} - \widehat{\mu}_h(x)\right]^2.$$

When only a fitted image is returned, the conditional theory treats the location and scale functions as externally specified. The fitted image provides the location component, and the scale component is specified either by a separate teacher output or by a deterministic external convention fixed before calibration. For each $h \in \mathcal{H}$, we let

$$S_{n,h}(x, y) = \frac{|h(y) - \widehat{\mu}_h(x)|}{\widehat{\sigma}_h(x) + \lambda}$$

denote the component predictor score. The predictor-informed conformity score is then defined as

$$R_n(x, y) = \sup_{h \in \mathcal{H}} S_{n,h}(x, y). \tag{S.26}$$

For a generic score $S : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and a threshold $t \in \mathbb{R}$, we define the level set $\mathcal{C}_t^S(x) = \{y \in \mathcal{Y} : S(x, y) \le t\}$. When S is nonnegative, this definition implies $\mathcal{C}_t^S(x) = \emptyset$ for $t < 0$. Given a threshold $q \ge 0$, we define the predictor score ball

$$\mathcal{C}_q(x) = \mathcal{C}_q^{R_n}(x) = \{y \in \mathcal{Y} : R_n(x, y) \le q\}.$$

S3 Calibration method

Two calibration constructions give coverage guarantees beyond the observed-group setting. The candidate-augmented solver is exact for finite-dimensional shift classes, and the generic discrepancy rule bounds undercoverage for broader classes. They are separate from the deterministic group rule in Algorithm 1 used in the medical and resolution analyses.

S3.1 Split calibration

We split the labeled source data into training and calibration parts $\mathcal{D}_{\mathrm{src}} = \mathcal{D}_{\mathrm{tr}} \cup \mathcal{D}_{\mathrm{cal}}$, where $\mathcal{D}_{\mathrm{cal}} = \{(X_i, Y_i)\}_{i \in \mathcal{I}_{\mathrm{cal}}}$ and $n_{\mathrm{cal}} = |\mathcal{I}_{\mathrm{cal}}|$. We use the training fold $\mathcal{D}_{\mathrm{tr}}$ to fix score-construction choices before the calibration responses enter the procedure. We query the teacher at every covariate value where the score is evaluated. For $i \in \mathcal{I}_{\mathrm{cal}}$, we let $R_i = R_n(X_i, Y_i)$ denote the calibration scores.

S3.2 FINITE-DIMENSIONAL TRANSDUCTIVE EXACT CALIBRATION

Under the exact finite-dimensional regime, we augment the calibration equations with the candidate deployment point to balance the covariate shift directly. Here we use a finite measurable score R_n with its pre-calibration construction randomness fixed. For each deployment covariate $x \in \mathcal{X}$ and candidate image $y \in \mathcal{Y}$, we add (x, y) to the calibration fold and solve a finite-dimensional moment equation on the augmented sample.

We define $B_i = B(X_i)$ and $R_i = R_n(X_i, Y_i)$ for $i \in \mathcal{I}_{\text{cal}}$, we define $B_\star(x) = B(x)$ and $R_\star(x, y) = R_n(x, y)$, we define $N = n_{\text{cal}} + 1$, and we let $\mathcal{Q}_{x,y}$ denote the finite threshold grid formed by the distinct augmented scores and two sentinel values below and above them. For $q \in \mathcal{Q}_{x,y}$, we define

$$V_{x,y}(q) = \frac{1}{N}\Big\{ \sum_{i \in \mathcal{I}_{\text{cal}}} B_i \mathbf{1}\{R_i > q\} + B_\star(x)\mathbf{1}\{R_\star(x, y) > q\}\Big\},\ \bar{B}_x = \frac{1}{N}\Big\{ \sum_{i \in \mathcal{I}_{\text{cal}}} B_i + B_\star(x)\Big\}. \tag{S.27}$$

We define $\Delta(\mathcal{Q}_{x,y})$ as the set of probability vectors on $\mathcal{Q}_{x,y}$. We then choose a feasible probability vector $\pi_{x,y} \in \Delta(\mathcal{Q}_{x,y})$ satisfying

$$\sum_{q \in \mathcal{Q}_{x,y}} \pi_{x,y}(q) V_{x,y}(q) = \alpha \bar{B}_x \tag{S.28}$$

by a fixed measurable permutation equivariant selection rule. The rule treats reordered versions of the same augmented sample identically. The moment equation (S.28) balances the augmented exceedance frequencies against the nominal level in every direction of the basis B. A solution exists because the sentinels place $\alpha\bar{B}_x$ in the convex hull of the vectors $V_{x,y}(q)$. The candidate-wise randomized exclusion probability and the prediction set obtained by this candidate-specific calibration are

$$\widehat{a}^+_{\text{fd}}(x, y) = \sum_{q \in \mathcal{Q}_{x,y}} \pi_{x,y}(q)\mathbf{1}\{R_n(x, y) > q\}, \qquad \widehat{\mathcal{C}}^+_{\text{fd}}(x; U_0) = \{y \in \mathcal{Y} : \widehat{a}^+_{\text{fd}}(x, y) \le U_0\}, \tag{S.29}$$

where $U_0 \sim \text{Unif}(0, 1)$ is drawn once, independently of all data. Each candidate evaluation solves a linear program over the augmented threshold grid. A feasible solution supported on at most $d + 2$ thresholds always exists. Appendix S6.1 specifies the threshold grid, the feasible selector, the algorithm, and the feasibility and Rao–Blackwellization properties of this construction.

The numerical comparison in Appendix S8 uses a static variant and not a candidate-wise solution. The static variant solves the same moment system once on the calibration fold alone and deploys a single mixture of thresholds at every covariate value. Proposition S16 shows that this variant is feasible for finite calibration scores. Under its stated conditions, coverage deviates from the nominal level by at most the calibration empirical-process bound with high probability. This deviation bound holds in both directions and for every shift in $\mathcal{F}_d$. Coverage here averages over the threshold drawn from the mixture and an independent target observation.

S3.3 GENERIC-CLASS CALIBRATION

For a generic shift class $\mathcal{F}$, we calibrate the prediction sets by selecting a threshold that controls the worst-case empirical undercoverage. We let $\Gamma^+_n(q; \mathcal{F})$ denote the one-sided empirical calibration discrepancy:

$$\Gamma^+_n(q; \mathcal{F}) = \sup_{f \in \mathcal{F}} \frac{1}{n_{\text{cal}}} \sum_{i \in \mathcal{I}_{\text{cal}}} f(X_i)\big\{\mathbf{1}\{R_n(X_i, Y_i) > q\} - \alpha\big\}.$$

The discrepancy $\Gamma^+_n(q; \mathcal{F})$ measures the maximum empirical violation of the nominal coverage level over the candidate shifts in $\mathcal{F}$. The one-sided criterion limits empirical undercoverage. It permits higher coverage. The radius bounds control set size separately. Given a tolerance level $\tau_n \ge 0$, we define the threshold

$$\widehat{q} = \inf\{q \ge 0 : \Gamma^+_n(q; \mathcal{F}) \le \tau_n\}. \tag{S.30}$$

The following proposition shows that the threshold (S.30) is well defined.

Proposition S7. *Suppose $R_n : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is measurable and every $f \in \mathcal{F}$ is nonnegative. Then, for every $\tau_n \ge 0$, the feasible set is nonempty, $\widehat{q} < \infty$, and the infimum is attained. If, in addition, $\mathcal{F}$ is pointwise measurable, then $\widehat{q}$ admits a measurable version.*

Proposition S7 holds because all strict-exceedance indicators equal zero beyond the largest calibration score. The empirical criterion changes value only on the finite grid of realized scores.

S3.4 Coverage constraints of the calibration rules

The candidate-augmented solver and the generic discrepancy rule address shift-weighted exceedance constraints but have different outputs and guarantees. The candidate-augmented solver is candidate dependent and attains exact finite-dimensional coverage after averaging over the calibration sample, the independent target draw, and U_0. The discrepancy rule returns one threshold and controls worst-case undercoverage up to its tolerance and empirical-process error. Neither construction is a component of Algorithm 1. The query-only assumption enters only through the conformity score R_n in (S.26). The method requires no access to the teacher's gradients, architecture, or training data.

S4 Theoretical guarantees

Calibration controls target coverage for the specified shifts. The size of the prediction set also depends on predictor accuracy and image regularity. The shift class $\mathcal{F}$ determines the range of protected target populations, and the image-functional class $\mathcal{H}$ determines the directions of variation that the prediction set controls.

We first show that a prediction set determined by a low-dimensional summary of the image has spatial radius at least half the largest spatial difference between images with the same accepted summary value. A positive spatial difference must be established for the representation class, and dimension reduction alone does not imply it. We then keep the image in its original space and establish exact coverage for finite-dimensional shift classes, threshold localization and one-sided coverage for generic shift classes, and efficiency and resolution bounds for the predictor-informed score. Appendices S11 and S12 contain the proofs.

S4.1 Spatial limits of image summaries

A set determined by a summary must accept every image with the same accepted summary value. The images sharing a summary value form its fiber. Differences among those images impose a lower bound on the spatial radius.

Definition S5. Let $\mathcal{S}_M$ be a class of measurable maps $S : \mathcal{Y} \to \mathbb{R}^M$, let $\widehat{S}$ be any random element of $\mathcal{S}_M$, and let $\widehat{A}(x) \subseteq \mathbb{R}^M$ be a measurable, data-dependent acceptance region indexed by $x \in \mathcal{X}$. The associated adaptive prediction set takes the form $\widehat{\mathcal{C}}^{(M)}(x) = \{y \in \mathcal{Y} : \widehat{S}(y) \in \widehat{A}(x)\}$, where, for each fixed x, the membership event $\{y \in \widehat{\mathcal{C}}^{(M)}(x)\}$ is measurable with respect to the sigma-field generated by $\widehat{S}(Y)$. We also assume that $(x, y, \omega) \mapsto \mathbf{1}\{y \in \widehat{\mathcal{C}}^{(M)}(x)(\omega)\}$ is jointly measurable, where ω indexes the candidate construction. The class-level fiber complexity is

$$\Delta_S(t) = \sup_{\substack{y,y' \in \mathcal{Y} \\ S(y)=S(y')=t}} \sup_{h \in \mathcal{H}} |h(y) - h(y')|, \qquad t \in S(\mathcal{Y}),$$

and $\Delta^\star_{\mathcal{S}_M} = \inf_{S \in \mathcal{S}_M} \inf_{t \in S(\mathcal{Y})} \Delta_S(t)$.

For any prediction set $\widehat{\mathcal{C}}^{(M)}$ factoring through an adaptive reduction as in Definition S5, we define $\Delta_{\widehat{S},\widehat{A}}(x) = \sup_{t \in \widehat{A}(x) \cap \widehat{S}(\mathcal{Y})} \Delta_{\widehat{S}}(t)$, with the convention that the supremum over the empty set is zero. The following theorem states an accepted-fiber inequality for a data-adaptive representation drawn from $\mathcal{S}_M$. It applies to nonlinear representations, but a positive lower bound requires a separate bound on the accepted fibers.

Theorem S6. *On the event $\widehat{\mathcal{C}}^{(M)}(X) \neq \emptyset$, it holds that*

$$\mathrm{rad}_{\mathcal{H}}\{\widehat{\mathcal{C}}^{(M)}(X)\} \geq \frac{1}{2}\Delta_{\widehat{S},\widehat{A}}(X) \geq \Delta^\star_{\mathcal{S}_M}/2 \tag{S.31}$$

almost surely.

Theorem S6 characterizes the limitations of data-dependent reductions. Along any sequence for which $\Delta^{\star}_{\mathcal{S}_M}$ and $\mathbb{P}\{\widehat{\mathcal{C}}^{(M)}(X) \neq \emptyset\}$ are bounded away from zero, the expected $\mathcal{H}$-radius cannot converge to zero. The result identifies accepted fibers that a reduction cannot distinguish. The positive regularity result controls these omitted directions through an explicit response class. Theorem S23 derives a complementary impossibility result for unrestricted shifts by reducing the problem to the scalar conditional coverage obstruction.

S4.2 Finite-dimensional exact validity

The class $\mathcal{F}_d$ is nonempty whenever the parameter set Θ_{aff} contains the constant weight vector $\theta = (1, 0, \ldots, 0)^{\top}$. Subgroup conditional applications impose that the normalized subgroup indicators belong to the linear class $\mathcal{F}_d$.

We fix any pre-calibration score-construction randomness used to construct a finite measurable score $R_n : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, and let $\widehat{\mathcal{C}}^{+}_{\mathrm{fd}}(x; U_0)$ denote the randomized set produced by Algorithm 2. The following theorem establishes exact finite-sample target-population validity for the candidate-augmented solver.

Theorem S7. *Suppose $\mathcal{F}_d$ is nonempty. Then, conditional on the training fold, for every $f \in \mathcal{F}_d$ and every independent target draw $(X^*, Y^*) \sim P_f$, it holds that*

$$\mathbb{E}\left[\mathbf{1}\{Y^* \notin \widehat{\mathcal{C}}^{+}_{\mathrm{fd}}(X^*; U_0)\} \mid \mathcal{D}_{\mathrm{tr}}\right] = \alpha, \quad \mathbb{P}\{Y^* \in \widehat{\mathcal{C}}^{+}_{\mathrm{fd}}(X^*; U_0) \mid \mathcal{D}_{\mathrm{tr}}\} = 1 - \alpha. \tag{S.32}$$

Theorem S7 shows that the candidate-augmented solver attains exactly the nominal coverage $1 - \alpha$ in finite samples for every shift in $\mathcal{F}_d$, without asymptotic sample sizes or score concentration. Gibbs et al. (2025) obtain exact coverage over finite-dimensional shift classes by randomized augmented quantile regression. Our solver targets the same guarantee through a mixture of thresholds that balances the basis moments. The identities in (S.32) hold for each fixed realization of the pre-calibration randomness and average over the calibration sample, the independent target draw, and the auxiliary randomization U_0. Corollary S3 states the resulting subgroup conditional validity whenever the normalized subgroup indicators belong to $\mathcal{F}_d$.

S4.3 Score perturbation and approximate validity

The validity analysis for generic shift classes compares the calibrated threshold with oracle quantiles of the score distribution under each shifted target population.

Definition S6. For $f \in \mathcal{F}$, we define

$$F^{\circ}_f(q) = P_f\{R^{\circ}(X, Y) \leq q\}, \quad q^{\circ}_{f,\alpha} = \inf\{q : F^{\circ}_f(q) \geq 1 - \alpha\},$$

and define the worst-case oracle quantile $q^{\circ}_{\alpha,\mathcal{F}} = \sup_{f \in \mathcal{F}} q^{\circ}_{f,\alpha}$. For a data-dependent prediction set $\widehat{\mathcal{C}}$, we define

$$\Delta_{\mathcal{F}}(\widehat{\mathcal{C}}) = \sup_{f \in \mathcal{F}} \left| P_f\{Y \in \widehat{\mathcal{C}}(X)\} - (1 - \alpha) \right|, \qquad \Delta^{-}_{\mathcal{F}}(\widehat{\mathcal{C}}) = \sup_{f \in \mathcal{F}} \left[(1 - \alpha) - P_f\{Y \in \widehat{\mathcal{C}}(X)\} \right]_{+}.$$

The worst-case oracle quantile protects all populations in the specified shift class. The two error criteria distinguish coverage shortfall from deviation in either direction.

Condition S1 (Common oracle quantile)**.** It holds that $q^{\circ}_{f,\alpha} = q^{\circ}_{\alpha,\mathcal{F}}$ for every $f \in \mathcal{F}$.

Condition S1 requires the same oracle quantile in every protected population. It gives the two-sided bound in Corollary S1 with a single threshold. The one-sided results below do not use it. The following proposition records the score error proved from the population moment and entropy conditions in Assumption S2. We let e_n be defined by (S.53).

Proposition S8. *Suppose Assumption S2 holds and the teacher random maps are independent of the calibration sample. Then, for every $\delta \in (0, 1)$,*

$$\mathbb{P}\left\{ \sup_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} |R_n(x, y) - R^{\circ}(x, y)| > e_n(\delta) \right\} \leq \delta. \tag{S.33}$$

Proposition S8 controls the error in the image score. Theorem S13 gives its complete proof. No sample-dependent score event is imposed as a primitive assumption.

The following assumption separates the population quantile from thresholds that are too small. It bounds the probability below the quantile and does not require a continuous score distribution.

Assumption S1 (Local quantile regularity)**.** The worst-case oracle quantile satisfies $q^{\circ}_{\alpha,\mathcal{F}} < \infty$. There exist constants $r_0, c_- > 0$ such that, for every $f \in \mathcal{F}$ with $q^{\circ}_{f,\alpha} \geq q^{\circ}_{\alpha,\mathcal{F}} - r_0$ and every $0 < t \leq r_0$, $F^{\circ}_f(q^{\circ}_{f,\alpha} - t) \leq 1 - \alpha - c_- t$.

Assumption S1 ensures that reducing a near-worst-case oracle threshold creates a definite coverage shortfall. A uniform positive lower bound for the densities on a common left neighborhood and $F^{\circ}_f(q^{\circ}_{f,\alpha}) = 1 - \alpha$ suffice, but are not required. The condition also allows an atom that raises the distribution function above $1 - \alpha$. Proposition S33 gives a discrete example. The following proposition controls the shift-weighted exceedance process using the population entropy condition. We let ψ_n be defined by (S.53).

Proposition S9. *Suppose Assumption S3 holds and the score is fixed independently of the i.i.d. calibration observations. Then, for every $\delta \in (0,1)$,*

$$\mathbb{P}\Big[\sup_{f\in\mathcal{F}}\sup_{q\in\mathbb{R}}\Big|\frac{1}{n_{\mathrm{cal}}}\sum_{i\in\mathcal{I}_{\mathrm{cal}}} f(X_i)\Big\{\mathbf{1}\{R_n(X_i,Y_i) > q\} - \alpha\Big\} - \Big(P_f\{R_n(X,Y) > q\} - \alpha\Big)\Big| > \psi_n(\delta,\mathcal{F})\Big] \leq \delta. \tag{S.34}$$

Proposition S9 permits evaluation at the data-dependent threshold $\widehat{q}$ because its bound is uniform in q. Theorem S14 gives the full proof from population boundedness and entropy. The following proposition gives an algebraic perturbation bound for the predictor-informed score.

Proposition S10. *Suppose*

$$\sup_{x,h}|\widehat{\mu}_h(x) - \mu^{\circ}_h(x)| \leq a, \qquad \sup_{x,h}|\widehat{\sigma}_h(x) - \sigma^{\circ}_h(x)| \leq b,$$

and suppose that, for every $x \in \mathcal{X}$ and $h \in \mathcal{H}$, $\widehat{\sigma}_h(x) \geq 0$, $\sigma^{\circ}_h(x) \geq 0$. Let

$$M_{\mathcal{H}} = \sup_{x\in\mathcal{X}}\sup_{y\in\mathcal{Y}}\sup_{h\in\mathcal{H}}|h(y) - \mu^{\circ}_h(x)| < \infty.$$

Then, it holds that

$$\sup_{x\in\mathcal{X}}\sup_{y\in\mathcal{Y}}|R_n(x,y) - R^{\circ}(x,y)| \leq \lambda^{-1}a + \lambda^{-2}M_{\mathcal{H}}b. \tag{S.35}$$

Appendix S12 verifies the high-probability bounds (S.33) and (S.34) for the max-score $R_n = \sup_{h\in\mathcal{H}} S_{n,h}$. The finite-screened teacher-quality and query-complexity results use the same theory on a finite queried covariate set. Their conclusions do not extrapolate score behavior to covariates at which the teacher has not been queried.

We fix $0 < \delta < 1$, define $e_{n,\delta} = e_n(\delta/2)$ and $\psi_{n,\delta} = \psi_n(\delta/2,\mathcal{F})$, and let $\widehat{q}$ denote the attained measurable version of (S.30) computed with a deterministic tolerance $\tau_n \geq \psi_{n,\delta}$. The following theorem localizes the empirical threshold around the worst-case oracle threshold.

Theorem S8. *Suppose $\mathcal{F}$ is pointwise measurable, the score perturbation bound (S.33), Assumption S1, and the calibration empirical-process bound (S.34) hold. There exist constants $c_0, C > 0$ depending only on c_- such that, if $e_{n,\delta} + \psi_{n,\delta} + \tau_n \leq c_0 r_0$, then, with probability at least $1 - \delta$, $\big|\widehat{q} - q^{\circ}_{\alpha,\mathcal{F}}\big| \leq C\{\tau_n + \psi_{n,\delta} + e_{n,\delta}\}$.*

Theorem S8 bounds threshold error by the tolerance, calibration error, and score error. It allows discrete scores under Assumption S1. The following corollary controls overcoverage when the populations share an oracle quantile and probability has a local upper bound near that quantile.

Corollary S1. *Suppose the conditions of Theorem S8 hold, the oracle $(1-\alpha)$-quantile is common across $\mathcal{F}$ as in Condition S1, and there exists $c_+ < \infty$ such that, for every $f \in \mathcal{F}$ and every $0 \leq t \leq r_0$, $F^{\circ}_f(q^{\circ}_{f,\alpha} + t) - F^{\circ}_f(q^{\circ}_{f,\alpha} - t) \leq c_+ t$. There exist constants $c_0', C' > 0$ depending only on c_- and c_+ such that, if $e_{n,\delta} + \psi_{n,\delta} + \tau_n \leq c_0' r_0$, then, with probability at least $1 - \delta$, $\Delta_{\mathcal{F}}(\widehat{\mathcal{C}}) \leq C'\{\tau_n + \psi_{n,\delta} + e_{n,\delta}\}$.*

Corollary S1 bounds both overcoverage and undercoverage under its stated conditions. When every shift in $\mathcal{F}$ has the same oracle quantile, the maximal coverage deviation in either direction has an upper bound of the same order as the threshold localization bound. The additional condition bounds the local increase in oracle score probability from above. Appendix S11 proves Corollary S1.

We fix $0 < \delta < 1$, define $\psi_{n,\delta} = \psi_n(\delta, \mathcal{F})$, and let $\widehat{q}$ denote the attained measurable version of (S.30) computed with a deterministic tolerance $\tau_n \geq 0$. The following theorem provides a one-sided validity bound that requires only the calibration empirical-process bound.

Theorem S9. *Suppose $\mathcal{F}$ is pointwise measurable and the calibration empirical-process bound (S.34) holds. Then, with probability at least $1 - \delta$,*

$$\Delta_{\mathcal{F}}^{-}(\widehat{\mathcal{C}}) \leq \tau_n + \psi_{n,\delta}. \tag{S.36}$$

In particular, if $\tau_n = \psi_{n,\delta}$, then $\Delta_{\mathcal{F}}^{-}(\widehat{\mathcal{C}}) \leq 2\psi_{n,\delta}$.

Theorem S9 implies that the proposed prediction set limits the largest coverage shortfall over the specified target populations. This guarantee uses the uniform empirical-process bound on the calibration sample and does not require the predictor score to concentrate for validity. The two-sided bound in Corollary S1 uses the additional common-quantile Condition S1 and the score perturbation bound.

S4.4 Rates and minimax optimality

Appendix S6.6 derives the calibration empirical-process bound (S.34) for finite-dimensional, VC-subgraph, and bounded kernel RKHS shift classes. For the finite-dimensional linear class $\mathcal{F}_d$, Proposition S21 verifies (S.34) with $\psi_n(\delta, \mathcal{F}_d)$ of order $B_d\sqrt{(d+1)/n_{\mathrm{cal}}}$, up to logarithmic factors. With $\tau_n = \psi_n(\delta, \mathcal{F}_d)$, Theorem S9 gives the same order for the one-sided validity bound.

Theorem S15 establishes a matching $\sqrt{d/n}$ lower bound for the subgroup-weighted exceedance calibration problem for even $d \geq 2$ and $n \geq d$. Proposition S24 extends this lower bound from the exceedance estimand to prediction sets. In the Gaussian-cell model, thresholds determined by the labeled source sample define one-sided intervals bounded above in each cell. For every such procedure, some model parameter gives expected worst-case target coverage error at least of order $\sqrt{d/n}$ over the stated bounded d-cell subgroup shift class. The lower bounds concern two-sided conditional coverage error. They do not give a lower bound for undercoverage alone, which an unrestricted set can make zero. The common-quantile condition in Corollary S1 must be checked separately before asserting a matching prediction-set upper bound.

S4.5 Predictor error and expected radius

The size analysis measures a prediction set through its worst-case expected $\mathcal{H}$-radius over the shift class. For linear image summaries on a convex response space, the two radius properties used below follow from the score construction.

Definition S7. For a prediction set $\widehat{\mathcal{C}}$, we define $\mathcal{S}_{\mathcal{F},\mathcal{H}}(\widehat{\mathcal{C}}) = \sup_{f\in\mathcal{F}} \mathbb{E}_f[\mathrm{rad}_{\mathcal{H}}\{\widehat{\mathcal{C}}(X)\}]$, with the expectation interpreted according to the convention following (S.23).

For the following two propositions, let $\mathcal{Y}$ be closed and convex and let every $h \in \mathcal{H}$ be linear. We consider all scores $S(x, y) = \sup_{h\in\mathcal{H}} |h(y) - h\{y_c(x)\}|/a_h(x)$ with $y_c(x) \in \mathcal{Y}$ and $0 < a_h(x) \leq A(x)$, where A is a deterministic measurable function satisfying $\sup_{f\in\mathcal{F}} \mathbb{E}_f A(X) \leq A_{\mathcal{F}} < \infty$. For any such score, we define $\mathcal{C}_q(x) = \{y \in \mathcal{Y} : S(x, y) \leq q\}$. These restrictions concern the response space and the allowed score parameters, not a realized calibration sample.

The first proposition converts a score threshold into an image-radius bound.

Proposition S11. *For every score in the stated family, $\mathrm{rad}_{\mathcal{H}}\{\mathcal{C}_q(x)\} \leq A(x)q$ for all $x \in \mathcal{X}$ and $q \geq 0$.*

Proof. The admissible center $y_c(x)$ satisfies $\sup_{h\in\mathcal{H}} |h(y)-h\{y_c(x)\}| \leq q \sup_h a_h(x) \leq qA(x)$ for every $y \in \mathcal{C}_q(x)$. Taking the infimum over centers proves the bound. This completes the proof. □

The second proposition controls the change in radius when calibration changes the threshold.

Proposition S12. *For every score in the stated family,* $|\mathrm{rad}_{\mathcal{H}}\{\mathcal{C}_{q_2}(x)\} - \mathrm{rad}_{\mathcal{H}}\{\mathcal{C}_{q_1}(x)\}| \le A(x)\,|q_2 - q_1|$ *for every* $x \in \mathcal{X}$ *and every* $q_1, q_2 \ge 0$.

Proof. We fix x and $0 \le q_1 < q_2$. For $y \in \mathcal{C}_{q_2}(x)$, convexity gives $y' = y_c(x) + (q_1/q_2)\{y - y_c(x)\} \in \mathcal{Y}$. Linearity gives $S(x, y') \le q_1$ and $\sup_h |h(y) - h(y')| \le A(x)(q_2 - q_1)$. We apply the triangle inequality at an η-optimal center of $\mathcal{C}_{q_1}(x)$ and let $\eta \downarrow 0$. Set inclusion gives the reverse inequality. Equality and interchanged thresholds cover $q_1 = q_2$ and $q_2 < q_1$. This completes the proof. □

Propositions S11 and S12 derive both radius properties. They are not assumptions about fitted sets. For the bounded linear image score, the predictor-sample mean and the population mean are admissible centers, and $A(x) = B_{\mathcal{H}} + \lambda$ bounds both scales. For nonlinear functionals or an additional intersection that excludes the center, these propositions do not apply automatically. Proposition 2 gives a separate exact calculation for the observed-coarse-image intersection used in the medical analyses.

We fix $0 < \delta < 1$, define $e_{n,\delta} = e_n(\delta/2)$ and $\psi_{n,\delta} = \psi_n(\delta/2, \mathcal{F})$, and let $\widehat{q}$ denote the attained measurable version of (S.30) computed with a deterministic tolerance $\tau_n \ge \psi_{n,\delta}$. The following theorem bounds the worst-case expected $\mathcal{H}$-radius of $\widehat{\mathcal{C}}$ in terms of the worst-case oracle quantile and the score perturbation rate.

Theorem S10. *Suppose Assumptions S2 and S3 hold, the predictor construction is independent of calibration, the score belongs to the family in Proposition S11, and* $q^{\circ}_{\alpha,\mathcal{F}} < \infty$. *Then, with probability at least* $1 - \delta$, $\mathcal{S}_{\mathcal{F},\mathcal{H}}(\widehat{\mathcal{C}}) \le A_{\mathcal{F}}\left[q^{\circ}_{\alpha,\mathcal{F}} + e_{n,\delta}\right]$.

Theorem S10 bounds the radius by the benchmark $A_{\mathcal{F}} q^{\circ}_{\alpha,\mathcal{F}}$ plus $A_{\mathcal{F}} e_{n,\delta}$. The benchmark is an upper bound, not necessarily the exact oracle-set radius. The next theorem compares the actual expected radii under additional regularity. It separates calibration error, systematic predictor error, and the error from finitely many predictor queries. Let $\mathcal{X}_q \subset \mathcal{X}$ be a nonempty finite set of queried covariate values and let $\mathcal{H}_q \subset \mathcal{H}$ be a nonempty finite class of included spatial functionals. We define

$$R_{n,K}(x, y) = \sup_{h \in \mathcal{H}_q} \frac{|h(y) - \widehat{\mu}_h(x)|}{\widehat{\sigma}_h(x) + \lambda}, \qquad R^{\circ}_q(x, y) = \sup_{h \in \mathcal{H}_q} \frac{|h(y) - \mu^{\circ}_h(x)|}{\sigma^{\circ}_h(x) + \lambda}.$$

For scores restricted to these covariates and functionals, we let $\mathcal{S}_{\mathcal{F},\mathcal{H}_q}$ denote the worst-case expected radius with $\mathrm{rad}_{\mathcal{H}}$ replaced by $\mathrm{rad}_{\mathcal{H}_q}$. For $0 < \delta < 1$, we define $\delta_q = \delta/3$, and let $\widehat{q}_{q,K}$ and $\widehat{q}^{\circ}_q$ denote the one-sided calibrated thresholds obtained from (S.30) with tolerance $\tau_{n,q} = 2\psi_{n,q}(\delta_q, \mathcal{F})$ after replacing the score by $R_{n,K}$ and R°_q, respectively. Here $\psi_{n,q}(\cdot, \mathcal{F})$ denotes the screened calibration empirical-process bound in Condition (C2) below. We define

$$\widehat{\mathcal{C}}_{q,K}(x) = \{y : R_{n,K}(x, y) \le \widehat{q}_{q,K}\}, \qquad \widehat{\mathcal{C}}^{\circ}_q(x) = \{y : R^{\circ}_q(x, y) \le \widehat{q}^{\circ}_q\}.$$

For an oracle screened score T, we define $q^{\circ}_{f,\alpha,T} = \inf\{q : P_f\{T(X, Y) \le q\} \ge 1 - \alpha\}$ and $q^{\circ}_{\alpha,\mathcal{F},T} = \sup_{f \in \mathcal{F}} q^{\circ}_{f,\alpha,T}$. When $T = R^{\circ}_q$, we define $q^{\circ}_{f,\alpha,q} = q^{\circ}_{f,\alpha,R^{\circ}_q}$ and $q^{\circ}_{\alpha,\mathcal{F},q} = q^{\circ}_{\alpha,\mathcal{F},R^{\circ}_q}$. For the screened teacher population moments, we define

$$\widetilde{\mu}_h(x) = \mathbb{E}_{Q^{\mathrm{T}}_x} h(Y), \qquad \widetilde{\sigma}_h(x) = \mathrm{Var}_{Q^{\mathrm{T}}_x}\{h(Y)\}^{1/2}, \qquad x \in \mathcal{X}_q,\ h \in \mathcal{H}_q.$$

For the screened teacher distance condition, we assume $\varepsilon_W, \varepsilon_K \in [0, \infty]$, with at least one finite, satisfy $\sup_{x \in \mathcal{X}_q} W_1(Q^{\mathrm{T}}_x, P_{Y|X=x}) \le \varepsilon_W$, $\sup_{x \in \mathcal{X}_q} D_{\mathrm{KL}}(P_{Y|X=x} \| Q^{\mathrm{T}}_x) \le \varepsilon_K$, where W_1 denotes the 1-Wasserstein distance induced by a reference metric $d_{\mathcal{Y}}$ on $\mathcal{Y}$ and D_{KL} denotes the Kullback–Leibler divergence. We let $B_{\mathcal{H}}$ denote a screened bound satisfying $\sup_{h \in \mathcal{H}_q} \sup_{y \in \mathcal{Y}} |h(y)| \le B_{\mathcal{H}}$, and let $L_{\mathcal{H}}$ denote a screened Lipschitz constant for $\mathcal{H}_q$ with respect to $d_{\mathcal{Y}}$ when the Wasserstein term is used. If the Lipschitz condition or the positive lower bound for the sum of the two scales is unavailable, we interpret $L_{\mathcal{H}}\varepsilon_W$ as $+\infty$. If the Kullback–Leibler term is not used, we interpret $B_{\mathcal{H}}\sqrt{\varepsilon_K}$ as $+\infty$. We define C_{sc} as the score perturbation constant in Theorem S16. It depends on a positive lower bound σ_0 for the sum of the two scales only for the linear Wasserstein term. Finally, we define

$$b_{q,K}(\eta) = \min\{L_{\mathcal{H}}\varepsilon_W,\ B_{\mathcal{H}}\sqrt{\varepsilon_K}\} + K^{-1/2}\sqrt{1 + \log(|\mathcal{X}_q||\mathcal{H}_q|/\eta)}. \tag{S.37}$$

For a score pair (S, T) on $\mathcal{X}_q \times \mathcal{Y}$, the conditions of Theorem S11 are grouped into the following screened conditions.

(C1) The query set and shift class satisfy $P_{\mathrm{src},X}(X \in \mathcal{X}_q) = 1$ and $\mathcal{F}$ is pointwise measurable.

(C2) Assumption S3 holds, and the score construction is independent of calibration. The image space is closed and convex, the functionals are linear, and S, T belong to the score family of Proposition S12 with a common scale bound A.

(C3) The bounded functional and screened teacher distance conditions and the teacher sampling condition hold on $\mathcal{X}_q \times \mathcal{H}_q$. For the linear Wasserstein term only, require $\inf_{x,h}\{\sigma_h^\circ(x) + \widetilde{\sigma}_h(x)\} \geq \sigma_0 > 0$. If this scale condition is unavailable, the Kullback–Leibler term is used.

Boundedness in (C3) gives $R_q^\circ \leq 2B_{\mathcal{H}}/\lambda$. This implies finite population quantiles. Theorem S14 derives the calibration bound from (C2), and Proposition S12 proves its radius property. The validity bound in (S.38) uses only the entropy and independence part of (C2). Proposition S7 gives nonemptiness and attainment of the calibrated thresholds. The screened expected-radius comparison in (S.39) uses the score perturbation in (C3) and both radius properties in (C2). Proposition S2 removes the need for a local quantile condition in this comparison.

The following theorem compares the worst-case expected radii of the screened predictor and oracle sets. We fix $0 < \delta < 1$ and define $\delta_q = \delta/3$.

Theorem S11. *Suppose Conditions (C1)–(C3) hold for $(S, T) = (R_{n,K}, R_q^\circ)$. Then there exists a constant $C < \infty$, depending only on $B_{\mathcal{H}}$ and λ, and on the screened scale lower bound when the linear Wasserstein term is used, such that, with probability at least $1 - \delta$,*

$$\Delta_{\mathcal{F}}^{-}(\widehat{\mathcal{C}}_{q,K}) \leq C\psi_{n,q}(\delta_q, \mathcal{F}) \tag{S.38}$$

and

$$\left|\mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\widehat{\mathcal{C}}_{q,K}) - \mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\widehat{\mathcal{C}}_q^\circ)\right| \leq A_{\mathcal{F}} C\left\{\psi_{n,q}(\delta_q, \mathcal{F}) + b_{q,K}(\delta_q)\right\}. \tag{S.39}$$

For the screened scores, Theorem S11 also gives

$$\begin{aligned}&\left|\mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\widehat{\mathcal{C}}_{q,K}) - \mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\widehat{\mathcal{C}}_q^\circ)\right| \\ &\quad\leq A_{\mathcal{F}} C\left[\psi_{n,q}(\delta/3, \mathcal{F}) + \min\{L_{\mathcal{H}}\varepsilon_W, B_{\mathcal{H}}\sqrt{\varepsilon_K}\} + K^{-1/2}\sqrt{1 + \log\{3|\mathcal{X}_q||\mathcal{H}_q|/\delta\}}\right].\end{aligned}$$

Here C has the same dependencies as in that theorem. The Kullback–Leibler term allows zero conditional variances. The linear Wasserstein term requires the positive scale condition in (C3).

Theorem S11 bounds the difference between the worst-case expected radii of the predictor and oracle sets. For the same calibration sample and rule, Proposition S2 gives the sharper bound $2A_{\mathcal{F}}C_{\mathrm{sc}}b_{q,K}(\delta_q)$ without a calibration-error term. Theorem S16 bounds the predictor term using Wasserstein and Kullback–Leibler distances. The query lower bound in Theorem S20 concerns scalar threshold estimation in a separate two-point subproblem.

S4.6 Spatial regularity and summary dimension

The bounds above hold for arbitrary responses in $\mathbb{R}^V$. We measure cost by $M = |\mathcal{H}_q|$, the number of image functionals included in the calibrated multiresolution summary, which equals the effective dimension of the finite-summary reduction in Definition S5. The following theorem controls the deterministic truncation error with a summary budget independent of V for the stated coefficient-decay class and gives a radius lower bound for Haar coordinate reductions on the full image box.

Theorem S12. *Let $\mathcal{H}^{\mathrm{MR}}$ be scale complete as in Definition S3.*

(i) Suppose Conditions (C1)–(C3) hold for the screened pair $(R_{n,K}, R_q^\circ)$ with $\mathcal{H}_q \subseteq \mathcal{H}^{\mathrm{MR}}$. Assume $\mathcal{Y} = \mathcal{Y}_R$ as defined in (S.24), fix $\epsilon \geq 2R\,2^{-sJ}$, take the smallest level cutoff $J' \in \{0, \ldots, J\}$ with $2^{sJ'} \geq 2R/\epsilon$, and set $\mathcal{H}_q = \mathcal{H}^{\mathrm{MR}}_{\leq J'}$. Then, with probability at least $1 - \delta$,

$$\mathcal{S}_{\mathcal{F},\mathcal{H}^{\mathrm{MR}}}(\widehat{\mathcal{C}}_{q,K}) \leq \mathcal{S}_{\mathcal{F},\mathcal{H}^{\mathrm{MR}}_{\leq J'}}(\widehat{\mathcal{C}}_q^\circ) + A_{\mathcal{F}}C\{\psi_{n,q}(\delta_q, \mathcal{F}) + b_{q,K}(\delta_q)\} + \epsilon, \tag{S.40}$$

and the budget satisfies $M = \left|\mathcal{H}^{\mathrm{MR}}_{\leq J'}\right| = 4^{J'} \leq \max\{1, 4(2R/\epsilon)^{2/s}\} = O\{1 + (R/\epsilon)^{2/s}\}$.

(ii) Assume $\mathcal{Y} = [0,1]^V$, fix $\rho_0 \in (0, 1/2]$, and let $\widehat{\mathcal{C}}^{(M)}$ factor through a reduction to a subclass $\mathcal{H}_q \subseteq \mathcal{H}^{\mathrm{MR}}$ of M Haar detail functionals with $1 \le M \le V-2$ that omits at least one detail functional, as in Definition S5. Then, with $c_\star = \rho_0/\sqrt{2}$, on the event that $\widehat{\mathcal{C}}^{(M)}(X)$ contains an image with coordinates in $[\rho_0, 1-\rho_0]^V$,

$$\mathrm{rad}_{\mathcal{H}^{\mathrm{MR}}}\{\widehat{\mathcal{C}}^{(M)}(X)\} \ge \tfrac{c_\star}{2}\sqrt{V/M}. \tag{S.41}$$

Theorem S12 establishes a resolution separation for image-valued prediction. Part (i) bounds the deterministic truncation remainder with a summary budget independent of V when R, s and the tolerance are fixed. Its statistical errors still depend on the screened class. Part (ii) applies to reductions that use Haar detail coordinates and accept an interior image. Within that class, the lower bound grows with resolution unless the number of included summaries also increases. The normalized result in Theorem 3 covers arbitrary linear reductions under its stated interior condition.

Corollary S6 gives finite-family selection over cutoff levels and an oracle inequality for the corresponding truncated-radius criteria. It does not establish adaptation to unknown (s, R).

S4.7 Shift-class misspecification and extensions

The one-sided validity bound also transfers to a misspecified calibration class. Theorem 4 shows that, when the true target ratio $r = dP_{\mathrm{tar},X}/dP_{\mathrm{src},X}$ is approximated by the specified class $\mathcal{F}$, it holds that

$$P_{\mathrm{tar}}\{Y \in \widehat{\mathcal{C}}(X)\} \ge 1 - \alpha - \Delta^{-}_{\mathcal{F}}(\widehat{\mathcal{C}}) - \kappa(r, \mathcal{F}), \qquad \kappa(r, \mathcal{F}) = \frac{1}{2}\inf_{f \in \mathcal{F}} \mathbb{E}_{P_{\mathrm{src},X}}|r(X) - f(X)|.$$

This bound shows that the loss relative to nominal target coverage is at most the sum of the calibrated one-sided error and the total-variation approximation error $\kappa(r, \mathcal{F})$ of the specified shift class.

Appendix S6 estimates shift weights semiparametrically and gives coverage bounds that hold at every prediction horizon under changing shifts. It also studies augmented calibration, selection from a finite family of image-functional classes, and the separate effects of calibration complexity and predictor accuracy.

S5 Numerical results

Appendix S7 gives the neural super-resolution study. The analyses of the five cohorts compare the group thresholds, their maximum as a common threshold, and the static randomized moment rule. Appendix S8 gives the disjoint medical reconstruction cohorts and direct conditional-conformal comparisons. Appendix S9 gives the normalized image-radius and summary-budget experiment. The comparisons report realized and randomization-averaged coverage separately for the static rule.

S6 Supporting results and extensions

Finite-dimensional shifts admit exact augmented moment balance. More general shift classes require empirical-process control. Distribution approximation and finitely many predictor queries affect the size comparisons.

S6.1 Candidate-augmented exact finite-dimensional calibration

The construction in (S.27)–(S.29) balances augmented exceedances on a finite threshold grid. The ordered grid $\mathcal{Q}_{x,y}$ consists of the distinct values in $\{R_i : i \in \mathcal{I}_{\mathrm{cal}}\} \cup \{R_\star(x,y)\}$ and of a lower sentinel $q_-(x,y)$ smaller than all augmented scores and an upper sentinel $q_+(x,y)$ greater than or equal to all augmented scores. The feasible vector $\pi_{x,y}$ in (S.28) is selected by a fixed measurable permutation equivariant rule. This ensures that relabeling the augmented sample relabels the induced pointwise noncoverage probabilities in the same way. Specifically, we take the selector to be the lexicographically smallest element of the feasible polytope after ordering $\mathcal{Q}_{x,y}$ increasingly. Because the polytope is closed in the finite simplex and is defined by the augmented sample through symmetric sums, this rule is measurable and permutation equivariant.

Algorithm 2 Transductive exact finite-dimensional calibration

1: Construct the predictor-informed score R_n from the training fold.
2: Draw $U_0 \sim \mathrm{Unif}(0,1)$ independently of all data.
3: For a deployment covariate x and candidate image y, augment the calibration fold by (x, y) and form $\mathcal{Q}_{x,y}$.
4: Solve the augmented moment equation (S.28) for $\pi_{x,y}$ using the lexicographic feasible selector.
5: Compute $\widehat{a}^{+}_{\mathrm{fd}}(x, y)$ from (S.29).
6: Output $\widehat{\mathcal{C}}^{+}_{\mathrm{fd}}(x; U_0) = \{y : \widehat{a}^{+}_{\mathrm{fd}}(x, y) \le U_0\}$.

Proposition S13. *For a fixed candidate (x, y) such that all augmented scores $\{R_i : i \in \mathcal{I}_{\mathrm{cal}}\} \cup \{R_\star(x, y)\}$ are finite real numbers, the augmented linear system is feasible if and only if $\alpha \bar{B}_x \in \mathrm{conv}\{V_{x,y}(q) : q \in \mathcal{Q}_{x,y}\}$.*

The condition always holds because the lower and upper sentinels are included in the definition of $\mathcal{Q}_{x,y}$.

Corollary S2. *For every candidate (x, y) satisfying the finite-score condition in Proposition S13, there exists a feasible solution $\pi_{x,y}$ supported on at most $d + 2$ thresholds.*

Corollary S2 bounds the support of an available feasible mixture in terms of the basis dimension. It does not require the chosen feasible selector to attain that support bound.

Proposition S14. *For a fixed candidate (x, y) satisfying the finite-score condition in Proposition S13, we define*

$$\widehat{a}_j(x, y) = \sum_{q \in \mathcal{Q}_{x,y}} \pi_{x,y}(q) \mathbf{1}\{R_j > q\}, \qquad j \in \mathcal{I}_{\mathrm{cal}} \cup \{\star\},$$

where $X_\star = x$ and $R_\star = R_n(x, y)$. Then for every $f \in \mathcal{F}_d$,

$$\frac{1}{N} \sum_{j \in \mathcal{I}_{\mathrm{cal}} \cup \{\star\}} f(X_j)\{\widehat{a}_j(x, y) - \alpha\} = 0. \tag{S.42}$$

Proposition S14 converts balance of the basis moments into balance for every shift in their linear span. This identity is the finite-sample step used in Theorem S7.

Proposition S15. *For every fixed (x, y) satisfying the finite-score condition in Proposition S13, let $\mathcal{G}_{\mathrm{sc}}$ denote the sigma-field generated by the score-construction randomness used by R_n. Conditional on $(\mathcal{D}_{\mathrm{tr}}, \mathcal{D}_{\mathrm{cal}}, x, y)$ and $\mathcal{G}_{\mathrm{sc}}$, it holds that*

$$\mathbb{E}_{U_0}\left[\mathbf{1}\{y \notin \widehat{\mathcal{C}}^{+}_{\mathrm{fd}}(x; U_0)\} \mid \mathcal{D}_{\mathrm{tr}}, \mathcal{D}_{\mathrm{cal}}, x, y, \mathcal{G}_{\mathrm{sc}}\right] = \widehat{a}^{+}_{\mathrm{fd}}(x, y). \tag{S.43}$$

Proposition S15 identifies $1 - \widehat{a}^{+}_{\mathrm{fd}}(x, y)$ as the conditional probability that the randomized set includes the candidate image, after averaging over the auxiliary uniform draw. The construction is candidate-dependent and does not return a single score threshold shared by all candidates. This distinguishes it from a static threshold-mixture solver.

Corollary S3. *Let $G_1, \ldots, G_L \subseteq \mathcal{X}$ be measurable subgroups, possibly overlapping, such that $P_{\mathrm{src},X}(G_\ell) > 0$ for every ℓ. Suppose the normalized indicators $f_\ell(x) = \mathbf{1}\{x \in G_\ell\}/P_{\mathrm{src},X}(G_\ell)$ for $\ell = 1, \ldots, L$ belong to $\mathcal{F}_d$. We fix the same pre-calibration randomness as in Theorem S7. Then, conditional on the training fold, for every $\ell = 1, \ldots, L$, it holds that*

$$\mathbb{E}\left[\mathbf{1}\{X^* \in G_\ell\}\left\{\mathbf{1}\{Y^* \notin \widehat{\mathcal{C}}^{+}_{\mathrm{fd}}(X^*; U_0)\} - \alpha\right\} \mid \mathcal{D}_{\mathrm{tr}}\right] = 0, \qquad (X^*, Y^*) \sim P_{\mathrm{src}}, \tag{S.44}$$

and equivalently

$$\mathbb{P}\{Y^* \in \widehat{\mathcal{C}}^{+}_{\mathrm{fd}}(X^*; U_0) \mid X^* \in G_\ell, \mathcal{D}_{\mathrm{tr}}\} = 1 - \alpha. \tag{S.45}$$

Corollary S3 protects each specified subgroup even when the groups overlap, provided their normalized indicators belong to the shift class. Here the probability in (S.45) averages over the calibration fold, the subgroup conditional target draw, and U_0.

The candidate-augmented solver runs one linear program per candidate pair (x, y). The following static variant solves a single moment system on the calibration fold and deploys one mixture of

thresholds at every covariate value. We let $\mathcal{Q}_{\mathrm{cal}}$ denote a finite ordered threshold-representative grid containing a lower sentinel $q_- < \min_{i\in\mathcal{I}_{\mathrm{cal}}} R_i$, an upper sentinel $q_+ \geq \max_{i\in\mathcal{I}_{\mathrm{cal}}} R_i$, and one representative in each open interval between consecutive distinct calibration scores, and we define

$$V_{\mathrm{cal}}(q) = \frac{1}{n_{\mathrm{cal}}} \sum_{i\in\mathcal{I}_{\mathrm{cal}}} B_i \mathbf{1}\{R_i > q\}, \qquad \bar{B}_{\mathrm{cal}} = \frac{1}{n_{\mathrm{cal}}} \sum_{i\in\mathcal{I}_{\mathrm{cal}}} B_i.$$

Proposition S16. *Suppose the calibration scores $\{R_i : i \in \mathcal{I}_{\mathrm{cal}}\}$ are finite real numbers. Then the moment system*

$$\sum_{q\in\mathcal{Q}_{\mathrm{cal}}} \pi(q) V_{\mathrm{cal}}(q) = \alpha \bar{B}_{\mathrm{cal}}, \qquad \pi \in \Delta(\mathcal{Q}_{\mathrm{cal}}), \tag{S.46}$$

admits a solution. Let $\widehat{\pi}$ be a measurable feasible selection. We sample $q^\star$ conditionally from $\widehat{\pi}$ using an independent uniform variable, and we define $\widehat{\mathcal{C}}_{\mathrm{mix}}(x; q^\star) = \{y \in \mathcal{Y} : R_n(x, y) \leq q^\star\}$. Suppose $\mathcal{F}_d$ is nonempty, Assumption S3 holds, and the score is fixed independently of calibration. Then, for every $0 < \delta < 1$, with probability at least $1 - \delta$,

$$\sup_{f\in\mathcal{F}_d} \left| P_f\{Y^* \in \widehat{\mathcal{C}}_{\mathrm{mix}}(X^*; q^\star) \mid \mathcal{D}_{\mathrm{cal}}\} - (1-\alpha) \right| \leq \psi_n(\delta, \mathcal{F}_d), \tag{S.47}$$

where the conditional probability averages over $q^\star$ and an independent target draw $(X^*, Y^*) \sim P_f$, holding the training fold and the score-construction randomness fixed.

Proposition S16 gives a two-sided coverage-deviation bound without Condition S1 and without the local quantile regularity of Assumption S1. The mixture balances the empirical moment equations exactly in every direction of the basis B. A single deterministic threshold cannot in general balance several directions. The candidate-augmented solver of Theorem S7 removes the $\psi_n(\delta, \mathcal{F}_d)$ term entirely. Appendix S8.4 describes the linear-programming implementation of (S.46) used in the numerical analyses.

S6.2 Semiparametric estimation of shift weights

For calibration with a general estimated shift, a density ratio $r(x) = \frac{dP_{\mathrm{tar},X}}{dP_{\mathrm{src},X}}(x)$ can be estimated from source and target covariates alone (Shimodaira, 2000; Kanamori et al., 2009; Sugiyama et al., 2012). The medical comparison uses empirical group-frequency ratios. The analysis below concerns a separate exponential-tilt estimator. We let $\Theta_{\mathrm{tilt}} \subseteq \mathbb{R}^d$ denote its parameter set and consider

$$r_\theta(x) = \exp\{\theta^\top \phi(x) - \Lambda(\theta)\}, \qquad \Lambda(\theta) = \log \mathbb{E}_{P_{\mathrm{src},X}}[\exp\{\theta^\top \phi(X)\}]. \tag{S.48}$$

Given source covariates $\{X_i\}_{i=1}^{n_{\mathrm{src}}}$ and unlabeled target covariates $\{X_j^{\mathrm{tar}}\}_{j=1}^{n_{\mathrm{tar}}}$, we define the empirical estimator

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta\in\Theta_{\mathrm{tilt}}} \left[\log\left\{ \frac{1}{n_{\mathrm{src}}} \sum_{i=1}^{n_{\mathrm{src}}} e^{\theta^\top \phi(X_i)} \right\} - \frac{1}{n_{\mathrm{tar}}} \sum_{j=1}^{n_{\mathrm{tar}}} \theta^\top \phi(X_j^{\mathrm{tar}}) \right]. \tag{S.49}$$

We let $\widehat{\Lambda}(\theta) = \log\{n_{\mathrm{src}}^{-1} \sum_{i=1}^{n_{\mathrm{src}}} e^{\theta^\top \phi(X_i)}\}$ denote the empirical normalizing constant, and we define $\widehat{r}(x) = \exp\{\widehat{\theta}^\top \phi(x) - \widehat{\Lambda}(\widehat{\theta})\}$. We use $\|\cdot\|$ for the Euclidean norm, $B(\theta, r) = \{u : \|u - \theta\| < r\}$ for the open ball, and $\overline{B}(\theta, r)$ for its closure. The diameter is $\mathrm{diam}(D) = \sup_{u,v\in D} \|u - v\|$. The following proposition gives a high probability rate for $\widehat{\theta}$.

Proposition S17. *Suppose the source covariates $\{X_i\}_{i=1}^{n_{\mathrm{src}}}$ and target covariates $\{X_j^{\mathrm{tar}}\}_{j=1}^{n_{\mathrm{tar}}}$ are independent, and each set is drawn i.i.d. from $P_{\mathrm{src},X}$ and $P_{\mathrm{tar},X}$, respectively. Suppose Θ_{tilt} is compact, $\theta_0 \in \mathrm{int}(\Theta_{\mathrm{tilt}})$ uniquely minimizes $M(\theta) = \Lambda(\theta) - \mathbb{E}_{P_{\mathrm{tar},X}}\{\theta^\top \phi(X)\}$, the feature map ϕ is bounded, and $\nabla^2 M(\theta_0)$ is positive definite. We fix $r_{\mathrm{loc}} > 0$ such that $\overline{B}(\theta_0, r_{\mathrm{loc}}) \subset \Theta_{\mathrm{tilt}}$ and M is m_{loc}-strongly convex on this ball, and define $\eta_{\mathrm{loc}} = \inf_{\theta\in\Theta_{\mathrm{tilt}}:\|\theta-\theta_0\|\geq r_{\mathrm{loc}}} \{M(\theta) - M(\theta_0)\}$. Then $\eta_{\mathrm{loc}} > 0$, and, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\|\widehat{\theta} - \theta_0\| \leq C\left\{ \sqrt{\frac{1 + \log(1/\delta)}{n_{\mathrm{src}}}} + \sqrt{\frac{1 + \log(1/\delta)}{n_{\mathrm{tar}}}} \right\},$$

where C depends only on a uniform bound for ϕ, $\sup_{\theta\in\Theta_{\mathrm{tilt}}} \|\theta\|$, $\mathrm{diam}(\Theta_{\mathrm{tilt}})$, m_{loc}, r_{loc}, η_{loc}, and a fixed Euclidean metric-entropy bound for Θ_{tilt}, and, for a fixed d-dimensional compact

parametrization, depends on d and on the covering regularity of Θ_{tilt}. If, in addition, $\widehat{\theta} \in \mathrm{int}(\Theta_{\mathrm{tilt}})$, then

$$\frac{1}{n_{\mathrm{src}}} \sum_{i=1}^{n_{\mathrm{src}}} \widehat{r}(X_i)\phi(X_i) - \frac{1}{n_{\mathrm{tar}}} \sum_{j=1}^{n_{\mathrm{tar}}} \phi(X_j^{\mathrm{tar}}) = 0.$$

Proposition S17 separates the estimation errors from source and target covariates. Its constants depend on the fixed parameter dimension and population curvature. It is not a dimension-free density-ratio bound.

S6.3 Radius bounds for score balls

The radius properties in Propositions S11 and S12 extend to any score whose level sets are centered $\mathcal{H}$-balls with the exact radius property stated below. For $y_0 \in \mathcal{Y}$ and $r \geq 0$, we define the centered $\mathcal{H}$-ball

$$\mathbb{B}_{\mathcal{H}}(y_0, r) = \left\{ y \in \mathcal{Y} : \sup_{h \in \mathcal{H}} |h(y) - h(y_0)| \leq r \right\}.$$

The following lemma presents a sufficient condition for the radial bound and radial Lipschitz properties in Propositions S11 and S12. It applies when each predictor-informed score in (S.26) has a single-center representation. This representation is a structural condition on the score. It does not follow from calibration.

Lemma S2. *Suppose there exists a measurable bound $A : \mathcal{X} \to [0, \infty)$ such that $\sup_{f \in \mathcal{F}} \mathbb{E}_f\{A(X)\} \leq A_{\mathcal{F}} < \infty$. Let $\mathcal{R}$ be a collection of scores on $\mathcal{X} \times \mathcal{Y}$. Suppose that, for each $S \in \mathcal{R}$, there exist measurable maps $y_S : \mathcal{X} \to \mathcal{Y}$ and $a_S : \mathcal{X} \to [0, \infty)$ such that $a_S(x) \leq A(x)$ and $\mathcal{C}_t^S(x) = \mathbb{B}_{\mathcal{H}}\{y_S(x), a_S(x)t\}$ for every $x \in \mathcal{X}$ and every $t \geq 0$. Suppose, in addition, that $\mathrm{rad}_{\mathcal{H}}\{\mathbb{B}_{\mathcal{H}}(y_0, r)\} = r$ for every $y_0 \in \mathcal{Y}$ and every $r \geq 0$. Then, for every $S \in \mathcal{R}$, $\mathrm{rad}_{\mathcal{H}}\{\mathcal{C}_t^S(x)\} \leq A(x)t$ and $\left|\mathrm{rad}_{\mathcal{H}}\{\mathcal{C}_{t_2}^S(x)\} - \mathrm{rad}_{\mathcal{H}}\{\mathcal{C}_{t_1}^S(x)\}\right| \leq A(x)\,|t_2 - t_1|$ for every $x \in \mathcal{X}$ and every $t, t_1, t_2 \geq 0$. In particular, if $\mathcal{R} = \{R_n\}$, then the radial bound and radial Lipschitz properties in Propositions S11 and S12 hold for R_n. If $\mathcal{R} = \{R_n, R^\circ\}$, then the shared radial Lipschitz property used in Theorem S22 also holds with the same bound A.*

The equality $\mathrm{rad}_{\mathcal{H}}\{\mathbb{B}_{\mathcal{H}}(y_0, r)\} = r$ in Lemma S2 is a structural condition on $(\mathcal{Y}, \mathcal{H})$. The same two conclusions hold without this equality if, for every S, x, y and $s \in [0, 1]$, there exists $y_s \in \mathcal{Y}$ with $h(y_s) = h\{y_S(x)\} + s[h(y) - h\{y_S(x)\}]$ for every $h \in \mathcal{H}$. Linear functionals on a convex image class satisfy this condition when $y_S(x) \in \mathcal{Y}$.

The center $y_S(x)$ gives $\mathrm{rad}_{\mathcal{H}}\{\mathcal{C}_t^S(x)\} \leq a_S(x)t$. For $0 \leq t_1 < t_2$, we contract by $s = t_1/t_2$. Each $y \in \mathcal{C}_{t_2}^S(x)$ gives $y_s \in \mathcal{C}_{t_1}^S(x)$ with $\sup_h |h(y) - h(y_s)| \leq a_S(x)(t_2 - t_1)$. At an ε-optimal center of the smaller set, the triangle inequality yields $\mathrm{rad}_{\mathcal{H}}\{\mathcal{C}_{t_2}^S(x)\} \leq \mathrm{rad}_{\mathcal{H}}\{\mathcal{C}_{t_1}^S(x)\} + \varepsilon + a_S(x)(t_2 - t_1)$. We let $\varepsilon \downarrow 0$ and use set inclusion. Equality and symmetry handle the remaining thresholds, including $a_S(x) = 0$. For $\mathcal{Y} = [0, 1]$, $h(y) = y$, and center $1/2$, the radius is $\min(r, 1/2)$. Proposition S18 allows different scales across a finite functional class.

Example S1. Let $\mathcal{Y} = \{0, 1\}$, let $\mathcal{H}_q = \{h\}$ defined by $h(y) = y$, and consider the normalized slab score $S(y) = |h(y) - 1/2|$. Then

$$\mathcal{C}_t^S = \{y \in \mathcal{Y} : |y - 1/2| \leq t\} = \begin{cases} \emptyset, & t < 1/2, \\ \{0, 1\}, & t \geq 1/2. \end{cases}$$

Because $\mathrm{rad}_{\mathcal{H}_q}(\emptyset) = 0$ and $\mathrm{rad}_{\mathcal{H}_q}(\{0, 1\}) = 1$, the map $t \mapsto \mathrm{rad}_{\mathcal{H}_q}(\mathcal{C}_t^S)$ has a jump at $t = 1/2$. This shows that finite intersections of normalized slabs do not, by themselves, imply the radial Lipschitz property in Proposition S12.

The next proposition avoids this discontinuity by requiring that feasible feature vectors can be contracted toward a common center.

For a finite screened class $\mathcal{H}_q = \{h_1, \ldots, h_J\}$, we let $T_q : \mathcal{Y} \to \mathbb{R}^J$ denote the coordinate map $T_q(y) = (h_1(y), \ldots, h_J(y))$, and let $d_q(y, y') = \|T_q(y) - T_q(y')\|_\infty$. For measurable functions $m_j : \mathcal{X} \to \mathbb{R}$ and $a_j : \mathcal{X} \to (0, \infty)$, we consider the normalized slab score

$$S_q(x, y) = \max_{1 \leq j \leq J} \frac{|h_j(y) - m_j(x)|}{a_j(x)}, \qquad \mathcal{C}_t^{S_q}(x) = \{y \in \mathcal{Y} : S_q(x, y) \leq t\},$$

and we define $a_\vee(x) = \max_{1\le j\le J} a_j(x)$.

Proposition S18. *Suppose that, for every $x \in \mathcal{X}$, there exists $y_c(x) \in \mathcal{Y}$ such that $T_q\{y_c(x)\} = (m_1(x), \dots, m_J(x))$ and, for every $z \in T_q(\mathcal{Y})$ and every $s \in [0,1]$, the point $T_q\{y_c(x)\} + s[z - T_q\{y_c(x)\}]$ belongs to $T_q(\mathcal{Y})$. Then, for every $x \in \mathcal{X}$ and every $t, t_1, t_2 \ge 0$, $\mathrm{rad}_{\mathcal{H}_q}\{\mathcal{C}_t^{S_q}(x)\} \le a_\vee(x)t$ and $\left|\mathrm{rad}_{\mathcal{H}_q}\{\mathcal{C}_{t_2}^{S_q}(x)\} - \mathrm{rad}_{\mathcal{H}_q}\{\mathcal{C}_{t_1}^{S_q}(x)\}\right| \le a_\vee(x)\,|t_2 - t_1|$. If, in addition, $\sup_{f\in\mathcal{F}} \mathbb{E}_f\{a_\vee(X)\} < \infty$, then the screened analogues of the radial bound and radial Lipschitz properties in Propositions S11 and S12 hold for S_q with $\mathcal{H}$ replaced by $\mathcal{H}_q$ and with bound $a_\vee$.*

For the regional-average, sparse-contrast, and multiscale wavelet classes, every functional h_j is linear in the image. If $\mathcal{Y}$ is a convex image space, such as a box constraint in $\mathbb{R}^V$ or all of $\mathbb{R}^V$, and the fitted center image $y_c(x)$ belongs to $\mathcal{Y}$ and satisfies $m_j(x) = h_j\{y_c(x)\}$, then $T_q(\mathcal{Y})$ is star-shaped about $T_q\{y_c(x)\}$. Proposition S18 then verifies the radial properties for the fully normalized finite screened slab score. It does not impose them.

S6.4 Multiresolution image regularity

The two bounds in Theorem S12 depend on different image properties. Haar coefficient decay controls the fine-scale contribution to the $\mathcal{H}^{\mathrm{MR}}$-radius through (S.24). On the full image box, omitted Haar coefficients instead give a lower bound through images with identical included coefficients. Both use the two-dimensional Haar system and the classes of Definitions S3 and S4.

The following lemma bounds the $\mathcal{H}^{\mathrm{MR}}$-radius by the radius of its level-J' truncation plus a deterministic fine-scale remainder.

Lemma S3. *Assume $\mathcal{Y} = \mathcal{Y}_R$ as defined in (S.24). Then, for every set $C \subseteq \mathcal{Y}_R$ and every $0 \le J' \le J$,*

$$\mathrm{rad}_{\mathcal{H}^{\mathrm{MR}}}(C) \le \mathrm{rad}_{\mathcal{H}^{\mathrm{MR}}_{\le J'}}(C) + 2R\,2^{-sJ'}. \tag{S.50}$$

The remainder is controlled by the candidate-image restriction and does not require calibration of the omitted fine-scale features.

Proof. For $C = \emptyset$, both radii are zero. We may thus assume $C \neq \emptyset$. We fix $\eta > 0$ and choose an η-optimal center $y_0^\star \in \mathcal{Y}_R$ for $\mathrm{rad}_{\mathcal{H}^{\mathrm{MR}}_{\le J'}}(C)$. For $h = h_{j,k}$ with $j > J'$ and $y \in C \subseteq \mathcal{Y}_R$, (S.24) gives $|h(y)|, |h(y_0^\star)| \le R2^{-sj}$, hence $|h(y) - h(y_0^\star)| \le 2R2^{-sj} \le 2R2^{-sJ'}$. We split $\mathcal{H}^{\mathrm{MR}}$ at level J' and use $\max(a,b) \le a + b$ for $a, b \ge 0$:

$$\mathrm{rad}_{\mathcal{H}^{\mathrm{MR}}}(C) \le \sup_{y\in C}\sup_{h\in\mathcal{H}^{\mathrm{MR}}} |h(y) - h(y_0^\star)| \le \sup_{y\in C}\sup_{h\in\mathcal{H}^{\mathrm{MR}}_{\le J'}} |h(y) - h(y_0^\star)| + 2R\,2^{-sJ'},$$

where the first term is at most $\mathrm{rad}_{\mathcal{H}^{\mathrm{MR}}_{\le J'}}(C) + \eta$. We let $\eta \downarrow 0$ to obtain (S.50). This completes the proof. □

The next proposition bounds the spatial differences that remain when a reduction uses only a strict subset of the Haar detail functionals. We restrict to reductions whose included functionals are detail functionals to ensure that the omitted level is an exact coordinate in the wavelet system. This gives the resolution lower bound in (S.41). Recall $\Delta_S(t)$ from Definition S5.

Proposition S19. *Let $\mathcal{Y} = [0,1]^V$ and let $\mathcal{H}^{\mathrm{MR}}$ be scale complete. Let $\mathcal{H}_q$ be a set of M Haar detail functionals with $1 \le M \le V - 2$ that omits at least one of the $V - 1$ detail functionals, and let $S = (h)_{h\in\mathcal{H}_q}$ denote the associated reduction. Fix $\rho_0 \in (0, 1/2]$ and define $c_\star = \rho_0/\sqrt{2}$. For every $t \in S(\mathcal{Y})$ that admits a preimage with coordinates in $[\rho_0, 1-\rho_0]^V$,*

$$\Delta_S(t) \ge c_\star\sqrt{V/M}. \tag{S.51}$$

Proof. Let j_0 be the coarsest omitted detail level and ψ_{j_0,k_0} an omitted functional. All $\sum_{j=1}^{j_0-1} 3\cdot 4^{j-1} = 4^{j_0-1} - 1$ coarser details lie in $\mathcal{H}_q$, hence $4^{j_0-1} - 1 \le M$, $4^{j_0-1} \le M + 1 \le 2M$, and $2^{j_0} \le 2\sqrt{2M}$. Definition S3 gives $\|\psi_{j_0,k_0}\|_\infty = 2^{j_0-1}/\sqrt{V}$. We choose a preimage $y^\dagger \in$

$[\rho_0, 1-\rho_0]^V$ of t, define $r = 2\rho_0\sqrt{V}/2^{j_0}$, and define $y^{\pm} = y^{\dagger} \pm \frac{r}{2}\psi_{j_0,k_0}$. The pixel perturbations are $\frac{r}{2}\|\psi_{j_0,k_0}\|_\infty = \rho_0/2 \le \rho_0$, hence $y^{\pm} \in [0,1]^V$. Orthonormality gives $h(y^+) = h(y^-)$ for every $h \in \mathcal{H}_q$, and thus $S(y^+) = S(y^-) = t$. Finally,

$$\sup_{h \in \mathcal{H}^{\mathrm{MR}}} \left|h(y^+) - h(y^-)\right| \ge \left|h_{j_0,k_0}(y^+) - h_{j_0,k_0}(y^-)\right| = r = 2\rho_0\sqrt{V}/2^{j_0} \ge \frac{\rho_0}{\sqrt{2}}\sqrt{V/M},$$

using $2^{j_0} \le 2\sqrt{2M}$. This gives (S.51). This completes the proof. □

The bound in (S.51) concerns coordinate selection in the Haar system. It is not asserted for arbitrary linear maps in the coefficientwise radius. Theorem 3 gives a separate, explicit lower bound for arbitrary linear maps in the normalized image norm.

S6.5 Uniform score and calibration bounds

Theorems S8–S10 require separate bounds for score approximation and calibration error. The following results establish these bounds from population conditions. The first controls the full predictor-informed score $R_n = \sup_{h\in\mathcal{H}} S_{n,h}$ uniformly over $\mathcal{X} \times \mathcal{Y}$, and the second controls the calibration empirical process for the level sets of this max-score.

We use $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For a class $\mathcal{A}$ of bounded real-valued functions, we let $N_\infty(\varepsilon, \mathcal{A})$ denote the smallest cardinality of a uniform ε-net. Such a net approximates every function in the class to the stated accuracy uniformly over its argument. For the maximum-score construction, we let $K = K_n$ and let $\widetilde{Z}_k : \mathcal{X} \to \mathcal{Y}$, $k = 1, \dots, K$, be the teacher random maps used to compute the plug-in moments $\widehat{\mu}_h$ and $\widehat{\sigma}_h$ in R_n. The construction used in Theorem S13 takes these maps to be i.i.d. with measurable sample paths, independent of $\mathcal{D}_{\mathrm{cal}}$, and with marginal distribution Q_x^{T} at x.

We define the oracle and teacher moments

$$\mu_h^{\mathrm{T}}(x) = \mathbb{E}_{Q_x^{\mathrm{T}}}\{h(Y)\}, \qquad m_{2,h}^{\mathrm{T}}(x) = \mathbb{E}_{Q_x^{\mathrm{T}}}\{h(Y)^2\}, \qquad m_{2,h}^{\circ}(x) = \mathbb{E}\{h(Y)^2 \mid X = x\}.$$

For $z : \mathcal{X} \to \mathcal{Y}$, we define the composite function classes

$$\mathcal{A}_{\mu,n} = \{z \mapsto h\{z(x)\} : x \in \mathcal{X}, h \in \mathcal{H}\}, \qquad \mathcal{A}_{2,n} = \{z \mapsto h\{z(x)\}^2 : x \in \mathcal{X}, h \in \mathcal{H}\}.$$

Assumption S2 (Primitive full-class score regularity)**.** There exist $B_{\mathcal{H}} > 0$, $A_{\mathcal{H}} \ge 2$, $v_{\mathcal{H}} \ge 0$, and deterministic $b_{\mu,n}, b_{2,n} \ge 0$ such that

$$\sup_{h\in\mathcal{H}}\sup_{y\in\mathcal{Y}}|h(y)| \le B_{\mathcal{H}}, \quad \sup_{x,h}|\mu_h^{\mathrm{T}}(x) - \mu_h^{\circ}(x)| \le b_{\mu,n}, \quad \sup_{x,h}|m_{2,h}^{\mathrm{T}}(x) - m_{2,h}^{\circ}(x)| \le b_{2,n}. \tag{S.52}$$

The class $\mathcal{A}_{\mu,n}$ is pointwise measurable and satisfies $N_\infty(\eta B_{\mathcal{H}}, \mathcal{A}_{\mu,n}) \le (A_{\mathcal{H}}/\eta)^{v_{\mathcal{H}}}$ for every $0 < \eta \le 1$. In addition, $\mathcal{H}$ contains a countable $\|\cdot\|_\infty$-dense subclass.

Assumption S2 separates population predictor bias from the complexity needed for uniform estimation of its moments. Neither term is a condition on realized calibration scores. For $|a|, |b| \le B_{\mathcal{H}}$, the inequality $|a^2 - b^2| \le 2B_{\mathcal{H}}|a - b|$ implies $N_\infty(\eta B_{\mathcal{H}}^2, \mathcal{A}_{2,n}) \le N_\infty(\eta B_{\mathcal{H}}/2, \mathcal{A}_{\mu,n})$. Squaring a countable pointwise dense subclass proves pointwise measurability of $\mathcal{A}_{2,n}$. Increasing $A_{\mathcal{H}}$ by at most a factor of two gives

$$N_\infty(\eta B_{\mathcal{H}}, \mathcal{A}_{\mu,n}) \vee N_\infty(\eta B_{\mathcal{H}}^2, \mathcal{A}_{2,n}) \le (A_{\mathcal{H}}/\eta)^{v_{\mathcal{H}}}.$$

Assumption S3 (Primitive shift entropy)**.** The class $\mathcal{F}$ is pointwise measurable, nonnegative, and uniformly bounded by B_F. There exist $A_F \ge 2$ and $v_F < \infty$ such that, for every probability measure Q on $\mathcal{X}$ and every $0 < \eta \le 1$,

$$N\{\eta B_F, \mathcal{F}, L_2(Q)\} \le (A_F/\eta)^{v_F}.$$

Assumption S3 bounds the number of functions needed to approximate the protected shift class at each accuracy. This approximation bound controls calibration error after the image score has been fixed.

For $0 < \delta < 1$, we define

$$\begin{aligned}
\eta_{\mu,n}(\delta) &= CB_{\mathcal{H}}\sqrt{\frac{1 + v_{\mathcal{H}}\log(A_{\mathcal{H}}K) + \log(4/\delta)}{K}},\\
\eta_{2,n}(\delta) &= CB_{\mathcal{H}}^2\sqrt{\frac{1 + v_{\mathcal{H}}\log(A_{\mathcal{H}}K) + \log(4/\delta)}{K}},\\
\eta_{\sigma,n}(\delta) &= \{b_{2,n} + \eta_{2,n}(\delta) + 2B_{\mathcal{H}}[b_{\mu,n} + \eta_{\mu,n}(\delta)]\}^{1/2},\\
e_n(\delta) &= \lambda^{-1}\{b_{\mu,n} + \eta_{\mu,n}(\delta)\} + 2B_{\mathcal{H}}\lambda^{-2}\eta_{\sigma,n}(\delta),\\
\psi_n(\delta, \mathcal{F}) &= CB_F\sqrt{\frac{1 + (v_F + 2)\log(A_F n_{\mathrm{cal}}) + \log(1/\delta)}{n_{\mathrm{cal}}}},
\end{aligned} \tag{S.53}$$

where C is a universal constant.

The following theorem establishes a high probability supremum-norm bound on the gap between the predictor-informed score R_n and the oracle score R°, which verifies the score perturbation bound (S.33).

Theorem S13. *Suppose that the teacher random maps in the primitive max-score construction are independent of the calibration fold and Assumption S2 holds. Then R_n and R° are measurable, and, for every $0 < \delta < 1$,*

$$\mathbb{P}\left\{\sup_{x\in\mathcal{X}}\sup_{y\in\mathcal{Y}}|R_n(x,y) - R^\circ(x,y)| \le e_n(\delta)\right\} \ge 1 - \delta. \tag{S.54}$$

Theorem S13 verifies the score perturbation bound (S.33) with $e_n(\delta)$ in (S.53). The bound separates errors in the predictor population moments from errors due to finitely many queries.

Concentration of empirical standard deviations for bounded observations is established by Maurer & Pontil (2009, Theorem 10). The standard-deviation estimate admits a sharper bound than the second-moment square-root argument in (S.53). The same mean-function entropy condition suffices for this bound.

We assume the boundedness, population bias, and countable dense-subclass conditions in Assumption S2, and the i.i.d. teacher random maps defined above. We require only that $\mathcal{A}_{\mu,n}$ is pointwise measurable and

$$N_\infty(\eta B_{\mathcal{H}}, \mathcal{A}_{\mu,n}) \le (A_{\mathcal{H}}/\eta)^{v_{\mathcal{H}}}, \qquad 0 < \eta \le 1,$$

where $A_{\mathcal{H}} \ge 2$ and $v_{\mathcal{H}} \ge 0$. Let $C_{B_{\mathcal{H}},\lambda}$ denote a positive constant depending only on $B_{\mathcal{H}}$ and λ.

Proposition S20. *Suppose the stated boundedness, bias, and entropy conditions hold. Then, for $K \ge 1$ and $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$\begin{aligned}
\sup_{x,y}|R_n(x,y) - R^\circ(x,y)| \le{}& \frac{b_{\mu,n}}{\lambda} + \frac{2B_{\mathcal{H}}}{\lambda^2}\sqrt{b_{2,n} + 2B_{\mathcal{H}}b_{\mu,n}}\\
&+ C_{B_{\mathcal{H}},\lambda}\sqrt{\frac{1 + v_{\mathcal{H}}\log(A_{\mathcal{H}}K) + \log(4/\delta)}{K}}.
\end{aligned} \tag{S.55}$$

For matching population moments, (S.55) has a $K^{-1/2}$ query term up to the entropy factor, including zero variances. The bound replaces $e_n(\delta)$ in (S.33) and in every subsequent result that uses that perturbation inequality. It does not remove the population predictor-bias terms.

Proof of Proposition S20. We choose a $B_{\mathcal{H}}/K$-net $\mathcal{N} \subset \mathcal{A}_{\mu,n}$ with $|\mathcal{N}| \le (A_{\mathcal{H}}K)^{v_{\mathcal{H}}}$, enlarging the covering constant by a universal factor if necessary. For each $g \in \mathcal{N}$, the $g(\widetilde{Z}_k)$ are i.i.d. and bounded by $B_{\mathcal{H}}$. Lemma S6 and a union bound give simultaneous mean and standard-deviation errors of size

$$CB_{\mathcal{H}}\sqrt{\frac{1 + v_{\mathcal{H}}\log(A_{\mathcal{H}}K) + \log(4/\delta)}{K}}.$$

Both the population and empirical mean maps are 1-Lipschitz in the uniform norm. The same holds for standard deviations because

$$|\,\mathrm{sd}_P(g) - \mathrm{sd}_P(g')| \le \|g - g'\|_{L_2(P)} \le \|g - g'\|_\infty,$$

by the distance-to-constants representation, also for the empirical measure. Extension from the net adds at most $2B_{\mathcal{H}}/K$ to each error, giving the rate uniformly over x, h. The population bias bounds give

$$|\widetilde{\sigma}_h(x) - \sigma_h^\circ(x)| \le \sqrt{|\widetilde{\sigma}_h(x)^2 - \sigma_h^\circ(x)^2|} \le \sqrt{b_{2,n} + 2B_{\mathcal{H}} b_{\mu,n}}.$$

Proposition S10 proves (S.55). This completes the proof. □

The following theorem establishes a uniform empirical-process bound on $\Gamma_n^+(\cdot;\mathcal{F})$ in terms of the metric entropy and uniform bound of $\mathcal{F}$, giving the calibration empirical-process bound (S.34) with rate $\psi_n(\delta,\mathcal{F})$. We write δ_z for the point mass at z. For a probability measure P and an integrable function g, we write $Pg = \int g\, dP$.

Theorem S14. *Suppose Assumption S3 holds. Let $R : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be measurable. Suppose that, conditional on R, the calibration observations $\{(X_i, Y_i)\}_{i\in\mathcal{I}_{\mathrm{cal}}}$ remain i.i.d. from the source distribution. Let*

$$P_n = n_{\mathrm{cal}}^{-1} \sum_{i\in\mathcal{I}_{\mathrm{cal}}} \delta_{(X_i,Y_i)}, \qquad P = P_{\mathrm{src}},$$

where P denotes the source joint distribution of a fresh calibration observation, conditional on the score-construction randomness. Then, for every $0 < \delta < 1$,

$$\mathbb{P}\Big[\sup_{f\in\mathcal{F}} \sup_{q\in\mathbb{R}} |(P_n - P)\,[f(X)\{\mathbf{1}\{R(X,Y) > q\} - \alpha\}]| > \psi_n(\delta,\mathcal{F})\Big] \le \delta. \tag{S.56}$$

In particular, under Assumption S3, Theorem S14 applies to $R = R_n = \sup_{h\in\mathcal{H}} S_{n,h}$ from Theorem S13 when the teacher random maps used to construct R_n are independent of the calibration fold. This implies that the calibration empirical-process bound (S.34) holds for this choice of $\psi_n(\delta,\mathcal{F})$.

For a subclass $\mathcal{G} \subseteq \mathcal{H}$, let $B_{\mathcal{G}}, L_{\mathcal{G}} > 0$ be fixed constants. We write $B_{\mathcal{H}}, L_{\mathcal{H}}$ when $\mathcal{G} = \mathcal{H}$. In finite-screened statements, these symbols denote the constants for $\mathcal{G} = \mathcal{H}_q$ unless a separate subscript is used.

Assumption S4 (Uniform boundedness for an image-functional class)**.** It holds that $\sup_{h\in\mathcal{G}} \sup_{y\in\mathcal{Y}} |h(y)| \le B_{\mathcal{G}}$. For Wasserstein bounds, also assume

$$|h(y) - h(y')| \le L_{\mathcal{G}} d_{\mathcal{Y}}(y, y'), \qquad h \in \mathcal{G}, \quad y, y' \in \mathcal{Y}.$$

Assumption S4 bounds the magnitude of the measured image features. Its Lipschitz part converts an error in the predictor distribution into moment error through Lemma S14; the Kullback–Leibler comparison in Lemma S15 uses boundedness alone.

S6.6 Calibration rates and lower bounds

To make Theorem S9 concrete, we derive the bound $\psi_n(\delta,\mathcal{F})$ for three canonical shift classes.

Proposition S21. *Suppose R_n is measurable and, conditional on R_n, the calibration observations remain i.i.d. from the source distribution. If $\mathcal{F} = \mathcal{F}_d$ is the finite-dimensional linear class (S.22) and $\sup_{f\in\mathcal{F}_d} \|f\|_\infty \le B_d$, then $A_d = 3$ satisfies*

$$\sup_Q N\{\eta B_d, \mathcal{F}_d, L_2(Q)\} \le (A_d/\eta)^{d+1}, \qquad 0 < \eta \le 1,$$

and the calibration empirical-process bound (S.34) holds with

$$\psi_n(\delta,\mathcal{F}) = C B_d \sqrt{\frac{1 + (d+3)\log(A_d e n_{\mathrm{cal}}) + \log(1/\delta)}{n_{\mathrm{cal}}}}.$$

Proposition S21 derives the entropy bound from finite dimension and bounded weights. These two quantities determine the calibration error.

Proposition S22. *Suppose R_n is measurable and, conditional on R_n, the calibration observations remain i.i.d. from the source distribution. If $\mathcal{F}$ is pointwise measurable, has VC-subgraph dimension v, and is uniformly bounded by B_F, then the calibration empirical-process bound (S.34) holds with*

$$\psi_n(\delta,\mathcal{F}) = C B_F \sqrt{\frac{1 + (v+2)\log(e n_{\mathrm{cal}}) + \log(1/\delta)}{n_{\mathrm{cal}}}}.$$

Proposition S22 permits a nonlinear shift class whose threshold complexity is bounded. The rate depends on that complexity and not on a coordinate representation of the images.

Proposition S23. *Suppose R_n is measurable and, conditional on R_n, the calibration observations remain i.i.d. from the source distribution. If $\mathcal{F}$ is a pointwise measurable subset of the unit ball of an RKHS $\mathcal{H}_k$ with reproducing kernel k satisfying $\sup_x k(x,x) \le \kappa^2$, then the calibration empirical-process bound (S.34) holds with*

$$\psi_n(\delta, \mathcal{F}) = C\kappa\sqrt{\frac{1+\log(1/\delta)}{n_{\mathrm{cal}}}}.$$

Proposition S23 controls an infinite-dimensional shift class through its kernel bound. The norm restriction is essential to this conclusion.

The following theorem establishes a class-indexed minimax lower bound for the worst-case calibration error over the subset shift class $\mathcal{F}_{\mathrm{sub}}$. In this construction, q_0 denotes a fixed score threshold.

Theorem S15. *For every even $d \ge 2$, every $n \ge d$, and every fixed $\alpha \in (0,1)$, there exists a covariate shift calibration model with source covariate distribution uniform on $\{1,\dots,d\}$, a shift class $\mathcal{F}_{\mathrm{sub}} = \{2\mathbf{1}\{x \in A\} : A \subset \{1,\dots,d\}, |A| = d/2\}$ bounded by 2, and Bernoulli exceedance variables $Z = \mathbf{1}\{R^\circ > q_0\}$ such that the following holds. If $p_j = P(Z = 1 \mid X = j)$ and an estimator $\widehat{p} = (\widehat{p}_1, \dots, \widehat{p}_d)$ is constructed from n labeled source observations, then*

$$\inf_{\widehat{p}} \sup_{p \in [\alpha/2, 1-\alpha/2]^d} \mathbb{E}_p \sup_{f \in \mathcal{F}_{\mathrm{sub}}} \left| \frac{1}{d} \sum_{j=1}^{d} f(j)\{\widehat{p}_j - p_j\} \right| \ge c_\alpha \sqrt{\frac{d}{n}}, \tag{S.57}$$

where $c_\alpha > 0$ depends only on α.

Theorem S15 shows that no estimator improves the $\sqrt{d/n}$ scale in the finite-dimensional calibration bound uniformly over bounded d-cell subgroup shift classes, and implies that the bound $\psi_n(\delta, \mathcal{F})$ in Theorem S14 is minimax rate optimal over these classes in terms of the worst-case calibration error, up to logarithmic factors.

For prediction sets defined by score thresholds, Areces et al. (2024) establish a minimax lower bound of order $\sqrt{d/n}$ for weighted coverage error over a binary function class of VC dimension d, with coverage evaluated conditionally on the calibration sample at a fixed nominal level. The lower bound in Theorem S15 concerns the estimation of subgroup-weighted exceedance probabilities. The following proposition converts it into a lower bound for prediction sets: in a Gaussian-cell calibration model, every procedure that outputs cell-wise upper intervals has expected worst-case target coverage error of at least this order for some parameter value. The restriction to threshold rules excludes degenerate randomized sets whose conditional coverage equals $1-\alpha$ identically without using the data.

Proposition S24. *Let $\alpha \in (0,1)$. For every even $d \ge 2$ and every $n \ge d$, consider the covariate shift calibration model with X uniform on $\{1,\dots,d\}$, conditional response $Y \mid X = j \sim N(\theta_j, 1)$ for $\theta \in \Theta = [-1,1]^d$, and shift class $\mathcal{F}_{\mathrm{sub}}$ as in Theorem S15. Let $\widehat{\mathcal{C}}$ range over all prediction sets of the cell-wise threshold form $\widehat{\mathcal{C}}(j) = \{y \in \mathbb{R} : y \le \widehat{q}_j\}$, where $\widehat{q} = (\widehat{q}_1, \dots, \widehat{q}_d)$ is measurable with respect to n labeled source observations. Then*

$$\inf_{\widehat{q}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \sup_{f \in \mathcal{F}_{\mathrm{sub}}} \left| P_{f,\theta}\{Y \in \widehat{\mathcal{C}}(X)\} - (1-\alpha) \right| \ge c'_\alpha \sqrt{\frac{d}{n}}, \tag{S.58}$$

where $P_{f,\theta}$ denotes the shifted target distribution $dP_{f,\theta} = f\, dP_{\mathrm{src},\theta}$ evaluated conditionally on the source sample, and $c'_\alpha > 0$ depends only on α.

Proposition S24 concerns one-sided intervals in the stated Gaussian model, with coverage evaluated conditionally on the calibration sample. It does not establish an unrestricted lower bound for image prediction sets or for undercoverage alone. The distinction from the calibration-averaged guarantee of Theorem 1 is essential.

S6.7 Teacher quality in Wasserstein and Kullback–Leibler distance

Errors in the predictor distribution change the score through its conditional moments. The following population condition measures this distribution error before the finite-query approximation is considered.

Assumption S5 (Teacher distance on a query set)**.** For a specified query set $\mathcal{X}_0 \subseteq \mathcal{X}$, there exist $\varepsilon_W, \varepsilon_K \in [0, \infty]$, at least one of which is finite, such that

$$\sup_{x \in \mathcal{X}_0} W_1(Q_x^{\mathrm{T}}, P_{Y|X=x}) \le \varepsilon_W \qquad \text{and} \qquad \sup_{x \in \mathcal{X}_0} D_{\mathrm{KL}}(P_{Y|X=x} \| Q_x^{\mathrm{T}}) \le \varepsilon_K,$$

where W_1 is the 1-Wasserstein distance on $\mathcal{Y}$ with respect to $d_{\mathcal{Y}}$.

For the finite Monte Carlo term below, we let $\mathcal{X}_q \subset \mathcal{X}$ be a nonempty finite-query set and let $\mathcal{H}_q \subset \mathcal{H}$ be a nonempty finite screened subclass, and we define

$$R_{n,K}(x, y) = \sup_{h \in \mathcal{H}_q} \frac{|h(y) - \widehat{\mu}_h(x)|}{\widehat{\sigma}_h(x) + \lambda}, \qquad R_q^\circ(x, y) = \sup_{h \in \mathcal{H}_q} \frac{|h(y) - \mu_h^\circ(x)|}{\sigma_h^\circ(x) + \lambda}.$$

For these finite-screened statements, let $\mathrm{rad}_{\mathcal{H}_q}$ and $\mathcal{S}_{\mathcal{F},\mathcal{H}_q}$ be defined by replacing $\mathcal{H}$ with $\mathcal{H}_q$ in (S.23) and in the definition of $\mathcal{S}_{\mathcal{F},\mathcal{H}}$. The finite set $\mathcal{X}_q$ is the queried covariate support. Whenever $R_{n,K}$, R_q°, $q^\circ_{f,\alpha,T}$, $q^\circ_{\alpha,\mathcal{F},T}$, $\psi_{n,q}$, or $\mathcal{S}_{\mathcal{F},\mathcal{H}_q}$ is used, the covariates entering the calibration and deployment problem are taken to satisfy $P_{\mathrm{src},X}(X \in \mathcal{X}_q) = 1$ and $P_f(X \in \mathcal{X}_q) = 1$ for every $f \in \mathcal{F}$. For a larger covariate space, the same formulas apply to the query-restricted problem on $\mathcal{X}_q$; a conclusion over all of $\mathcal{X}$ requires either teacher queries outside $\mathcal{X}_q$ or a separate smoothness model transferring score information across covariates.

For population-quantile estimation only, we define $q^\circ_{f,\alpha,T} = \inf\{q : P_f\{T(X,Y) \le q\} \ge 1 - \alpha\}$ and $q^\circ_{\alpha,\mathcal{F},T} = \sup_{f \in \mathcal{F}} q^\circ_{f,\alpha,T}$. When $T = R_q^\circ$, we define $q^\circ_{f,\alpha,q} = q^\circ_{f,\alpha,R_q^\circ}$ and $q^\circ_{\alpha,\mathcal{F},q} = q^\circ_{\alpha,\mathcal{F},R_q^\circ}$.

The local quantile condition is used to compare a calibrated threshold with a population quantile. Specifically, the oracle score T satisfies Assumption S1 with $q^\circ_{\alpha,\mathcal{F}}$, r_0, c_-, and F_f° replaced by $q^\circ_{\alpha,\mathcal{F},T}$, $r_{0,T}$, $c_{-,T}$, and the distribution function of $T(X,Y)$ under P_f; in the special case $T = R_q^\circ$, we define $r_{0,q} = r_{0,R_q^\circ}$ and $c_{-,q} = c_{-,R_q^\circ}$. Proposition S25 gives a primitive density condition for this local quantile condition.

For coverage control, the analogue of the calibration empirical-process bound (S.34) holds for the score S with bound $\psi_{n,q}(\delta, \mathcal{F})$, and also for T when an oracle screened threshold is compared. This empirical-process part follows from Theorem S14 whenever the screened scores are measurable and constructed independently of the calibration fold. The population-quantile conclusions of Theorem S20 use the empirical-process and Monte Carlo events through a union bound. The paired radius conclusion of Theorem S16 uses only the Monte Carlo event.

For a paired radius comparison, the radial Lipschitz property in Proposition S12 holds for the score ball families generated by both S and T with the same bound. The contraction argument after Lemma S2 establishes this radial property for centered $\mathcal{H}_q$-balls on a convex image class with linear functionals and a feasible center; Proposition S18 gives a sufficient condition by finite slab intersection under the star-shaped coordinate-image condition.

The following proposition gives a primitive density condition for the screened local quantile condition.

Proposition S25. *Let T be an oracle screened score, and let $F_{f,T}$ denote the distribution function of $T(X,Y)$ under P_f. Suppose $q^\circ_{\alpha,\mathcal{F},T} < \infty$. If there exist constants $r_{0,T} > 0$ and $c_{-,T} > 0$ such that, for every $f \in \mathcal{F}$ satisfying $q^\circ_{f,\alpha,T} \ge q^\circ_{\alpha,\mathcal{F},T} - r_{0,T}$, one has $F_{f,T}(q^\circ_{f,\alpha,T}) = 1 - \alpha$ and $F_{f,T}$ is absolutely continuous on $[q^\circ_{f,\alpha,T} - r_{0,T}, q^\circ_{f,\alpha,T}]$ with density at least $c_{-,T}$, then T satisfies the screened local quantile condition with radius $r_{0,T}$ and constant $c_{-,T}$.*

Proposition S25 gives a sufficient density condition for Assumption S1. The assumption itself is weaker and allows discrete scores. Neither statement assumes concentration of an estimated quantile.

The following proposition shows that a lower bound on threshold error gives a lower bound on radius error within a fixed score family when the radius increases at a positive rate with the threshold.

Proposition S26. *Let T be an oracle screened score and let $I \subseteq \mathbb{R}$. Suppose there exist a measurable function $a : \mathcal{X}_q \to [0, \infty)$ and a constant $a_0 > 0$ such that $\inf_{f \in \mathcal{F}} \mathbb{E}_f\{a(X)\} \geq a_0$ and, for every $x \in \mathcal{X}_q$ and every $s, t \in I$ with $t \geq s$,*

$$\mathrm{rad}_{\mathcal{H}_q}\{\mathcal{C}_t^T(x)\} - \mathrm{rad}_{\mathcal{H}_q}\{\mathcal{C}_s^T(x)\} \geq a(x)(t-s).$$

Assume also that

$$\mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\mathcal{C}_u^T) < \infty \qquad \text{for every } u \in I.$$

Then, for every $s, t \in I$,

$$\left|\mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\mathcal{C}_t^T) - \mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\mathcal{C}_s^T)\right| \geq a_0 \left|t - s\right|.$$

Proposition S26 requires a positive increase in radius with the threshold. A threshold lower bound alone does not imply an image-radius lower bound.

When the Lipschitz or positive-scale condition for the linear Wasserstein term is not imposed, we interpret $L_{\mathcal{H}}\varepsilon_W$ as $+\infty$ in Theorem S16.

The following theorem bounds the score and radius errors through Wasserstein and Kullback–Leibler distances. A separate example under only Wasserstein control gives a radius error proportional to the predictor distribution error. In this scalar example, $h_0(y) = y$.

We let $\widetilde{\mu}_h(x) = \mathbb{E}_{Q_x^{\mathrm{T}}} h(Y)$ and $\widetilde{\sigma}_h(x) = \mathrm{Var}_{Q_x^{\mathrm{T}}}\{h(Y)\}^{1/2}$. For the linear Wasserstein term only, we assume $\inf_{x \in \mathcal{X}_q, h \in \mathcal{H}_q}\{\sigma_h^\circ(x) + \widetilde{\sigma}_h(x)\} \geq \sigma_0 > 0$. In both distance comparisons, we suppose that, conditional on $\{Q_x^{\mathrm{T}} : x \in \mathcal{X}_q\}$, $\{\widetilde{Y}^{(k)}(x) : k = 1, \ldots, K\}$ are i.i.d. from Q_x^{T} for each $x \in \mathcal{X}_q$. We define $e_{n,q} = \sup_{x \in \mathcal{X}_q} \sup_{y \in \mathcal{Y}} \left|R_{n,K}(x,y) - R_q^\circ(x,y)\right|$.

Theorem S16. *Suppose Assumptions S4 and S5 hold on $\mathcal{X}_q \times \mathcal{H}_q$, where the Lipschitz Wasserstein part of Assumption S4 is imposed only when the $L_{\mathcal{H}}\varepsilon_W$ term is used. Fix $0 < \delta < 1$ and define $\delta_q = \delta/3$. Then there exists a constant $C_{\mathrm{sc}} > 0$, depending only on $B_{\mathcal{H}}$ and λ, and also on σ_0 when the linear Wasserstein term is used, such that, after defining*

$$\bar{e}_{n,q}(\delta_q) = C_{\mathrm{sc}} \left\{ \min\{L_{\mathcal{H}}\varepsilon_W, B_{\mathcal{H}}\sqrt{\varepsilon_K}\} + K^{-1/2}\sqrt{1 + \log(|\mathcal{X}_q||\mathcal{H}_q|/\delta_q)} \right\},$$

with probability at least $1 - \delta_q$,

$$e_{n,q} \leq \bar{e}_{n,q}(\delta_q). \tag{S.59}$$

If, in addition, $\mathcal{F}$ is pointwise measurable and the two scores belong to the common family of Proposition S12, calibrate both scores on the same observations using (S.30) and the same nonnegative tolerance $\tau_{n,q} = 2\psi_{n,q}(\delta_q, \mathcal{F})$. Then, with probability at least $1 - \delta$, for an absolute constant $C_{\mathrm{rad}} > 0$,

$$\left|\mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\widehat{\mathcal{C}}_{q,K}) - \mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\widehat{\mathcal{C}}_q^\circ)\right| \leq A_{\mathcal{F}} C_{\mathrm{rad}} \{\bar{e}_{n,q}(\delta_q) + \psi_{n,q}(\delta_q, \mathcal{F})\}. \tag{S.60}$$

In contrast, for every $\varepsilon_W \in (0, 1/2]$ and every deterministic tolerance $0 \leq \tau_n < 1 - \alpha$, let $(\widehat{q}_{\tau_n}, \widehat{\mathcal{C}}_{\tau_n})$ and $(\widehat{q}_{\tau_n}^\circ, \widehat{\mathcal{C}}_{\tau_n}^\circ)$ denote the one-sided rules obtained from (S.30) with tolerance τ_n and scores $R_{n,K}$ and R_q°, respectively. Without the screened local-quantile, Kullback–Leibler, or nondegenerate-scale conditions, there exists a one-parameter teacher family on $\mathcal{Y} = [0,1]$, where the reference metric is Euclidean, $K = \infty$, $\mathcal{F} = \{1\}$, $\mathcal{H}_q = \mathcal{H} = \{h_0\}$, and $L_{\mathcal{H}} = 1$, such that

$$\mathcal{S}_{\mathcal{F},\mathcal{H}}(\widehat{\mathcal{C}}_{\tau_n}) - \mathcal{S}_{\mathcal{F},\mathcal{H}}(\widehat{\mathcal{C}}_{\tau_n}^\circ) \geq \varepsilon_W. \tag{S.61}$$

In (S.60), (S.37) gives $\bar{e}_{n,q}(\delta_q) = C_{\mathrm{sc}} b_{q,K}(\delta_q)$. When no screened local-quantile, Kullback–Leibler, or nondegenerate-scale regularity is imposed, the lower-bound part of Theorem S16 shows that the radius difference is at least of order $L_{\mathcal{H}}\varepsilon_W$ through the Wasserstein term alone. This example has zero variances and infinite Kullback–Leibler divergence. It does not give a matching minimax lower bound for either distance branch of (S.60).

The linear Wasserstein rate requires a positive lower bound on the sum of the source and teacher standard deviations. Without it, a square-root Wasserstein term still controls the score. This case includes functional values that are determined by the observed coarse image. We write $\|h\|_{\mathrm{Lip}}$ for the smallest L such that $|h(y) - h(y')| \leq L d_{\mathcal{Y}}(y, y')$ for all $y, y' \in \mathcal{Y}$.

Proposition S27. *Suppose the finite queried classes satisfy* $|h| \le B_{\mathcal{H}}$, $\|h\|_{\mathrm{Lip}} \le L_{\mathcal{H}}$, *the teacher sampling condition, and* $\sup_x W_1(Q_x^{\mathrm{T}}, P_{Y|X=x}) \le \varepsilon_W$. *For* $0 < \eta < 1$, *with probability at least* $1-\eta$,

$$e_{n,q} \le \frac{L_{\mathcal{H}}\varepsilon_W}{\lambda} + \frac{4B_{\mathcal{H}}\sqrt{B_{\mathcal{H}}L_{\mathcal{H}}\varepsilon_W}}{\lambda^2} + C_{B_{\mathcal{H}},\lambda}\sqrt{\frac{1+\log(|\mathcal{X}_q||\mathcal{H}_q|/\eta)}{K}}. \tag{S.62}$$

Unlike a lower variance bound, the ridge $\lambda > 0$ is part of the specified score. Equation (S.62) uses it to accommodate deterministic coarse summaries. The conclusion is a score bound, not a claim that the empirical medical predictors satisfy a known Wasserstein bound.

Proof. Lemma S14 and $|\sqrt{u} - \sqrt{v}| \le \sqrt{|u-v|}$ give

$$\sup_{x,h} |\widetilde{\mu}_h(x) - \mu_h^{\circ}(x)| \le L_{\mathcal{H}}\varepsilon_W, \qquad \sup_{x,h} |\widetilde{\sigma}_h(x) - \sigma_h^{\circ}(x)| \le 2\sqrt{B_{\mathcal{H}}L_{\mathcal{H}}\varepsilon_W}.$$

Lemma S6 adds a term of order $\sqrt{\{1+\log(|\mathcal{X}_q||\mathcal{H}_q|/\eta)\}/K}$ to each moment error. For $|h(y)|, |\mu|, |\mu'| \le B_{\mathcal{H}}$ and $a, a' \ge \lambda$,

$$\left| \frac{|h(y)-\mu|}{a} - \frac{|h(y)-\mu'|}{a'} \right| \le \lambda^{-1}|\mu - \mu'| + 2B_{\mathcal{H}}\lambda^{-2}|a - a'|.$$

Taking the supremum proves (S.62). This completes the proof. □

The square-root dependence cannot in general be replaced by a linear Wasserstein rate. For $h(y) = y$, $P = \delta_0$, and $Q = (1-\varepsilon)\delta_0 + \varepsilon\delta_1$, one has $W_1(P,Q) = \varepsilon$, $\sigma_P = 0$, and $\sigma_Q = \sqrt{\varepsilon(1-\varepsilon)}$. At $y = 1$ and fixed $\lambda > 0$, the score difference satisfies

$$\frac{1}{\sqrt{\varepsilon}} \left\{ \frac{1}{\lambda} - \frac{1-\varepsilon}{\lambda + \sqrt{\varepsilon(1-\varepsilon)}} \right\} \longrightarrow \frac{1}{\lambda^2}.$$

Thus a uniform linear Wasserstein score bound fails when zero variances are allowed, whereas (S.86) gives the square-root Kullback–Leibler rate.

Remark S1. Because the calibration empirical-process closures impose bounded shift classes through Assumption S3, the theory applies directly to bounded shift classes and to bounded normalized truncations that are explicitly included in the specified class $\mathcal{F}$. Unbounded or heavy tailed density ratios require a separate empirical-process argument beyond (S.34).

S6.8 Sequential deployment and online validity under drift

For a fixed calibrated set, average undercoverage is controlled by the average distance of the deployment shifts from the protected class.

We let $\widehat{\mathcal{C}}$ denote the prediction set calibrated once on $\mathcal{D}_{\mathrm{cal}}$ via the generic-class calibration rule (S.30), and we consider a stream of target observations $\{(X_{n+t}, Y_{n+t})\}_{t\ge1}$ with unobserved time-varying shift weights $f_t^\star$ satisfying $\mathbb{E}_{P_{\mathrm{src},X}} f_t^\star(X) = 1$. We let $\mathcal{G}_n = \sigma\{\mathcal{D}_{\mathrm{tr}}, \mathcal{D}_{\mathrm{cal}}, \text{score-construction randomness}\}$. This implies that $\widehat{\mathcal{C}}$ is $\mathcal{G}_n$-measurable. Throughout this subsection, $P_{f_t^\star}\{\cdot \mid \mathcal{G}_n\}$ denotes the conditional one-step marginal distribution of (X_{n+t}, Y_{n+t}) given $\mathcal{G}_n$, where the covariate shift weight is $f_t^\star$ relative to P_{src}.

The following theorem establishes a conditional bound on coverage shortfall that holds at every prediction horizon for the once-calibrated set $\widehat{\mathcal{C}}$ when each time-varying weight $f_t^\star$ remains in the specified class $\mathcal{F}$, without stationarity, exchangeability, or mixing.

Theorem S17. *Fix* $0 < \delta < 1$. *Suppose* $\mathcal{F}$ *is pointwise measurable and the calibration empirical-process bound* (S.34) *holds. Let* $\psi_{n,\delta} = \psi_n(\delta, \mathcal{F})$, *and let* $\widehat{\mathcal{C}}$ *be calibrated with a deterministic tolerance* $\tau_n \ge 0$ *using the attained measurable version of* (S.30). *If* $f_t^\star \in \mathcal{F}$ *for every* $t \ge 1$, *then for every horizon* $T \ge 1$, *with probability at least* $1-\delta$,

$$T^{-1}\sum_{t=1}^{T} \left[(1-\alpha) - P_{f_t^\star}\{Y_{n+t} \in \widehat{\mathcal{C}}(X_{n+t}) \mid \mathcal{G}_n\} \right]_+ \le \tau_n + \psi_{n,\delta},$$

uniformly in T *and without stationarity, exchangeability, or mixing.*

The following theorem extends Theorem S17 to the misspecified-shift regime in which $f_t^\star \notin \mathcal{F}$, by adding a penalty proportional to the per-step $L^1(P_{\mathrm{src},X})$ drift κ_t.

Theorem S18. *Fix $0 < \delta < 1$. Suppose $\mathcal{F}$ is pointwise measurable and the calibration empirical-process bound (S.34) holds. Let $\psi_{n,\delta} = \psi_n(\delta, \mathcal{F})$, and let $\widehat{\mathcal{C}}$ be calibrated with a deterministic tolerance $\tau_n \geq 0$ using the attained measurable version of (S.30). If $f_t^\star$ is not necessarily in $\mathcal{F}$ and has L^1 drift $\kappa_t = \inf_{f\in\mathcal{F}} \mathbb{E}_{P_{\mathrm{src},X}} |f_t^\star(X) - f(X)|$, then for every horizon $T \geq 1$, with probability at least $1-\delta$,*

$$T^{-1}\sum_{t=1}^{T}\left[(1-\alpha) - P_{f_t^\star}\{Y_{n+t} \in \widehat{\mathcal{C}}(X_{n+t}) \mid \mathcal{G}_n\}\right]_+ \leq \tau_n + \psi_{n,\delta} + \bar{\kappa}_T, \qquad \bar{\kappa}_T = T^{-1}\sum_{t=1}^{T}\kappa_t.$$

Theorem S18 bounds the average coverage shortfall by the calibration error and the average $L^1(P_{\mathrm{src},X})$ distance $\bar{\kappa}_T$ from $\mathcal{F}$. The probabilities condition on $\mathcal{G}_n$; neither sequential theorem asserts concentration of the realized fraction of uncovered images.

S6.9 Cross-fitted augmented calibration

We estimate the density ratio and the predictor moments on observations outside each calibration fold. These estimates are nuisance quantities, and their error contributes a separate term to the coverage bound.

We partition $\mathcal{I}_{\mathrm{cal}}$ into $K_{\mathrm{CV}} \geq 2$ folds $\mathcal{I}_{\mathrm{cal}}^{(1)}, \ldots, \mathcal{I}_{\mathrm{cal}}^{(K_{\mathrm{CV}})}$ of equal size. For each fold k, we let $\widehat{r}^{(-k)}$ be the density-ratio estimator fit on $\bigcup_{j\neq k}\mathcal{I}_{\mathrm{cal}}^{(j)}$ and the unlabeled target sample, and let $\widehat{\mu}_h^{(-k)}, \widehat{\sigma}_h^{(-k)}$ be teacher-plug-in estimators trained on the same out-of-fold data. For $i \in \mathcal{I}_{\mathrm{cal}}^{(k)}$, we define the cross-fitted score

$$R_n^{\mathrm{CF}}(X_i, Y_i) = R_n^{(-k)}(X_i, Y_i) = \sup_{h\in\mathcal{H}} \frac{|h(Y_i) - \widehat{\mu}_h^{(-k)}(X_i)|}{\widehat{\sigma}_h^{(-k)}(X_i) + \lambda}.$$

We let $k(i)$ denote the index of the fold containing observation i. We define $\widehat{r}_i = \widehat{r}^{(-k(i))}(X_i)$ and $\overline{\widehat{r}} = n_{\mathrm{cal}}^{-1}\sum_i \widehat{r}_i$. The augmented one-sided discrepancy is

$$\Gamma_n^{\mathrm{CF,DR}}(q;\mathcal{F}) = \sup_{f\in\mathcal{F}} n_{\mathrm{cal}}^{-1}\sum_i \left[f(X_i) + \widehat{r}_i - \overline{\widehat{r}}\right]\{\mathbf{1}\{R_n^{\mathrm{CF}}(X_i,Y_i) > q\} - \alpha\}. \tag{S.63}$$

We let $\widehat{q}_{\mathrm{DR}} = \inf\{q \geq 0 : \Gamma_n^{\mathrm{CF,DR}}(q;\mathcal{F}) \leq \tau_n\}$. The vector $\{\mathbf{1}\{R_n^{\mathrm{CF}}(X_i,Y_i) > q\} : i \in \mathcal{I}_{\mathrm{cal}}\}$ takes only finitely many values as q varies because the calibration fold is finite. Under the strict-exceedance convention, each realized indicator vector is attained at the left endpoint of the corresponding interval. Suppose that the realized cross-fitted calibration scores are finite. Then, for any q greater than or equal to all realized cross-fitted scores, it holds that $\Gamma_n^{\mathrm{CF,DR}}(q;\mathcal{F}) = \sup_{f\in\mathcal{F}}[-\alpha n_{\mathrm{cal}}^{-1}\sum_i f(X_i)] \leq 0 \leq \tau_n$, where we have used the condition that $n_{\mathrm{cal}}^{-1}\sum_i(\widehat{r}_i - \overline{\widehat{r}}) = 0$ and $f \geq 0$. Under this finite-score condition, the feasible set is nonempty and the infimum is attained. Throughout the augmented-calibration results that follow, $\widehat{q}_{\mathrm{DR}}$ denotes an attained measurable version of this infimum. For deployment, we draw $K^\star$ uniformly from $\{1, \ldots, K_{\mathrm{CV}}\}$ independently of all data, and output

$$\widehat{\mathcal{C}}_{\mathrm{DR}}(x) = \{y : R_n^{(-K^\star)}(x,y) \leq \widehat{q}_{\mathrm{DR}}\}.$$

We evaluate coverage by averaging its indicator over the independently selected fold $K^\star$. Throughout this subsection, target probabilities in the cross-fitted statements are evaluated conditional on the out-of-fold nuisance fits whenever the result conditions on those fits. In $\mathbb{E}_{K^\star}[P_{\mathrm{tar}}\{\cdot \mid K^\star, \mathcal{D}_{\mathrm{cal}}\}]$, the calibration data and nuisance fits are fixed. The inner probability evaluates the set from the selected fold $K^\star$. The outer expectation averages over the independent uniform draw of $K^\star$.

For $q \in \mathbb{R}$, we define

$$M_k(q) = \mathbf{1}\{R_n^{(-k)}(X,Y) > q\}, \qquad \bar{M}(q) = K_{\mathrm{CV}}^{-1}\sum_{k=1}^{K_{\mathrm{CV}}} M_k(q).$$

For the following theorem, we define $\bar{r}_n^P = K_{\mathrm{CV}}^{-1} \sum_{k=1}^{K_{\mathrm{CV}}} \mathbb{E}_{P_{\mathrm{src},X}}\{\widehat{r}^{(-k)}(X)\}$ and

$$\rho_n^{\mathrm{aug}} = \inf_{f\in\mathcal{F}} \sup_{q\in\mathbb{R}} \mathbb{E}_{P_{\mathrm{src}}} \left| r(X)\bar{M}(q) - K_{\mathrm{CV}}^{-1} \sum_{k=1}^{K_{\mathrm{CV}}} \{f(X) + \widehat{r}^{(-k)}(X) - \bar{\widehat{r}}\} M_k(q) \right| + \alpha \left| \bar{\widehat{r}} - \bar{r}_n^P \right|. \tag{S.64}$$

We fix $0 < \delta < 1$, and let $\widehat{q}_{\mathrm{DR}}$ be the attained infimum in the augmented rule. We define $\psi_n^{\mathrm{DR}}(\delta)$ for the bound proved in Proposition S28 in the fully sample-split case, and for $\psi_{n,K}^{\mathrm{DR}}(\delta)$ from Proposition S29 in the foldwise case. The corresponding empirical-process inequality is

$$\begin{aligned} \sup_{f\in\mathcal{F}} \sup_{q\in\mathbb{R}} \Bigg| & n_{\mathrm{cal}}^{-1} \sum_i \big[f(X_i) + \widehat{r}_i - \bar{\widehat{r}}\big]\{\mathbf{1}\{R_n^{\mathrm{CF}}(X_i, Y_i) > q\} - \alpha\} \\ & - K_{\mathrm{CV}}^{-1} \sum_{k=1}^{K_{\mathrm{CV}}} \mathbb{E}_{P_{\mathrm{src}}} \Big[\{f(X) + \widehat{r}^{(-k)}(X) - \bar{\widehat{r}}\}\{\mathbf{1}\{R_n^{(-k)}(X, Y) > q\} - \alpha\}\Big] \Bigg| \le \psi_n^{\mathrm{DR}}(\delta) \end{aligned} \tag{S.65}$$

Proposition S28 proves (S.65) conditional on independently trained nuisance fits. Proposition S29 proves it for the ordinary foldwise construction after averaging over the training data. In the latter case, we define $\psi_n^{\mathrm{DR}}(\delta) = \psi_{n,K}^{\mathrm{DR}}(\delta)$. The following theorem gives the corresponding coverage bounds.

Theorem S19. *Suppose (S.21) holds, the calibration scores are finite, and the conditions of Proposition S28 or Proposition S29 hold. For a deterministic tolerance $\tau_n \ge 0$, with probability at least $1 - \delta$,*

$$\Big[(1-\alpha) - \mathbb{E}_{K^\star}\Big[P_{\mathrm{tar}}\{Y \in \widehat{\mathcal{C}}_{\mathrm{DR}}(X) \mid K^\star, \mathcal{D}_{\mathrm{cal}}\}\Big]\Big]_+ \le \tau_n + \psi_n^{\mathrm{DR}}(\delta) + \rho_n^{\mathrm{aug}}. \tag{S.66}$$

In the fully sample-split case, this probability bound also holds conditional on the nuisance fits. For the fitted procedure held fixed, the inequality holds on the same event for every normalized nonnegative target ratio r, with P_{tar} and ρ_n^{aug} evaluated at that ratio.

The proof also gives (S.66) for every realization satisfying (S.65). Thus any other valid conditional or unconditional probability bound for (S.65) gives the same probability bound for (S.66).

Theorem S19 gives an additive bound, and not an exact double robustness bound. The bound is additive because the augmentation in (S.63) replaces the target weight r by a fold-dependent surrogate, and the remaining discrepancy equals the remainder term (S.64). Exact orthogonal double robustness requires an additional outcome-regression augmentation for the exceedance process beyond the density-ratio augmentation. For marginal target coverage under a single covariate shift, Yang et al. (2024) construct such an orthogonal augmentation and prove asymptotic validity when either the density ratio or the conditional score distribution is estimated consistently.

The supremum in (S.63) ranges over surrogate weights indexed by $\mathcal{F}$. Theorem S19 gives a common calibration event for all target ratios, but its remainder ρ_n^{aug} depends on the ratio being evaluated. A common coverage bound over the original target class $\mathcal{F}$ requires uniform control of these remainders. Indexing the surrogate weights by $\mathcal{F}$ alone does not provide this control. The density-ratio augmentation here differs from the orthogonal score of Yang et al. (2024).

The following proposition gives a primitive sufficient condition for (S.65) in a fully sample-split variant where all nuisance fits are trained on data independent of the calibration sample entering the empirical average.

Proposition S28. *Suppose that the nuisance fits used in (S.63) are trained on data independent of the calibration observations appearing in the empirical average. Conditional on these nuisance fits, suppose that the fold-specific scores $R_n^{(-k)} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, $k = 1, \dots, K_{\mathrm{CV}}$, are measurable, and suppose that the class*

$$\mathcal{W}_{\mathrm{DR},0} = \Big\{x \mapsto f(x) + \widehat{r}^{(-k)}(x) : f \in \mathcal{F},\ k = 1, \dots, K_{\mathrm{CV}}\Big\}$$

is pointwise measurable, uniformly bounded by $B_{\mathrm{DR},0} \ge 1$, and, for deterministic constants $A_{\mathrm{DR}} \ge 2$ and $v_{\mathrm{DR}} < \infty$, satisfies

$$\sup_Q N\{\eta B_{\mathrm{DR},0}, \mathcal{W}_{\mathrm{DR},0}, L_2(Q)\} \le (A_{\mathrm{DR}}/\eta)^{v_{\mathrm{DR}}} \qquad 0 < \eta \le 1.$$

Assume also that $|\overline{\widehat{r}}| \le B_{\mathrm{DR},0}$ *almost surely. Then, there exists an absolute constant* $C > 0$ *such that, for every* $0 < \delta < 1$*, the bound* (S.65) *holds with conditional probability at least* $1 - \delta$ *for*

$$\psi_n^{\mathrm{DR}}(\delta) = C B_{\mathrm{DR},0} \sqrt{\frac{(v_{\mathrm{DR}} + 3)\log(A_{\mathrm{DR}} n_{\mathrm{cal}}) + \log(K_{\mathrm{CV}}/\delta)}{n_{\mathrm{cal}}}}.$$

Under the conditions of Proposition S28, it holds that $|f + \widehat{r}^{(-k)} - \overline{\widehat{r}}| \le 2B_{\mathrm{DR},0}$, and this uniform bound verifies the empirical-process bound (S.65).

Corollary S4. *Consider the fully sample-split implementation covered by Proposition S28. Suppose that* $\{R_n^{\mathrm{CF}}(X_i, Y_i) : i \in \mathcal{I}_{\mathrm{cal}}\} \subset \mathbb{R}$*, and let* $\widehat{q}_{\mathrm{DR}}$ *denote the attained measurable version of* $\inf\{q \ge 0 : \Gamma_n^{\mathrm{CF,DR}}(q; \mathcal{F}) \le \tau_n\}$*. Fix* $0 < \delta < 1$*, and define the fully sample-split bound* $\psi_{n,\mathrm{fs}}^{\mathrm{DR}}(\delta)$ *by*

$$\psi_{n,\mathrm{fs}}^{\mathrm{DR}}(\delta) = C B_{\mathrm{DR},0} \sqrt{\frac{(v_{\mathrm{DR}} + 3)\log(A_{\mathrm{DR}} n_{\mathrm{cal}}) + \log(K_{\mathrm{CV}}/\delta)}{n_{\mathrm{cal}}}}.$$

Suppose that the remainder definition (S.64) *is evaluated for the same fully sample-split nuisance fits. If* $\widehat{q}_{\mathrm{DR}}$ *is computed with a deterministic tolerance* $\tau_n \ge 0$*, then, conditional on the nuisance fits, it holds with probability at least* $1 - \delta$ *that*

$$\left[(1-\alpha) - \mathbb{E}_{K^\star}\left[P_{\mathrm{tar}}\{Y \in \widehat{\mathcal{C}}_{\mathrm{DR}}(X) \mid K^\star, \mathcal{D}_{\mathrm{cal}}\}\right]\right]_+ \le \tau_n + \psi_{n,\mathrm{fs}}^{\mathrm{DR}}(\delta) + \rho_n^{\mathrm{aug}}.$$

Corollary S4 gives a calibration guarantee for independently trained nuisance functions, including the additive approximation term.

Proof. Conditional on the nuisance fits, Proposition S28 verifies (S.65) with the bound $\psi_{n,\mathrm{fs}}^{\mathrm{DR}}(\delta)$ and conditional probability at least $1 - \delta$. Applying Theorem S19 on this event gives the stated bound. This completes the proof. □

The following proposition proves the empirical-process bound for the K_{CV}-fold construction. Using overlapping out-of-fold nuisance fits introduces an explicit factor $K_{\mathrm{CV}}^{1/2}$ in the sufficient bound.

Proposition S29. *Let* $n_{\mathrm{fold}} = n_{\mathrm{cal}}/K_{\mathrm{CV}}$*, and assume that the folds have equal size. For each fold* k*, let* $\mathcal{A}_k$ *be the sigma-field generated by the calibration observations outside* $\mathcal{I}_{\mathrm{cal}}^{(k)}$*, the unlabeled target covariates, and the exogenous score-construction randomness used to construct* $\widehat{r}^{(-k)}$ *and* $R_n^{(-k)}$*. Conditional on* $\mathcal{A}_k$*, suppose that the fold-*k *observations* $\{(X_i, Y_i) : i \in \mathcal{I}_{\mathrm{cal}}^{(k)}\}$ *are independent and identically distributed according to* P_{src}*,* $R_n^{(-k)} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ *is fixed and measurable, and the class*

$$\mathcal{W}_{\mathrm{DR},k} = \{x \mapsto f(x) + \widehat{r}^{(-k)}(x) : f \in \mathcal{F}\}$$

is pointwise measurable, uniformly bounded by $B_{\mathrm{DR},K} \ge 1$*, and, for deterministic constants* $A_{\mathrm{DR}} \ge 2$ *and* $v_{\mathrm{DR}} < \infty$*, satisfies almost surely*

$$\sup_Q N\{\eta B_{\mathrm{DR},K}, \mathcal{W}_{\mathrm{DR},k}, L_2(Q)\} \le (A_{\mathrm{DR}}/\eta)^{v_{\mathrm{DR}}} \qquad 0 < \eta \le 1.$$

Assume also that $|\overline{\widehat{r}}| \le B_{\mathrm{DR},K}$ *almost surely. Then, there exists an absolute constant* $C > 0$ *such that, for every* $0 < \delta < 1$*, the bound* (S.65) *holds with probability at least* $1 - \delta$ *for*

$$\psi_{n,K}^{\mathrm{DR}}(\delta) = C B_{\mathrm{DR},K} \sqrt{\frac{K_{\mathrm{CV}}\{(v_{\mathrm{DR}} + 3)\log(A_{\mathrm{DR}} n_{\mathrm{cal}}) + \log(K_{\mathrm{CV}}/\delta)\}}{n_{\mathrm{cal}}}}.$$

The following proposition verifies the multiplier-class conditions in Propositions S28 and S29 for compact exponential-tilt ratio fits. Let B_ϕ be a finite bound for $\sup_{x \in \mathcal{X}} \|\phi(x)\|$. Define $M_\Theta = \sup_{\theta \in \Theta_{\mathrm{tilt}}} \|\theta\|$ and $B_{\mathrm{tilt}} = \exp(2M_\Theta B_\phi)$. For each fold k, write

$$\widehat{r}^{(-k)}(x) = \exp\{(\widehat{\theta}^{(-k)})^\top \phi(x) - \widehat{\Lambda}^{(-k)}(\widehat{\theta}^{(-k)})\},$$

where $\widehat{\Lambda}^{(-k)}$ is the empirical normalizing constant used in (S.49) for the foldwise fit.

Proposition S30. *Suppose that Assumption S3 holds,* $\sup_{x\in\mathcal{X}}\|\phi(x)\| \le B_\phi < \infty$, *and* Θ_{tilt} *is compact. Put*

$$B_{\mathrm{DR}}^{\mathrm{tilt}} = 1 + B_F + B_{\mathrm{tilt}}, \qquad A_{\mathrm{DR}}^{\mathrm{tilt}} = 2K_{\mathrm{CV}}(A_F \vee 1), \qquad v_{\mathrm{DR}}^{\mathrm{tilt}} = v_F + 1.$$

Then, for every realization of the foldwise training samples, it holds that $|\bar{\widehat{r}}| \le B_{\mathrm{DR}}^{\mathrm{tilt}}$, *and the classes*

$$\mathcal{W}_{\mathrm{DR},0}^{\mathrm{tilt}} = \left\{x \mapsto f(x) + \widehat{r}^{(-k)}(x) : f \in \mathcal{F},\ k = 1,\dots,K_{\mathrm{CV}}\right\}$$

and

$$\mathcal{W}_{\mathrm{DR},k}^{\mathrm{tilt}} = \left\{x \mapsto f(x) + \widehat{r}^{(-k)}(x) : f \in \mathcal{F}\right\}$$

are pointwise measurable, uniformly bounded by $B_{\mathrm{DR}}^{\mathrm{tilt}}$, *and satisfy, for every* $0 < \eta \le 1$,

$$\sup_Q N\{\eta B_{\mathrm{DR}}^{\mathrm{tilt}}, \mathcal{W}_{\mathrm{DR},0}^{\mathrm{tilt}}, L_2(Q)\} \le (A_{\mathrm{DR}}^{\mathrm{tilt}}/\eta)^{v_{\mathrm{DR}}^{\mathrm{tilt}}},$$

and

$$\sup_Q N\{\eta B_{\mathrm{DR}}^{\mathrm{tilt}}, \mathcal{W}_{\mathrm{DR},k}^{\mathrm{tilt}}, L_2(Q)\} \le (A_{\mathrm{DR}}^{\mathrm{tilt}}/\eta)^{v_{\mathrm{DR}}^{\mathrm{tilt}}}, \qquad k = 1,\dots,K_{\mathrm{CV}}.$$

This proposition implies that, for compact exponential-tilt ratio fits, the multiplier-class boundedness and entropy conditions in Propositions S28 and S29 hold with $B_{\mathrm{DR},0} = B_{\mathrm{DR},K} = B_{\mathrm{DR}}^{\mathrm{tilt}}$, $A_{\mathrm{DR}} = A_{\mathrm{DR}}^{\mathrm{tilt}}$, and $v_{\mathrm{DR}} = v_{\mathrm{DR}}^{\mathrm{tilt}}$.

Corollary S5. *Under the conditions of Proposition S29, suppose that* $\{R_n^{\mathrm{CF}}(X_i, Y_i) : i \in \mathcal{I}_{\mathrm{cal}}\} \subset \mathbb{R}$, *and let* $\widehat{q}_{\mathrm{DR}}$ *denote the attained measurable version of* $\inf\{q \ge 0 : \Gamma_n^{\mathrm{CF,DR}}(q;\mathcal{F}) \le \tau_n\}$. *Fix* $0 < \delta < 1$. *If* $\widehat{q}_{\mathrm{DR}}$ *is computed with a deterministic tolerance* $\tau_n \ge 0$, *then, it holds with probability at least* $1-\delta$ *that*

$$\left[(1-\alpha) - \mathbb{E}_{K^\star}\left[P_{\mathrm{tar}}\{Y \in \widehat{\mathcal{C}}_{\mathrm{DR}}(X) \mid K^\star, \mathcal{D}_{\mathrm{cal}}\}\right]\right]_+ \le \tau_n + \psi_{n,K}^{\mathrm{DR}}(\delta) + \rho_n^{\mathrm{aug}}.$$

Corollary S5 includes the foldwise factor from the empirical-process bound. Reusing observations in nuisance training does not justify the fully independent-sample rate.

The following proposition gives a sufficient condition under which the additive remainder in (S.64) is negligible.

Proposition S31. *Conditional on the out-of-fold nuisance fits, suppose that there exist* $f_{\mathrm{aug}} \in \mathcal{F}$ *and deterministic quantities* $\eta_n^{\mathrm{aug}}, \eta_n^{\mathrm{cent}} \ge 0$ *such that*

$$K_{\mathrm{CV}}^{-1}\sum_{k=1}^{K_{\mathrm{CV}}} \mathbb{E}_{P_{\mathrm{src},X}}\left|r(X) - f_{\mathrm{aug}}(X) - \widehat{r}^{(-k)}(X) + \bar{\widehat{r}}\right| \le \eta_n^{\mathrm{aug}}$$

and $\left|\bar{\widehat{r}} - \bar{r}_n^P\right| \le \eta_n^{\mathrm{cent}}$. *Then, it holds that* $\rho_n^{\mathrm{aug}} \le \eta_n^{\mathrm{aug}} + \alpha\eta_n^{\mathrm{cent}}$.

This bound implies that the additive remainders in Corollaries S4 and S5 are negligible whenever $\eta_n^{\mathrm{aug}} + \eta_n^{\mathrm{cent}} = o(1)$.

The following proposition gives a high probability sufficient condition for ρ_n^{aug} under the compact exponential-tilt ratio model. Let $n_{\mathrm{src},k}$ and $n_{\mathrm{tar},k}$ denote the source and target covariate sample sizes used to construct $\widehat{r}^{(-k)}$, and recall that $n_{\mathrm{fold}} = n_{\mathrm{cal}}/K_{\mathrm{CV}}$. For $0 < \delta < 1$, put

$$\zeta_n^{\mathrm{tilt}}(\delta) = K_{\mathrm{CV}}^{-1}\sum_{k=1}^{K_{\mathrm{CV}}}\left\{\sqrt{\frac{1+\log(6K_{\mathrm{CV}}/\delta)}{n_{\mathrm{src},k}}} + \sqrt{\frac{1+\log(6K_{\mathrm{CV}}/\delta)}{n_{\mathrm{tar},k}}}\right\} + \sqrt{\frac{K_{\mathrm{CV}}\log(6K_{\mathrm{CV}}/\delta)}{n_{\mathrm{cal}}}}.$$

Proposition S32. *Suppose that* $1 \in \mathcal{F}$, $r = r_{\theta_0}$ *for* (S.48), $\sup_{x\in\mathcal{X}}\|\phi(x)\| \le B_\phi < \infty$, *and* Θ_{tilt} *is compact. Suppose that each* $\widehat{r}^{(-k)}$ *is the foldwise exponential-tilt estimator in* (S.49), *the conditions of Proposition S17 hold for each fold with common constants, and, conditional on the data used to construct* $\widehat{r}^{(-k)}$, *the covariates* $\{X_i : i \in \mathcal{I}_{\mathrm{cal}}^{(k)}\}$ *are independent and identically distributed according to* $P_{\mathrm{src},X}$. *Then, there exists a constant* $C < \infty$, *depending only on the constants in Proposition S17,* B_ϕ, *and* Θ_{tilt}, *such that, for every* $0 < \delta < 1$, *it holds with probability at least* $1-\delta$ *that*

$$\rho_n^{\mathrm{aug}} \le C\zeta_n^{\mathrm{tilt}}(\delta).$$

The fully sample-split implementation yields the bound in Corollary S4, whereas the K_{CV}-fold implementation is covered by Corollary S5 at the foldwise factor $K_{\mathrm{CV}}^{1/2}$. Proposition S31 shows that the remaining augmented-calibration term is controlled by an $L^1(P_{\mathrm{src},X})$ approximation of the residual density ratio after the fold-specific ratio augmentation.

S6.10 Teacher query complexity

Finite predictor queries can create sampling error even when the predictor has the correct conditional moments.

We assume that every teacher query used to construct $R_{n,K}$ lies in $\mathcal{X}_q$. We recall that $q^\circ_{f,\alpha,q}$ and $q^\circ_{\alpha,\mathcal{F},q}$ denote the screened oracle quantiles for R°_q, and let $\psi_{n,q}(\delta,\mathcal{F})$ denote the empirical-process bound in the screened regularity conditions. For $0<\delta<1$, we define $\delta_q=\delta/3$ and $L_{q,K,\delta}=1+\log(|\mathcal{X}_q||\mathcal{H}_q|/\delta_q)$.

The following theorem gives upper bounds for score, threshold, and radius errors from finitely many predictor queries. It also gives a lower bound for scalar threshold estimation in a separate bounded two-point subproblem. The bounded-query upper bound permits zero oracle variance; moment matching and the fixed ridge suffice.

For the upper bounds, we suppose $0<B_{\mathcal{H}}=\sup_{h\in\mathcal{H}_q,y\in\mathcal{Y}}|h(y)|<\infty$, and use conditionally i.i.d. teacher samples at each queried covariate. We assume the population moment equalities

$$\mathbb{E}_{Q_x^{\mathrm{T}}}h(Y)=\mu_h^\circ(x),\qquad \mathrm{Var}_{Q_x^{\mathrm{T}}}\{h(Y)\}^{1/2}=\sigma_h^\circ(x)$$

for every $x\in\mathcal{X}_q$ and $h\in\mathcal{H}_q$. The score bound uses only these conditions. The population-quantile bound also uses the screened local-quantile condition, calibration independence, and Assumption S3; Theorem S14 then gives both screened empirical-process bounds. We let $(\widehat{q}_K,\widehat{\mathcal{C}}_K)$ and $(\widehat{q}^\circ,\widehat{\mathcal{C}}^\circ)$ be the thresholds and sets obtained from $R_{n,K}$ and R°_q on the same observations by (S.30), with the same tolerance $\tau_{n,q}=2\psi_{n,q}(\delta_q,\mathcal{F})$. For population-quantile estimation, we require

$$3\psi_{n,q}(\delta_q,\mathcal{F})+C_{\mathrm{mc}}\sqrt{\frac{L_{q,K,\delta}}{K}}\le c_0 r_{0,q},$$

where $C_{\mathrm{mc}}>0$ is sufficiently large and depends only on $B_{\mathcal{H}}$ and λ, and $c_0,r_{0,q},c_{-,q}$ are the constants of Theorem S8 for the screened oracle score. The paired radius comparison instead uses the common radial property in Proposition S12; it needs neither this smallness inequality nor the local-quantile condition.

For the lower bound, we consider a bounded threshold-functional two-point query subproblem where the screened class is a singleton $\mathcal{H}_q=\{h_0\}$ and the response space contains two points y_- and y_+ such that $h_0(y_-)=-B_{\mathcal{H}}$ and $h_0(y_+)=B_{\mathcal{H}}$. We suppose that there exist $q_0\in\mathbb{R}$ and $x_0\in\mathcal{X}_q$ such that, under P_ϑ, $\{\widetilde{Y}^{(k)}(x_0):k=1,\dots,K\}$ are independent and identically distributed and

$$\mathcal{P}_{\mathrm{bin}}=\left\{\begin{array}{l}P_\vartheta: h_0\{\widetilde{Y}^{(k)}(x_0)\}\in\{-B_{\mathcal{H}},B_{\mathcal{H}}\},\\ \quad P_\vartheta\{h_0(\widetilde{Y}^{(k)}(x_0))=B_{\mathcal{H}}\}=\dfrac{1}{2}+\dfrac{\vartheta}{2B_{\mathcal{H}}},\\ \quad q^\circ_{\alpha,\mathcal{F},q}=q_0+\vartheta,\quad \vartheta\in\{-c_1B_{\mathcal{H}}K^{-1/2},c_1B_{\mathcal{H}}K^{-1/2}\}\end{array}\right\},$$

where $c_1>0$ is sufficiently small. We let $\mathcal{Q}_K$ be the class of threshold estimators measurable with respect to the K teacher samples.

This lower-bound experiment is separate from the moment-matched upper-bound experiment. Lemma S7 constructs its source distributions and verifies the threshold functional.

Theorem S20. *Under the bounded moment-matched query conditions, with probability at least* $1-\delta_q$,

$$\sup_{x\in\mathcal{X}_q}\sup_{h\in\mathcal{H}_q}|\widehat{\mu}_h(x)-\mu_h^\circ(x)|+\sup_{x\in\mathcal{X}_q}\sup_{h\in\mathcal{H}_q}|\widehat{\sigma}_h(x)-\sigma_h^\circ(x)|\le C_1\sqrt{\frac{L_{q,K,\delta}}{K}},\tag{S.67}$$

$$e_{n,q}\le C_1\sqrt{\frac{L_{q,K,\delta}}{K}}.\tag{S.68}$$

Here C_1 depends only on $B_{\mathcal{H}}$ and λ. Under the additional population-quantile conditions, with probability at least $1-\delta$,

$$\left|\widehat{q}_K - q^{\circ}_{\alpha,\mathcal{F},q}\right| \le C_2 \left\{ \psi_{n,q}(\delta_q, \mathcal{F}) + \sqrt{\frac{L_{q,K,\delta}}{K}} \right\}, \tag{S.69}$$

where C_2 depends only on $B_{\mathcal{H}}$, λ, and $c_{-,q}$. Under the bounded query and paired radial conditions alone, the following bound holds with probability at least $1-\delta$ for C_2 depending only on $B_{\mathcal{H}}$ and λ:

$$\left|\mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\widehat{\mathcal{C}}_K) - \mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\widehat{\mathcal{C}}^{\circ})\right| \le A_{\mathcal{F}} C_2 \left\{ \sqrt{\frac{L_{q,K,\delta}}{K}} + \psi_{n,q}(\delta_q, \mathcal{F}) \right\}. \tag{S.70}$$

In the separate two-point query experiment, for a constant $c > 0$,

$$\inf_{\widetilde{q} \in \mathcal{Q}_K} \sup_{P \in \mathcal{P}_{\mathrm{bin}}} \mathbb{E}_P \left|\widetilde{q} - q^{\circ}_{\alpha,\mathcal{F},q}\right| \ge c B_{\mathcal{H}} K^{-1/2}. \tag{S.71}$$

Theorem S20 bounds the query error when the teacher and source moments agree. Increasing K reduces the sampling term in the threshold and radius bounds; systematic predictor error is treated separately in Theorem S16. The lower bound concerns estimators that observe only the K teacher samples in the stated two-point subproblem. That subproblem does not impose the moment-matching condition of the upper bound and does not establish a minimax lower bound for image-set radius. The coverage guarantee uses calibration and does not require an accurate predictor.

The preceding query lower bound specifies a threshold functional in a separate two-point experiment. The following construction uses the oracle score defined from the same distribution that generates the queries. It derives the dependence of the oracle threshold on the query distribution.

We fix $0 < \alpha < 1/2$, $B > 0$, and $\lambda > 0$. We let $\mathcal{X} = \{x_0\}$, $\mathcal{F} = \{1\}$, $\mathcal{Y} = \{-B, B\}$, and $\mathcal{H} = \{y \mapsto y\}$. Under P_p, the source and teacher distributions both assign probability p to B. We define $p_0 = 3/4 - \alpha/2$ and $p_{\pm} = p_0 \pm \eta K^{-1/2}$, where the fixed positive η is sufficiently small that $1/2 < p_- < p_+ < 1 - \alpha$ for every $K \ge 1$. We let $q_\alpha(p)$ be the $1-\alpha$ quantile of the oracle score under P_p, and let $\mathcal{Q}_K$ contain estimators based only on K independent teacher outputs. For a distribution P, we write P^K for the joint distribution of K independent draws from P.

Proposition S33. *For the stated family,*

$$q_\alpha(p) = \frac{2Bp}{2B\sqrt{p(1-p)} + \lambda}, \qquad p \in \{p_-, p_+\}, \tag{S.72}$$

and there is $c_{\alpha,B,\lambda} > 0$ such that

$$\inf_{\widetilde{q} \in \mathcal{Q}_K} \max_{p \in \{p_-, p_+\}} \mathbb{E}_p |\widetilde{q} - q_\alpha(p)| \ge c_{\alpha,B,\lambda} K^{-1/2}. \tag{S.73}$$

Proposition S33 gives the query rate for an oracle threshold with exact teacher moments and a positive population scale. Its discrete score distribution satisfies Assumption S1 uniformly in K. A positive score density is not needed for this comparison. The information restriction excludes calibration responses; the result is not a lower bound for procedures that also use labeled calibration data or for image-set radius.

Proof. The source mean is $B(2p-1)$ and its standard deviation is $2B\sqrt{p(1-p)}$. The score at B is $2B(1-p)/\{2B\sqrt{p(1-p)} + \lambda\}$, and the score at $-B$ is the larger value in (S.72). Because $p < 1-\alpha$, that larger value is the $1-\alpha$ quantile. For every $t > 0$, $P_p\{R^\circ \le q_\alpha(p) - t\} \le p \le p_0 + \eta$. Thus Assumption S1 holds with $r_0 = 1$ and $c_- = 1 - \alpha - p_0 - \eta > 0$, independently of K. On a fixed neighborhood of p_0 inside $(1/2, 1-\alpha)$, its derivative is at least $2B/(B+\lambda)$. Hence, with $\Delta = q_\alpha(p_+) - q_\alpha(p_-)$,

$$\Delta \ge \frac{4B\eta}{B+\lambda} K^{-1/2}, \qquad D_{\mathrm{KL}}(P_{p_+}^K \| P_{p_-}^K) \le \frac{4\eta^2}{p_-(1-p_-)}.$$

We choose η such that the last bound is at most $1/2$. Pinsker's inequality gives total variation at most $1/2$. The nearest-threshold test has estimation error at least $\Delta/2$ whenever it errs. Thus the maximum expected error is at least $\Delta\{1 - \|P_{p_+}^K - P_{p_-}^K\|_{\mathrm{TV}}\}/4 \ge \Delta/8$, proving (S.73). This completes the proof. □

S6.11 Adaptive selection of image-functional classes

We select the number of spatial summaries by comparing estimated radii and bound the coverage shortfall after selection.

We let $\mathcal{H}_1, \dots, \mathcal{H}_J$ be a finite family of candidate image-functional classes. For each j, we let $\widehat{\mathcal{C}}_j$ denote the prediction set obtained by applying the same calibration rule with $\mathcal{H}_j$, and let $S_j = \mathcal{S}_{\mathcal{F},\mathcal{H}_j}(\widehat{\mathcal{C}}_j)$. We suppose that an additional independent selection sample is used to construct estimators $\widehat{S}_j$ of S_j, and let $\widehat{j}$ denote the smallest index in $\operatorname{argmin}_{1\le j\le J}\{\widehat{S}_j + \mathrm{pen}_j\}$, where $\mathrm{pen}_1, \dots, \mathrm{pen}_J$ are nonnegative penalties measurable with respect to the sigma-field generated by the candidate construction and independent of the selection sample.

We let $\mathcal{D}^X_{\mathrm{sel}} = \{X_i^{\mathrm{sel}}\}_{i=1}^{n_{\mathrm{sel}}}$ be an additional selection sample consisting of independent observations from $P_{\mathrm{src},X}$, independent of the candidate construction. For each j, we let $\rho_j(x) = \mathrm{rad}_{\mathcal{H}_j}\{\widehat{\mathcal{C}}_j(x)\}$ and define the size estimator

$$\widehat{S}_j = \sup_{f\in\mathcal{F}} \frac{1}{n_{\mathrm{sel}}} \sum_{i=1}^{n_{\mathrm{sel}}} f(X_i^{\mathrm{sel}})\rho_j(X_i^{\mathrm{sel}}).$$

The identity $dP_{f,X} = f\, dP_{\mathrm{src},X}$ implies that $S_j = \sup_{f\in\mathcal{F}} \mathbb{E}_{P_{\mathrm{src},X}}\{f(X)\rho_j(X)\}$. The following proposition establishes a sufficient condition for the penalty event (S.75) that holds conditionally on the candidate construction.

For the size comparison, we define the product classes

$$\mathcal{G}_j = \{x \mapsto f(x)\rho_j(x) : f \in \mathcal{F}\}, \qquad j = 1, \dots, J,$$

where a countable pointwise dense subclass of $\mathcal{F}$ gives one for each $\mathcal{G}_j$ by multiplication by ρ_j.

Proposition S34. *Suppose $\mathcal{F}$ is pointwise measurable, bounded by B_F, and has VC-subgraph dimension $v_{\mathcal{F}}$. Let $\mathcal{D}^X_{\mathrm{sel}}$ be i.i.d. from $P_{\mathrm{src},X}$ and independent of the candidate construction. Conditional on the candidate sets, suppose every ρ_j is measurable and $\sup_{1\le j\le J}\sup_{x\in\mathcal{X}}\rho_j(x) \le B_\rho$ for a finite bound $B_\rho < \infty$ measurable with respect to the candidate construction. Then, there exists an absolute constant $C > 0$ such that, for every $0 < \delta < 1$, it holds with conditional probability at least $1-\delta$ over the selection sample $\mathcal{D}^X_{\mathrm{sel}}$ that*

$$\left|\widehat{S}_j - S_j\right| \le C B_F B_\rho \sqrt{\frac{1 + (v_{\mathcal{F}}+1)\log(e n_{\mathrm{sel}}) + \log(J/\delta)}{n_{\mathrm{sel}}}} \qquad \textit{for every } j = 1, \dots, J. \quad \text{(S.74)}$$

In particular, the penalty event (S.75) holds with pen_j equal to the right-hand side of (S.74). In addition, if each candidate set $\widehat{\mathcal{C}}_j$ is obtained from a score-threshold rule with threshold $\widehat{q}_j$ and satisfies $\rho_j(x) \le A_j(x)\widehat{q}_j$ pointwise with $\sup_{1\le j\le J}\sup_{x\in\mathcal{X}} A_j(x) \le B_A$, then the bound $\sup_{1\le j\le J}\sup_{x\in\mathcal{X}}\rho_j(x) \le B_\rho$ holds with $B_\rho = B_A \max_{1\le j\le J}\widehat{q}_j$.

By the tower property, the conditional probability statement in Proposition S34 also gives the penalty event (S.75) with unconditional probability at least $1-\delta$ whenever $B_\rho < \infty$ almost surely.

Algorithm 3 Adaptive calibration over image-functional classes

1: Fix candidate classes $\mathcal{H}_1, \dots, \mathcal{H}_J$.
2: **for** $j = 1, \dots, J$ **do**
3: construct $\widehat{\mathcal{C}}_j$ by applying the same calibration rule with $\mathcal{H}_j$;
4: estimate its size by $\widehat{S}_j$ on an independent selection sample.
5: **end for**
6: choose the smallest $\widehat{j} \in \operatorname{argmin}_{1\le j\le J}\{\widehat{S}_j + \mathrm{pen}_j\}$.
7: output $\widehat{\mathcal{C}}_{\widehat{j}}$.

The following theorem compares selection over the finite family $\mathcal{H}_1, \dots, \mathcal{H}_J$ with the best candidate after allowing for error in estimated set size. The selected set satisfies the one-sided validity bound $\Delta^-_{\mathcal{F}}(\widehat{\mathcal{C}}_{\widehat{j}}) \le \varepsilon_n$ when every candidate satisfies this bound at level ε_n. Each candidate is measured using its own functional class.

Theorem S21. *Fix $\delta_1, \delta_2 \in (0,1)$ with $\delta_1 + \delta_2 < 1$. Suppose that J is a deterministic finite integer and that, with probability at least $1-\delta_1$, $\Delta_{\mathcal{F}}^{-}(\widehat{\mathcal{C}}_j) \le \varepsilon_n$ for every $j = 1, \dots, J$. If, with probability at least $1-\delta_2$,*

$$\left|\widehat{S}_j - S_j\right| \le \mathrm{pen}_j \qquad \text{for every } j = 1, \dots, J, \tag{S.75}$$

then, it holds with probability at least $1-\delta_1-\delta_2$ that

$$\Delta_{\mathcal{F}}^{-}(\widehat{\mathcal{C}}_{\widehat{j}}) \le \varepsilon_n \tag{S.76}$$

and

$$\mathcal{S}_{\mathcal{F},\mathcal{H}_{\widehat{j}}}(\widehat{\mathcal{C}}_{\widehat{j}}) \le \inf_{1\le j\le J} \left\{ \mathcal{S}_{\mathcal{F},\mathcal{H}_j}(\widehat{\mathcal{C}}_j) + 2\mathrm{pen}_j \right\}. \tag{S.77}$$

Theorem S21 compares each candidate using the functional class applied to construct that candidate. It does not compare all candidates through one common full image-functional class.

The following proposition establishes the coverage and size bounds after selection from a compact image class. It derives both probability bounds in Theorem S21 from the score construction and sampling conditions.

We let $\mathcal{Y} \subset \mathbb{R}^V$ be nonempty and compact, and let each $\mathcal{H}_j$ be a nonempty finite class of linear functionals. We define $B_{\mathcal{H}} = \max_{j,h\in\mathcal{H}_j} \sup_{y\in\mathcal{Y}} |h(y)|$ and $B_\rho = 2B_{\mathcal{H}}$.

For each candidate, we use a finite measurable score R_j that is continuous in the image argument, and we calibrate its sublevel sets in $\mathcal{Y}$ by (S.30) with deterministic tolerance $\tau_{n,j}$. All scores are fixed jointly before and independently of the i.i.d. source calibration sample. The selection sample is independent of the candidate construction.

We suppose that the fixed class $\mathcal{F}$ is pointwise measurable, bounded by B_F, and has finite VC-subgraph dimension. We use the rate ψ_n in Proposition S22. We define $\varepsilon_n = \max_j\{\tau_{n,j} + \psi_n(\delta_1/J, \mathcal{F})\}$ and take every pen_j to be the right-hand side of (S.74) with $B_\rho = 2B_{\mathcal{H}}$ and confidence parameter δ_2.

Proposition S35. *Suppose the candidate sets are constructed as above. Then each ρ_j is measurable, $0 \le \rho_j \le 2B_{\mathcal{H}}$, and, with probability at least $1-\delta_1-\delta_2$, the selected set satisfies (S.76) and (S.77).*

Proposition S35 obtains a uniform radius bound from the fixed image class. It requires no score density or predictor-accuracy condition. If all tolerances equal $2\psi_n(\delta_1/J, \mathcal{F})$, then $\varepsilon_n = 3\psi_n(\delta_1/J, \mathcal{F})$.

Proof. Proposition S7 gives finite measurable thresholds. We fix a countable dense subset D of $\mathcal{Y}$ and define $d_j(y,c) = \max_{h\in\mathcal{H}_j} |h(y) - h(c)|$. For every $c \in D$, continuity and compactness give

$$\sup_{y\in\widehat{\mathcal{C}}_j(x)} d_j(y,c) = \inf_{m\ge 1} \sup_{y\in D} d_j(y,c)\mathbf{1}\{R_j(x,y) < \widehat{q}_j + m^{-1}\},$$

with zero on the left for an empty set. The right side is measurable, and continuity in c permits the infimum over D. Hence ρ_j is measurable and $0 \le \rho_j \le 2B_{\mathcal{H}}$. Multiplying the countable dense subclass of $\mathcal{F}$ by ρ_j preserves pointwise measurability.

We condition on the candidate scores. Proposition S22 at level δ_1/J, a union bound, and Theorem S9 give $\Delta_{\mathcal{F}}^{-}(\widehat{\mathcal{C}}_j) \le \varepsilon_n$ simultaneously with probability at least $1-\delta_1$. Proposition S34 gives (S.75) with conditional, hence unconditional, probability at least $1-\delta_2$. We apply Theorem S21. This completes the proof. □

The deterministic family $\{\mathcal{H}_{\le j}^{\mathrm{MR}}\}_{0\le j\le J}$ of level truncations from Definition S3 can be used under the conditions of Theorem S21. Proposition S35 verifies these conditions for compact image classes and the stated score and sampling construction. The resulting comparison concerns the truncated-radius criterion at each candidate level.

Corollary S6. *We apply the adaptive rule of this subsection to the candidate family $\mathcal{H}_j = \mathcal{H}_{\le j}^{\mathrm{MR}}$, $0 \le j \le J$, with $\rho_j(x) = \mathrm{rad}_{\mathcal{H}_{\le j}^{\mathrm{MR}}}\{\widehat{\mathcal{C}}_j(x)\} \le A_j(x)\widehat{q}_j$ and $B_\rho = B_A \max_{0\le j\le J} \widehat{q}_j < \infty$. Under the conditions of Theorem S21 and Proposition S34, it holds with probability at least $1-\delta_1-\delta_2$ that the selected set $\widehat{\mathcal{C}}_{\widehat{j}}$ satisfies $\Delta_{\mathcal{F}}^{-}(\widehat{\mathcal{C}}_{\widehat{j}}) \le \varepsilon_n$ and*

$$\mathcal{S}_{\mathcal{F},\mathcal{H}_{\le \widehat{j}}^{\mathrm{MR}}}(\widehat{\mathcal{C}}_{\widehat{j}}) \le \inf_{0\le j\le J} \left\{ \mathcal{S}_{\mathcal{F},\mathcal{H}_{\le j}^{\mathrm{MR}}}(\widehat{\mathcal{C}}_j) + 2\,\mathrm{pen}_j \right\}.$$

Proof. Theorem S21 gives the validity and oracle bounds on the intersection of the candidate-validity and penalty events. Proposition S34 establishes the penalty bound for the stated candidates and penalties. Substituting $\mathcal{H}_j = \mathcal{H}^{\mathrm{MR}}_{\le j}$ proves the corollary. This completes the proof. □

When $\mathcal{Y} = \mathcal{Y}_R$ as defined in (S.24), Lemma S3 bounds the full-class radius of the selected set by its selected truncated radius plus $2R2^{-s\widehat{J}}$. The selection criterion does not contain this remainder, which depends on the unknown regularity. Corollary S6 hence does not establish adaptation to unknown (s, R).

The family has $J + 1 = \log_4 V + 1$ deterministic candidates, and its selection penalty depends on $\log\log V$. The spatial approximation budget in Theorem S12(i) is independent of V for a fixed regularity class and a cutoff chosen using (s, R). A full-class adaptive guarantee would require a selection criterion that controls the truncation remainder and the sampling penalty.

S6.12 Separation between calibration complexity and teacher quality

Calibration complexity bounds the validity error. Predictor error controls the paired radius comparison in the following theorem.

Theorem S22. *Fix $0 < \delta < 1$, define $e_{n,\delta} = e_n(\delta/3)$ and $\psi_{n,\delta} = \psi_n(\delta/3, \mathcal{F})$, and define $\tau_n = 2\psi_{n,\delta}$. Let $\widehat{q}^\circ$ be the attained measurable version of (S.30) computed with score R° and tolerance τ_n, and let $\widehat{\mathcal{C}}^\circ(x) = \{y : R^\circ(x, y) \le \widehat{q}^\circ\}$. Suppose that $\mathcal{F}$ is pointwise measurable, that (S.33), the empirical-process bound (S.34) for both R_n and R° with bound $\psi_n(\cdot, \mathcal{F})$, and the radial Lipschitz property in Proposition S12 for both score ball families with common bound A hold. Then there exists an absolute constant $C < \infty$ such that, with probability at least $1 - \delta$,*

$$\Delta^-_{\mathcal{F}}(\widehat{\mathcal{C}}^\circ) \le C\psi_{n,\delta}, \qquad \Delta^-_{\mathcal{F}}(\widehat{\mathcal{C}}) \le C\psi_{n,\delta}, \tag{S.78}$$

and

$$\left|\mathcal{S}_{\mathcal{F},\mathcal{H}}(\widehat{\mathcal{C}}) - \mathcal{S}_{\mathcal{F},\mathcal{H}}(\widehat{\mathcal{C}}^\circ)\right| \le A_{\mathcal{F}} C\{e_{n,\delta} + \psi_{n,\delta}\}. \tag{S.79}$$

Theorem S22 shows that the validity error is bounded by the calibration complexity. Proposition S2 gives the sharper radius comparison $2A_{\mathcal{F}}e_{n,\delta}$. The calibration term in (S.79) is unnecessary for this paired comparison.

S6.13 Additional structural consequences

Uniform protection against every covariate shift restricts the size of a prediction set when an image functional is unbounded.

For a set-valued rule $\widehat{\mathcal{C}}$, we let $\mathrm{Gr}(\widehat{\mathcal{C}}) = \{(x, y) \in \mathcal{X} \times \mathcal{Y} : y \in \widehat{\mathcal{C}}(x)\}$ denote its graph. We let $\mathbb{E}^*$ denote outer expectation. When the radius is measurable, this coincides with the standard expectation.

For this result, $\mathcal{X}$ and $\mathcal{Y}$ are standard Borel spaces. The same prediction rule is used for every source distribution and protected weight. It is a function of $N < \infty$ i.i.d. source observations and auxiliary randomization in a standard Borel space, independent of the observations and with a fixed distribution that does not depend on the source distribution. Its graph is jointly analytic in the observations, randomization, covariate and image. The distribution-free premise below ranges over all source distributions on $\mathcal{X} \times \mathcal{Y}$, including local changes in the conditional response distribution.

We let $\mathcal{F}_{\mathrm{all}}$ be the class of measurable functions $f : \mathcal{X} \to [0, \infty)$ with $\mathbb{E}_{P_{\mathrm{src},X}} f(X) = 1$. The following theorem shows that an unbounded image functional prevents finite expected radius under uniform distribution-free coverage over this class.

Theorem S23. *Suppose some measurable $h_0 \in \mathcal{H}$ has an unbounded range. If the prediction rule specified above satisfies $P_f\{Y \in \widehat{\mathcal{C}}(X)\} \ge 1 - \alpha$ for all $f \in \mathcal{F}_{\mathrm{all}}$ under every source distribution and its covariate shifts, then, for every source distribution with a nonatomic covariate marginal,* $\mathbb{E}^*\left[\mathrm{rad}_{\mathcal{H}}\{\widehat{\mathcal{C}}(X)\}\right] = \infty$.

Theorem S23 requires neither a density for the response nor a nondegenerate conditional response distribution. The obstruction follows from the unbounded functional and protection against every shift. It does not apply to the bounded image box with bounded spatial functionals.

S7 NEURAL SUPER-RESOLUTION STUDY

The neural study asks how informative the exact ranges are when a fitted neural ensemble provides the returned images, how much the exact calculation tightens separate intervals, and what alternative guarantees deliver at the same data and predictor. It has two designs with the same predictor, pools, and replications. The initial design certifies all $21{,}845$ dyadic averages down to single pixels. Its design, including every comparison and metric, was recorded before any predictor output was computed. Two amendments are recorded with it. Before any output, the number of training epochs was reduced from 30 to 20 for computing time. After a preliminary computational check with three members and five replications, we added a center constructed by successive projections of the regional and pixel midpoints. We report both this construction and the constructive fill in the proof. In the initial design, the $16{,}384$ pixel residuals set the threshold, and the radius was close to its largest possible value $1/2$. The widest ranges were nearly unrestricted, which limits their information about regional averages. We then specified the regional design of Section 4.1, which certifies the $5{,}461$ averages over squares with side at least two pixels. It was recorded after the results of the initial design had been seen, and each of its choices was made on the validation pool before any regional computation on calibration or target images. It reuses the fitted ensemble, the pools, α, the sample counts, and the replication seeds, which pairs the two designs replication by replication. Each design was evaluated once, and we report both in full.

S7.1 DATA, POOLS, AND SAMPLING

PneumoniaMNIST contains $5{,}856$ pediatric chest radiographs labeled normal or pneumonia (Kermany et al., 2018; Yang et al., 2023). We use its 224×224 release, resize every image to 128×128 by antialiased bilinear interpolation, and keep intensities in $[0, 1]$. The observation Z is the 32×32 array of averages over 4×4 pixel blocks, and the group G is the diagnostic label. A fixed seed divides the official training images into a predictor pool of $3{,}000$ images, a validation pool of 300 images, and a remainder of $1{,}408$ images. The predictor pool is used only to fit the ensemble. The validation pool is used only for early stopping and for the choice of λ in both designs. In the controlled design, calibration and target images are drawn without replacement within each group from the union of the remainder and the official validation and test images, which contains 733 normal and $1{,}823$ pneumonia images. Each replication draws 240 normal and 160 pneumonia calibration images and 100 normal and 300 pneumonia target images. In the official design, calibration images come from the remainder and the official validation images, and target images come from the official test images, with the same counts. The MedMNIST release does not include patient identifiers, and the pools are divided by image. Images of one patient can hence occur in both the predictor pool and a calibration or target sample, which does not affect exchangeability within groups in the controlled design. The ensemble is fitted once, and the replications resample calibration and target images. The reported variability hence excludes refitting the predictor. The official test images form a separate pool whose within-group distribution can differ from the calibration pool. Proposition 3 bounds the effects of covariate and conditional-image changes through $\kappa(r, \mathcal{F})$ and ζ, respectively.

S7.2 PREDICTOR

Each ensemble member is a U-Net (Ronneberger et al., 2015) with four resolutions of 32, 64, 128, and 256 channels, group normalization, and 1.93 million parameters. It receives the nearest-neighbour upsampled observation and a constant channel equal to the group indicator, and it predicts a residual that is added to the upsampled observation. Outputs are clipped to $[0, 1]$. Members differ only in their random initialization and data order, with seeds 0 to 4. Each member is trained for 20 epochs with the ℓ_1 loss, Adam, a one-cycle learning-rate schedule with maximum rate 2×10^{-3}, and batch size 16, and the state with the smallest validation ℓ_1 error is kept. The five members reached validation ℓ_1 errors between 0.01089 and 0.01092. The $K = 5$ member outputs are the returned images. Their sample mean and standard deviation, with divisor K, enter (3). The predictor and validation pools determine fitting and checkpoint selection. Theorem 1 gives group coverage conditional on the fitted ensemble and the score selected before calibration.

S7.3 SPATIAL FUNCTIONALS, RIDGE, AND EXACT RANGES

The regional class contains the averages over all dyadic squares with sides from 2 to 128 pixels, for 5,461 functionals. Averages over squares with side at least four pixels are fixed by Z, and their score constraints are satisfied by every candidate or by none. Within each observed block, the four 2×2 regions form the single unknown level of Proposition 2, which gives their exact ranges over the set intersected with the observation. The accepted reconstruction attains the optimal error $r_Z(x)$ of Proposition S1, computed by (S.2). Among the optimal accepted centers, we report the one closest to the range midpoints in Euclidean distance. In each block, it is the projection of the midpoints onto $\{c : \max(L_j, U_j - r) \le c_j \le \min(U_j, L_j + r),\ \sum_j c_j = 4z_b\}$ with $r = r_Z(x)$. The certificate concerns regional averages. The full-resolution image in Figure 1 adjusts the ensemble-mean pixels of each 2×2 region to the accepted regional average by Euclidean projection within $[0, 1]$.

The initial class contains the averages over all dyadic squares, from the whole image to single pixels, for 21,845 functionals. Within each observed block, the four 2×2 regions and sixteen pixels form a tree with two unknown levels. Theorem S1 gives their exact ranges over the set intersected with the observation. The accepted reconstruction uses the bisection after that theorem with tolerance 10^{-4}, and the projected center first projects the regional midpoints and then the pixel midpoints conditional on the chosen regional means. In both classes, the separate intervals report $[\max\{0, \widehat{\mu}_h(x) - q\{\widehat{\sigma}_h(x) + \lambda\}\}, \min\{1, \widehat{\mu}_h(x) + q\{\widehat{\sigma}_h(x) + \lambda\}\}]$ for each functional without the observation or the nested constraints. Both reports describe the same calibrated set and have the same coverage.

In both designs, λ minimizes the mean exact half-width at the finest unknown level on the validation pool. We split this pool into two group-stratified halves, calibrate the group rule on one half, and measure the criterion on the other. The criterion averages both directions, and ties go to the larger value. In the initial design, the grid $\{0.0025, 0.005, 0.01, 0.02, 0.04, 0.08\}$ and the pixel half-width gave $\lambda = 0.0025$, the lower end of the grid (Table 5). The regional design uses the 2×2 half-width and extends the grid on both sides to the doubling grid from 0.000625 to 0.32. Table 4 reports the criterion, the radius, and the coverage, and $\lambda = \infty$ denotes the constant scale. The criterion is again smallest at $\lambda = 0.0025$, now inside the grid, with half-width 0.105 and radius 0.354. The radius decreases as λ grows and equals 0.130 for the constant scale. The choice of λ hence trades the typical range against the worst-case range. A radius criterion on the same halves selects the constant scale, and we report the proposed set with both scales. The registered design lists the constant scale as a comparison, and the radius criterion was stated after the evaluation. On the validation halves, no set was empty, and the observed coarser averages never attained the largest standardized residual.

Table 4: Choice of the ridge for the regional design on the validation pool.

λ	0.000625	0.00125	0.0025	0.005	0.01	0.02	0.04	0.08	0.16	0.32	∞
2×2 half-width	0.129	0.112	0.105	0.106	0.108	0.114	0.117	0.121	0.125	0.126	0.129
Radius	0.464	0.420	0.354	0.285	0.220	0.182	0.153	0.141	0.135	0.132	0.130
Coverage	0.890	0.900	0.903	0.907	0.903	0.897	0.887	0.884	0.890	0.890	0.894

Table 5: Choice of the ridge for the initial design on the validation pool.

λ	0.0025	0.005	0.01	0.02	0.04	0.08
Pixel half-width	0.183	0.188	0.189	0.199	0.219	0.226
Coverage	0.903	0.910	0.900	0.893	0.913	0.907

We checked the implementation against linear programs. In 200 random blocks with random pixel and region intervals and a random observed mean, the recursion reproduced every linear-programming minimum and maximum and every infeasibility, with errors below 10^{-8}. In 100 further blocks, the bisection reproduced the smallest worst-case error of an accepted image computed by linear programming, with errors below 2×10^{-6}. For the regional class, in 180 random blocks, the ranges of Proposition 2 and the error (S.2) reproduced the linear-programming ranges, the one infeasible block, and the smallest worst-case error of an accepted center, with errors below 10^{-8}.

S7.4 COMPARISONS

All comparisons use the same ensemble outputs. The separate intervals describe the proposed set without the exact calculation. The constant-scale score replaces $\widehat{\sigma}_h(x) + \lambda$ by one, which removes the dependence on the ensemble spread, and it uses corrected group thresholds and exact ranges. The

set of the standardized score approaches this set as λ grows. The pooled rule uses one corrected threshold for all calibration images. The largest group threshold is the common threshold of Section 4. The uncalibrated ensemble band uses the standardized score at the normal maximum quantile for the number of functionals, $q = 4.27$ for 5,461 and $q = 4.57$ for 21,845, without labeled calibration. Risk control calibrates intervals $\widehat{\mu}_h(x) \pm t\{\widehat{\sigma}_h(x) + \lambda\}$ at the finest unknown level within each group. Under exchangeability within groups, it ensures that the expected fraction of uncovered 2×2 regions, or of uncovered pixels in the initial design, is at most 0.1, by conformal risk control (Angelopoulos et al., 2024) on a grid of step 0.01 for t. If no grid value satisfies the corrected empirical-risk bound, the intervals equal $[0, 1]$. For pixels, this loss is the miscoverage fraction controlled for image-to-image regression by Angelopoulos et al. (2022) and Teneggi et al. (2023). For risk control, the reported coverage is the fraction of target images whose regions, or pixels, are all covered, and the uncovered fraction is reported in the text.

S7.5 Results

Tables 6 and 7 report the regional design, and Tables 9 and 10 report the initial design, as means and standard deviations over 100 replications. Numbers in parentheses represent standard deviations. The radius is the largest exact half-range of a 2×2 region in the regional design and of a pixel in the initial design, averaged over target images with empty sets counted as zero. The half-widths are averages over regions or pixels and over target images with nonempty sets. For risk control, the radius and half-width describe its intervals. The group columns give coverage for normal and pneumonia images, and Tables 8 and 11 compare the accepted reconstructions with the ensemble mean. Their final column is the fraction of covered images whose reconstruction error is at most the certified error r.

In the controlled regional design, the standardized set contains 0.901 of target images, with group coverage 0.905 and 0.900, and no set is empty. The constant-scale set contains 0.899, with group coverage 0.905 and 0.897. The pooled threshold covers 0.895 and 0.913, and the largest group threshold gives coverage 0.919 and mean 2×2 half-width 0.108. The uncalibrated band uses $q = 4.27$ and contains no target image. Its sets are empty for a fraction 0.002 of target images, because the ensemble moments are incompatible with the observed coarse image at that threshold. Regionwise risk control leaves 9.5% of regions uncovered, and none of its target images has all regions covered. The exact ranges have mean half-width 0.1040 against 0.1042 for separate intervals, and the radius differs by less than 0.001. Relative to the initial design on the same replications, the mean group thresholds fall from 37.4 and 30.9 to 21.3 and 20.3. The mean 2×2 half-width falls from 0.163 to 0.104, by between 29% and 43% in individual replications. The following summaries were computed after the evaluation from the same calibration and target samples. With single pixels also certified, the largest exact 2×2 half-range is on average 0.437 and 0.249 for the standardized and constant scales. The regional design lowers it by between 16% and 26% and between 35% and 50% in individual replications, and by 21% and 44% on average in the official split. The certified error of the optimal accepted center is on average 1.01 times the radius for the standardized scale and 1.02 times for the constant scale, and every covered target image satisfies the bound for both scales. The accepted centers match the observed block averages exactly. For the standardized scale, their largest and mean absolute regional errors, 0.095 and 0.0082, are close to 0.087 and 0.0081 for the ensemble mean, and for the constant scale they are 0.087 and 0.0081. A conformal interval at level $1 - \alpha/m$ for each of the $m = 4{,}096$ unknown regions is narrower than $[0, 1]$ only when the group has at least $m/\alpha - 1 = 40{,}959$ calibration images. The adjusted full-resolution image has mean absolute pixel error 0.0105, against 0.0103 for the ensemble mean. In the official split, the standardized set contains 0.921 of target images, with group coverage 0.933 and 0.917, and the constant scale covers 0.874 of normal images with certified error 0.143 on average.

Table 6: Regional design of the neural study, controlled design.

Method	Coverage	Normal	Pneumonia	Radius	2×2 half-width
Standardized: exact ranges	0.901 (0.027)	0.905 (0.038)	0.900 (0.032)	0.348 (0.013)	0.104 (0.005)
Standardized: separate intervals	0.901 (0.027)	0.905 (0.038)	0.900 (0.032)	0.348 (0.013)	0.104 (0.005)
Constant scale: exact ranges	0.899 (0.027)	0.905 (0.035)	0.897 (0.031)	0.140 (0.011)	0.139 (0.011)
Constant scale: separate intervals	0.899 (0.027)	0.905 (0.035)	0.897 (0.031)	0.140 (0.011)	0.139 (0.011)
One pooled threshold	0.909 (0.022)	0.895 (0.040)	0.913 (0.020)	0.352 (0.009)	0.105 (0.003)
Largest group threshold	0.919 (0.021)	0.909 (0.038)	0.923 (0.019)	0.359 (0.011)	0.108 (0.004)
Uncalibrated ensemble	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.078 (0.001)	0.022 (0.000)
Regionwise risk control	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.064 (0.001)	0.018 (0.000)

Table 7: Regional design of the neural study, official split.

Method	Coverage	Normal	Pneumonia	Radius	2×2 half-width
Standardized: exact ranges	0.921 (0.013)	0.933 (0.020)	0.917 (0.015)	0.346 (0.011)	0.105 (0.004)
Standardized: separate intervals	0.921 (0.013)	0.933 (0.020)	0.917 (0.015)	0.346 (0.011)	0.105 (0.004)
Constant scale: exact ranges	0.889 (0.018)	0.874 (0.029)	0.894 (0.021)	0.140 (0.009)	0.139 (0.009)
Constant scale: separate intervals	0.889 (0.018)	0.874 (0.029)	0.894 (0.021)	0.140 (0.009)	0.139 (0.009)
One pooled threshold	0.926 (0.009)	0.928 (0.022)	0.925 (0.009)	0.350 (0.007)	0.106 (0.003)
Largest group threshold	0.931 (0.010)	0.935 (0.020)	0.930 (0.010)	0.357 (0.008)	0.109 (0.003)
Uncalibrated ensemble	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.077 (0.001)	0.021 (0.000)
Regionwise risk control	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.064 (0.001)	0.018 (0.000)

Table 8: Reconstructions in the regional design: certified error, errors of 2×2 averages, and agreement with the observation.

Design	Reconstruction	Certified r	Largest error	Mean absolute error	Largest mismatch	Covered images within r
Controlled	Accepted, standardized	0.351 (0.013)	0.095 (0.002)	0.0082 (0.0001)	0.0000	1.000
	Accepted, constant scale	0.143 (0.011)	0.087 (0.002)	0.0081 (0.0001)	0.0000	1.000
	Ensemble mean	–	0.087 (0.002)	0.0081 (0.0001)	0.0031 (0.0001)	–
Official split	Accepted, standardized	0.350 (0.011)	0.093 (0.002)	0.0081 (0.0001)	0.0000	1.000
	Accepted, constant scale	0.143 (0.009)	0.086 (0.001)	0.0080 (0.0001)	0.0000	1.000
	Ensemble mean	–	0.086 (0.001)	0.0080 (0.0001)	0.0030 (0.0001)	–

In the controlled initial design, the proposed set contains 0.899 of target images, with group coverage 0.901 and 0.898, and no set is empty. The largest group threshold gives coverage 0.937 and mean pixel half-width 0.202. The uncalibrated band uses $q = 4.57$ and contains no target image. Its sets are empty for a fraction 0.009 of target images, because the ensemble moments are incompatible with the observed coarse image at that threshold. The exact ranges have mean pixel half-width 0.179 against 0.179 for separate intervals and mean 2×2 half-width 0.163 against 0.164, and the radius is identical to three decimals. The certified error of the optimal accepted center is on average 1.08 times the radius. The successive projections give an accepted center within the stated tolerance of the optimal certified error. Its largest pixel error is 0.215 on average. The constructive fill in the proof of Theorem S1 attains the same certified error and has mean absolute error 0.1479. Every covered target image satisfies the certified bound for both constructions. In the official split, the proposed set contains 0.915 of target images, with group coverage 0.917 and 0.915, and the pooled threshold covers 0.898 and 0.941. The official test pool did not lower group coverage in this study, unlike four of the five cohorts of Appendix S8.7.

Table 9: Initial design of the neural study, controlled design.

Method	Coverage	Normal	Pneumonia	Radius	Pixel half-width	2×2 half-width
Proposed: exact ranges	0.899 (0.027)	0.901 (0.036)	0.898 (0.032)	0.488 (0.005)	0.179 (0.009)	0.163 (0.008)
Proposed: separate intervals	0.899 (0.027)	0.901 (0.036)	0.898 (0.032)	0.488 (0.005)	0.179 (0.009)	0.164 (0.008)
Constant scale	0.899 (0.024)	0.911 (0.034)	0.895 (0.029)	0.249 (0.015)	0.241 (0.014)	0.241 (0.014)
One pooled threshold	0.921 (0.020)	0.879 (0.039)	0.935 (0.018)	0.494 (0.003)	0.189 (0.010)	0.172 (0.009)
Largest group threshold	0.937 (0.018)	0.901 (0.036)	0.949 (0.016)	0.496 (0.002)	0.202 (0.011)	0.185 (0.010)
Uncalibrated ensemble	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.124 (0.002)	0.026 (0.000)	0.023 (0.000)
Pixelwise risk control	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	–	0.023 (0.000)	–

Table 10: Initial design of the neural study, official split.

Method	Coverage	Normal	Pneumonia	Radius	Pixel half-width	2×2 half-width
Proposed: exact ranges	0.915 (0.016)	0.917 (0.026)	0.915 (0.021)	0.490 (0.004)	0.180 (0.008)	0.164 (0.007)
Proposed: separate intervals	0.915 (0.016)	0.917 (0.026)	0.915 (0.021)	0.490 (0.004)	0.180 (0.008)	0.164 (0.007)
Constant scale	0.902 (0.017)	0.906 (0.027)	0.901 (0.020)	0.251 (0.013)	0.244 (0.013)	0.243 (0.013)
One pooled threshold	0.930 (0.016)	0.898 (0.031)	0.941 (0.014)	0.496 (0.002)	0.189 (0.008)	0.172 (0.008)
Largest group threshold	0.950 (0.014)	0.918 (0.027)	0.960 (0.013)	0.498 (0.001)	0.203 (0.010)	0.185 (0.009)
Uncalibrated ensemble	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.123 (0.001)	0.025 (0.000)	0.023 (0.000)
Pixelwise risk control	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	–	0.023 (0.000)	–

Table 11: Reconstructions in the initial design: largest and mean absolute pixel errors.

Design	Reconstruction	Largest error	Mean absolute error	Covered images within r
Controlled	Accepted, projected	0.215 (0.008)	0.0125 (0.0004)	1.000
	Accepted, constructive fill	0.409 (0.010)	0.1479 (0.0065)	1.000
	Ensemble mean	0.144 (0.003)	0.0103 (0.0001)	–
Official split	Accepted, projected	0.209 (0.007)	0.0123 (0.0003)	1.000
	Accepted, constructive fill	0.406 (0.008)	0.1496 (0.0056)	1.000
	Ensemble mean	0.140 (0.002)	0.0102 (0.0001)	–

S8 MEDICAL IMAGE RECONSTRUCTION AND DIRECT CALIBRATION COMPARISONS

The medical experiments predict fine image features from a coarse observation and observed clinical status. They evaluate uncertainty for reconstruction, not the accuracy of disease classification. The controlled prevalence experiments and the separate source and test pools of Appendix S8.7 answer different questions and are reported separately.

S8.1 OBSERVED INFORMATION AND RESPONSE

Let Z be the 4×4 array of nonoverlapping block averages of the fine image. The query consists of Z and the observed clinical group G. A queried predictor receives only this information. The response is the fine retinal or chest image. Calibration and test responses are used only to evaluate conformity scores and empirical coverage, respectively.

The image-functional class contains averages over the 2×2, 4×4, and 8×8 dyadic partitions, giving $4 + 16 + 64 = 84$ functionals. The 8×8 partition includes features that are not determined by Z. Two images can have the same 4×4 observation while differing within each observed block. Thus, passing Z to the predictor does not pass the fine response or all evaluated features.

The primary design fixes group counts in source and target samples and selects distinct images within each group. Within each group, a single draw without replacement gives the source images, and the remaining drawn images form the target sample. The source sample is then split equally within each group into training and calibration folds. No image occurs in two folds of one replication. Conditional on the benchmark and the training observations, calibration and test images remain exchangeable within each group. The rank proof of Theorem 1 thus applies to this finite-population design.

The same finite benchmark is sampled repeatedly with independent random seeds. The reported variability concerns this repeated-sampling experiment; it is not variability across 100 independently collected patient populations. Some patients contribute several images to ChestMNIST and RetinaMNIST, and the controlled design samples images. Images of one patient can hence occur in both the training fold and the target sample. This overlap leaves calibration and target images exchangeable within groups, but it can make the predictor more accurate on these images than on new patients.

S8.2 COHORTS AND SAMPLE SIZES

RetinaMNIST contains 1,600 retinal fundus images in five diabetic-retinopathy grades (Yang et al., 2023; Liu et al., 2022). The 64×64 analysis uses the green channel of the corresponding MedMNIST distribution. The 128×128 analysis resamples the green channel of its 224×224 distribution. The source counts for grades zero through four are $450, 100, 160, 130, 60$, and the target counts are $90, 35, 90, 65, 20$. Each replication thus uses 900 source and 300 target images.

ChestMNIST contains 112,120 radiographs derived from NIH ChestX-ray14 (Wang et al., 2017; Yang et al., 2023). We pad the 28×28 images to a 32×32 grid by edge replication. The observed group distinguishes no finding from at least one of the fourteen text-mined findings. Source counts are 1,600 and 800, and target counts are 400 and 1,200.

The NLM Indiana University collection is distributed through Open-i (Demner-Fushman et al., 2016). We use 3,838 decoded images with matched reports and select the first available image in filename order from each report. This gives 2,893 report-level observations, resampled to 64×64. Normal status means that the report contains the MeSH term "normal" and at most two MeSH terms; the remaining reports form the abnormal group. Source counts are 400 normal and 400 abnormal, and target counts are 100 normal and 300 abnormal. The extraction is a benchmark subset and is not presented as a representative sample of a hospital population.

The public MAE analysis uses a fixed ChestMNIST subset of 3,000 images, selected by group and a seed before predictor queries. The subset contains 1,800 images from the official source pool and 1,200 from the official test pool. Each replication uses source counts $400, 200$ and target counts $100, 300$. The checkpoint is the public chest-image masked autoencoder distributed at https://huggingface.co/adelelsayed1991/mae, revision `a35bb09dbd4f6e1ebf3d34f502e393d8f2787cde`. Its repository specifies a CheXpert training configuration. We use the public checkpoint and do not

fit its parameters. The masked-autoencoder architecture follows the reconstruction approach of He et al. (2022).

S8.3 Queried predictors

The non-MAE analyses use a controlled image predictor fitted only on the source training fold. Its design contains the observed group indicators and the sixteen coarse block averages. Predictor fitting standardizes these covariates using training observations and uses ridge image regression with penalty 10^{-3}. The predictor returns six images per query by adding independent Gaussian pixel noise with standard deviation 0.04 to its fitted image and clipping pixel intensities to $[0, 1]$. Only the returned images are used to calculate score moments. This controlled predictor isolates the effects of the calibration rule and the image score.

The MAE receives the upsampled coarse observation, never the fine response. Intensities are transformed using the repository's mean 0.5 and standard deviation 0.5. The query image is resized to 384×384, and two independently masked reconstructions are returned at each query. Visible patches contain the coarse observation, not the unobserved fine image. Reconstructions are mapped back to $[0, 1]$ and resized to 32×32. Their regional means and standard deviations use divisor K. Predictor outputs are computed once for each benchmark image with a fixed seed and then reused over the calibration splits. The repeated-sampling analysis is conditional on these returned images.

All analyses use $\alpha = 0.1$ and ridge $\lambda = 0.05$ in (3). Preprocessing and predictor fitting use no calibration or test response. The acquisition covariates used by a comparison calibrator are brightness, contrast, and centrally weighted intensity of the coarse observation, standardized on the training fold.

The 20 coarser regional averages are determined by the observed image and have zero conditional variance. Lemma S6 and the Kullback–Leibler bound in (S.86) allow zero variances; Proposition S27 gives the corresponding Wasserstein bound. These results do not establish a numerical distribution-error bound for the fitted medical predictors. Proposition S4 applies to their exact regional radii when both compared intersections are nonempty.

S8.4 Comparison procedures and size measurement

The proposed procedure is Algorithm 1. The common-threshold comparison uses $\max_g \widehat{q}_g$ with the same predictor and image score. Marginal split conformal uses the corrected global order statistic. The ratio-weighted comparison uses the target-to-source empirical group-frequency ratio with additive count smoothing 0.5, estimated from source training labels and target covariates. It is not described as oracle-weighted. It uses the weighted empirical quantile of the calibration scores without the test-point mass of Tibshirani et al. (2019), and the kernel-weighted comparison below does the same. A kernel-weighted comparison uses a Gaussian kernel with bandwidth one on the three standardized acquisition summaries.

The two direct Kandinsky comparisons minimize the pinball loss for the 0.9 score quantile, as in Algorithms 1–2 of Bairaktari et al. (2025). The first basis contains only clinical-group indicators. The second adds the three acquisition summaries. Calibration and test scores receive independent uniform perturbations on $[-10^{-8}, 10^{-8}]$ to avoid score ties. The linear programs are solved by the HiGHS implementation in SciPy. Both comparisons use the identical image predictor and maximum score. The group-indicator comparison differs from (4) only in the finite-sample order correction and tie selection. Any difference between them is not evidence for a new conditional-calibration principle.

The group-L_2 comparison uses the Euclidean norm of the vector of standardized regional residuals, with the same group-specific finite-sample calibration. It measures the effect of the score on the reported regional size bound. The plug-in comparison uses the normal maximum quantile without labeled calibration. The static-mixture comparison solves the moment equation on the calibration fold and samples one threshold for the deployed set. It is different from the candidate-augmented solver of Theorem S7.

For regional averages of images in $[0, 1]^V$ and a nonnegative threshold $q(x)$, a simple radius upper bound is

$$r(x) = \min\left\{\frac{1}{2}, q(x) \max_{h \in \mathcal{H}} \{\widehat{\sigma}_h(x) + \lambda\}\right\}. \tag{S.80}$$

A negative threshold gives an empty set with radius zero. The fitted mean image gives the second bound in (S.80), and the constant image $1/2$ gives the first. We improve this calculation by intersecting the candidate image set with the observed coarse averages. Proposition 2 then gives the exact regional radius for the maximum-score comparisons. For the group-L_2 score, the same coordinate-range calculation gives an upper bound, identified as such in the tables. Empty intersections have radius zero and are reported separately. An ordering based on the L_2 upper bound does not imply the same ordering for its exact radius.

For each replication, we record pooled coverage, coverage in every target group, maximum group undercoverage, maximum absolute group deviation, and the mean radius from Proposition 2. Tables report averages over 100 replications. Numbers in parentheses represent standard deviations. The supplementary interval comparisons resample paired replication differences; they quantify variation over the prespecified repeated-sampling experiment and not uncertainty across institutions.

Table 12: Retina 64×64 reconstruction under controlled prevalence changes.

Method	Coverage	Max. under.	Max. dev.	Radius	Empty
Proposed	0.906 (0.020)	0.065 (0.053)	0.091 (0.039)	0.128 (0.004)	0.000
Common threshold	0.967 (0.016)	0.021 (0.041)	0.094 (0.016)	0.160 (0.013)	0.000
Kandinsky: groups	0.893 (0.020)	0.093 (0.066)	0.104 (0.057)	0.124 (0.004)	0.000
Kandinsky: linear	0.891 (0.023)	0.090 (0.061)	0.102 (0.052)	0.122 (0.004)	0.000
Group-L_2	0.910 (0.018)	0.056 (0.056)	0.090 (0.042)	0.309 (0.008)	0.000
Marginal	0.895 (0.021)	0.116 (0.068)	0.120 (0.063)	0.124 (0.004)	0.000
Ratio-weighted	0.901 (0.021)	0.108 (0.067)	0.114 (0.061)	0.126 (0.005)	0.000
Kernel-weighted	0.895 (0.020)	0.105 (0.064)	0.109 (0.060)	0.124 (0.004)	0.000
Plug-in predictor	0.988 (0.006)	0.002 (0.011)	0.100 (0.002)	0.185 (0.000)	0.000
Static mixture	0.897 (0.299)	0.091 (0.271)	0.180 (0.241)	0.207 (0.069)	0.026

Table 13: Retina 64×64 coverage within observed groups.

Method	Grade 0	Grade 1	Grade 2	Grade 3	Grade 4
Proposed	0.908 (0.030)	0.907 (0.065)	0.903 (0.043)	0.909 (0.051)	0.904 (0.082)
Common threshold	0.979 (0.017)	0.950 (0.037)	0.974 (0.019)	0.964 (0.025)	0.917 (0.070)
Kandinsky: groups	0.904 (0.032)	0.890 (0.066)	0.891 (0.046)	0.892 (0.060)	0.868 (0.094)
Kandinsky: linear	0.898 (0.035)	0.891 (0.061)	0.891 (0.048)	0.889 (0.059)	0.875 (0.093)
Group-L_2	0.910 (0.033)	0.904 (0.057)	0.910 (0.037)	0.913 (0.047)	0.915 (0.083)
Marginal	0.932 (0.025)	0.857 (0.056)	0.906 (0.032)	0.880 (0.040)	0.803 (0.084)
Ratio-weighted	0.935 (0.026)	0.865 (0.054)	0.912 (0.031)	0.887 (0.039)	0.810 (0.083)
Kernel-weighted	0.921 (0.028)	0.865 (0.054)	0.910 (0.030)	0.881 (0.040)	0.818 (0.082)
Plug-in predictor	0.995 (0.007)	0.983 (0.020)	0.989 (0.010)	0.989 (0.011)	0.965 (0.041)
Static mixture	0.899 (0.298)	0.897 (0.301)	0.897 (0.300)	0.894 (0.299)	0.889 (0.299)

The maximum-score radii in Table 12 use Proposition 2; the group-L_2 radius is an upper bound. Table 13 reports each group separately.

Table 14: Retina 128×128 reconstruction under controlled prevalence changes.

Method	Coverage	Max. under.	Max. dev.	Radius	Empty
Proposed	0.906 (0.021)	0.063 (0.056)	0.092 (0.042)	0.124 (0.004)	0.000
Common threshold	0.967 (0.017)	0.021 (0.040)	0.094 (0.016)	0.156 (0.013)	0.000
Kandinsky: groups	0.893 (0.021)	0.095 (0.066)	0.107 (0.058)	0.121 (0.004)	0.000
Kandinsky: linear	0.891 (0.022)	0.092 (0.058)	0.102 (0.050)	0.118 (0.004)	0.000
Group-L_2	0.911 (0.018)	0.057 (0.058)	0.089 (0.045)	0.303 (0.008)	0.000
Marginal	0.894 (0.021)	0.122 (0.073)	0.126 (0.068)	0.121 (0.004)	0.000
Ratio-weighted	0.899 (0.021)	0.110 (0.068)	0.115 (0.063)	0.123 (0.004)	0.000
Kernel-weighted	0.895 (0.020)	0.109 (0.069)	0.113 (0.065)	0.121 (0.004)	0.000
Plug-in predictor	0.984 (0.006)	0.004 (0.016)	0.099 (0.003)	0.173 (0.000)	0.000
Static mixture	0.907 (0.285)	0.082 (0.259)	0.172 (0.230)	0.206 (0.064)	0.020

Table 15: Retina 128×128 coverage within observed groups.

Method	Grade 0	Grade 1	Grade 2	Grade 3	Grade 4
Proposed	0.909 (0.030)	0.907 (0.062)	0.901 (0.043)	0.907 (0.053)	0.907 (0.083)
Common threshold	0.980 (0.016)	0.948 (0.040)	0.974 (0.018)	0.964 (0.026)	0.919 (0.069)
Kandinsky: groups	0.905 (0.031)	0.890 (0.068)	0.890 (0.044)	0.893 (0.057)	0.865 (0.096)
Kandinsky: linear	0.899 (0.033)	0.893 (0.064)	0.888 (0.051)	0.889 (0.059)	0.875 (0.087)
Group-L_2	0.910 (0.033)	0.903 (0.059)	0.911 (0.035)	0.915 (0.047)	0.913 (0.085)
Marginal	0.931 (0.025)	0.855 (0.055)	0.904 (0.031)	0.878 (0.040)	0.799 (0.089)
Ratio-weighted	0.935 (0.025)	0.862 (0.055)	0.909 (0.030)	0.884 (0.040)	0.810 (0.084)
Kernel-weighted	0.921 (0.026)	0.868 (0.055)	0.909 (0.029)	0.880 (0.042)	0.812 (0.085)
Plug-in predictor	0.992 (0.009)	0.976 (0.025)	0.986 (0.011)	0.985 (0.013)	0.956 (0.046)
Static mixture	0.910 (0.283)	0.908 (0.286)	0.908 (0.285)	0.904 (0.286)	0.900 (0.286)

The maximum-score radii in Table 14 use Proposition 2; the group-L_2 radius is an upper bound. Table 15 reports each group separately.

Table 16: ChestMNIST reconstruction under controlled prevalence changes.

Method	Coverage	Max. under.	Max. dev.	Radius	Empty
Proposed	0.902 (0.015)	0.012 (0.012)	0.021 (0.010)	0.274 (0.006)	0.002
Common threshold	0.909 (0.013)	0.006 (0.009)	0.021 (0.010)	0.278 (0.006)	0.002
Kandinsky: groups	0.900 (0.015)	0.014 (0.012)	0.022 (0.010)	0.273 (0.006)	0.002
Kandinsky: linear	0.899 (0.015)	0.014 (0.013)	0.021 (0.010)	0.272 (0.006)	0.001
Group-L_2	0.901 (0.015)	0.013 (0.013)	0.022 (0.011)	0.500 (0.000)	0.001
Marginal	0.903 (0.014)	0.010 (0.011)	0.019 (0.010)	0.274 (0.005)	0.002
Ratio-weighted	0.901 (0.015)	0.011 (0.013)	0.020 (0.011)	0.273 (0.006)	0.002
Kernel-weighted	0.898 (0.013)	0.013 (0.011)	0.018 (0.010)	0.270 (0.005)	0.001
Plug-in predictor	0.698 (0.012)	0.215 (0.018)	0.215 (0.018)	0.210 (0.000)	0.006
Static mixture	0.899 (0.110)	0.030 (0.100)	0.057 (0.097)	0.290 (0.057)	0.003

Table 17: ChestMNIST coverage within observed groups.

Method	No finding	Abnormal
Proposed	0.900 (0.020)	0.903 (0.018)
Common threshold	0.905 (0.019)	0.911 (0.014)
Kandinsky: groups	0.899 (0.020)	0.900 (0.018)
Kandinsky: linear	0.898 (0.019)	0.899 (0.018)
Group-L_2	0.900 (0.020)	0.901 (0.019)
Marginal	0.899 (0.020)	0.905 (0.014)
Ratio-weighted	0.896 (0.020)	0.903 (0.015)
Kernel-weighted	0.895 (0.019)	0.899 (0.013)
Plug-in predictor	0.690 (0.023)	0.700 (0.013)
Static mixture	0.894 (0.111)	0.900 (0.110)

The maximum-score radii in Table 16 use Proposition 2; the group-L_2 radius is an upper bound. Table 17 reports each group separately.

Table 18: NLM report cohort reconstruction under controlled prevalence changes.

Method	Coverage	Max. under.	Max. dev.	Radius	Empty
Proposed	0.903 (0.024)	0.022 (0.024)	0.039 (0.019)	0.261 (0.009)	0.002
Common threshold	0.912 (0.023)	0.011 (0.015)	0.036 (0.019)	0.266 (0.010)	0.002
Kandinsky: groups	0.897 (0.024)	0.026 (0.025)	0.039 (0.019)	0.259 (0.009)	0.002
Kandinsky: linear	0.895 (0.023)	0.028 (0.028)	0.040 (0.024)	0.244 (0.008)	0.001
Group-L_2	0.904 (0.023)	0.019 (0.023)	0.035 (0.019)	0.500 (0.001)	0.000
Marginal	0.899 (0.022)	0.019 (0.020)	0.034 (0.017)	0.259 (0.008)	0.002
Ratio-weighted	0.901 (0.024)	0.019 (0.021)	0.036 (0.016)	0.260 (0.009)	0.002
Kernel-weighted	0.905 (0.020)	0.014 (0.018)	0.032 (0.017)	0.251 (0.007)	0.001
Plug-in predictor	0.631 (0.022)	0.290 (0.031)	0.290 (0.031)	0.183 (0.001)	0.019
Static mixture	0.897 (0.117)	0.038 (0.107)	0.073 (0.101)	0.278 (0.062)	0.005

Table 19: NLM report cohort coverage within observed groups.

Method	Normal	Abnormal
Proposed	0.899 (0.038)	0.904 (0.028)
Common threshold	0.916 (0.033)	0.910 (0.024)
Kandinsky: groups	0.894 (0.038)	0.899 (0.029)
Kandinsky: linear	0.893 (0.040)	0.895 (0.028)
Group-L_2	0.897 (0.036)	0.906 (0.026)
Marginal	0.903 (0.034)	0.898 (0.024)
Ratio-weighted	0.905 (0.036)	0.899 (0.025)
Kernel-weighted	0.906 (0.034)	0.905 (0.020)
Plug-in predictor	0.635 (0.047)	0.629 (0.026)
Static mixture	0.900 (0.118)	0.896 (0.118)

The maximum-score radii in Table 18 use Proposition 2; the group-L_2 radius is an upper bound. Table 19 reports each group separately.

The maximum-score radii in Table 20 use Proposition 2; the group-L_2 radius is an upper bound. Table 21 reports each group separately.

The full cohort tables above give the group-level results and empty-set frequencies.

S8.5 Paired comparisons with one threshold and with Kandinsky calibration

Figure 3 compares methods within the same replications. The common threshold is the largest group threshold, which keeps the guarantee of Theorem 1 in every group with one threshold. Each

Table 20: ChestMNIST MAE reconstruction under controlled prevalence changes.

Method	Coverage	Max. under.	Max. dev.	Radius	Empty
Proposed	0.898 (0.026)	0.027 (0.025)	0.040 (0.019)	0.447 (0.006)	0.000
Common threshold	0.926 (0.019)	0.015 (0.022)	0.044 (0.015)	0.454 (0.005)	0.000
Kandinsky: groups	0.888 (0.029)	0.035 (0.029)	0.044 (0.023)	0.444 (0.006)	0.000
Kandinsky: linear	0.890 (0.030)	0.033 (0.029)	0.042 (0.024)	0.437 (0.007)	0.000
Group-L_2	0.895 (0.028)	0.029 (0.027)	0.042 (0.021)	0.500 (0.000)	0.000
Marginal	0.913 (0.018)	0.026 (0.027)	0.041 (0.019)	0.451 (0.004)	0.000
Ratio-weighted	0.896 (0.025)	0.042 (0.035)	0.050 (0.029)	0.446 (0.005)	0.000
Kernel-weighted	0.916 (0.018)	0.021 (0.026)	0.040 (0.019)	0.446 (0.004)	0.000
Plug-in predictor	0.027 (0.008)	0.884 (0.009)	0.884 (0.009)	0.243 (0.002)	0.000
Static mixture	0.885 (0.246)	0.080 (0.228)	0.138 (0.210)	0.452 (0.070)	0.000

Table 21: ChestMNIST MAE coverage within observed groups.

Method	No finding	Abnormal
Proposed	0.895 (0.036)	0.899 (0.034)
Common threshold	0.897 (0.035)	0.935 (0.018)
Kandinsky: groups	0.889 (0.037)	0.887 (0.038)
Kandinsky: linear	0.891 (0.035)	0.890 (0.038)
Group-L_2	0.898 (0.038)	0.894 (0.034)
Marginal	0.881 (0.035)	0.924 (0.018)
Ratio-weighted	0.861 (0.039)	0.907 (0.025)
Kernel-weighted	0.886 (0.034)	0.926 (0.018)
Plug-in predictor	0.019 (0.013)	0.030 (0.010)
Static mixture	0.868 (0.247)	0.890 (0.246)

group threshold is at most this value. For nonempty sets the exact radius is nondecreasing in the threshold, and an empty set has radius zero. The proposed-minus-common radius difference is hence nonpositive in every replication, and the left panel reports its size. Maximum group undercoverage is the largest shortfall of empirical group coverage below 0.9. Kandinsky's group variant differs from (4) only in the finite-sample order correction and tie selection, and the right panel reports the change in maximum undercoverage produced by that correction. Compared with both Kandinsky variants, the proposed rule has a slightly larger mean radius.

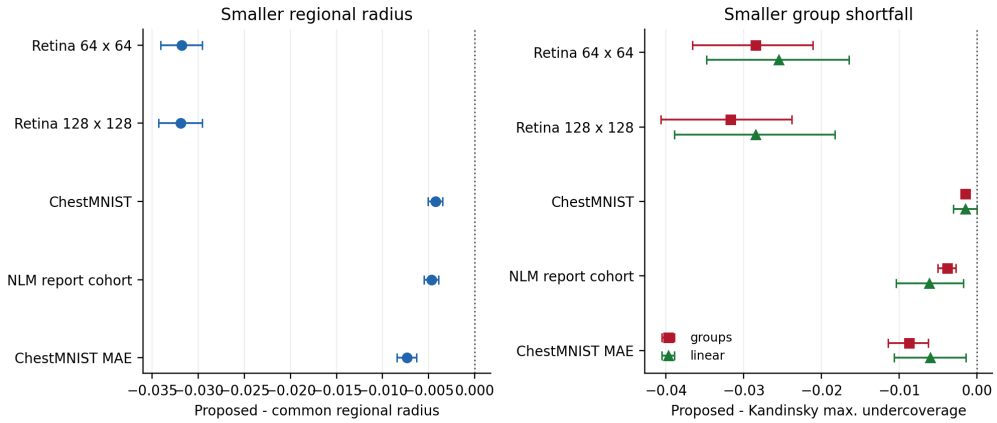


Figure 3: Paired proposed-minus-comparison differences over 100 replications. Bars are 95% bootstrap intervals; negative values favor the proposed rule.

S8.6 Deterministic and randomized deployment

We evaluate the static mixture both after drawing its deployed threshold and after averaging over that draw. Table 22 reports these results. Randomization-averaged coverage does not describe the coverage realized by every returned set. The standard deviation of realized static-mixture coverage ranges from 0.110 to 0.299, against 0.015 to 0.026 for the deterministic group rule. The expected static-mixture radius is larger in all five settings. Numbers in parentheses represent standard deviations.

Table 22: Realized and randomization-averaged coverage of the static mixture.

Cohort	Group rule	Static realized	Static averaged	Expected radius
Retina 64×64	0.906 (0.020)	0.897 (0.299)	0.898 (0.004)	0.208 (0.024)
Retina 128×128	0.906 (0.021)	0.907 (0.285)	0.898 (0.004)	0.203 (0.024)
ChestMNIST	0.902 (0.015)	0.899 (0.110)	0.901 (0.012)	0.292 (0.035)
NLM report cohort	0.903 (0.024)	0.897 (0.117)	0.896 (0.017)	0.279 (0.035)
ChestMNIST MAE	0.898 (0.026)	0.885 (0.246)	0.907 (0.014)	0.458 (0.009)

S8.7 Separate source and test pools

The controlled design samples calibration and target images from one pool, which enforces the within-group exchangeability of Theorem 1. A second analysis keeps the original pools separate. For the MedMNIST cohorts, the official training and validation images form the source pool and the official test images form the target pool. For the NLM cohort, a fixed hash of the report identifier assigns 70% of the reports to the source pool and the remainder to the target pool, and images from one report cannot occur in both pools. Group counts, predictors, scores, and calibration rules are unchanged. Separate pools can differ in both covariate frequencies and conditional image distributions. Proposition 3 bounds their effects through $\kappa(r, \mathcal{F})$ and ζ.

Table 23 reports pooled and group coverage. Numbers in parentheses represent standard deviations. The proposed rule falls below 0.9 in some groups of four cohorts. Its smallest mean group coverage is 0.876 for retinal grade zero at both resolutions, 0.869 for abnormal NLM reports, and 0.848 for ChestMNIST images without findings under the MAE predictor. In the NLM cohort its pooled coverage equals that of marginal calibration. These results show that calibration within groups does not protect against differences in the within-group distribution between pools. Such differences require control of ζ.

Table 23: Coverage with separate source and test pools.

Cohort	Method	Pooled coverage	Group coverages
Retina 64×64	Proposed	0.914 (0.014)	0.876; 0.968; 0.898; 0.965; 0.892
	Common threshold	0.975 (0.015)	0.982; 0.984; 0.973; 0.984; 0.905
	Kandinsky: groups	0.902 (0.016)	0.873; 0.947; 0.884; 0.953; 0.861
	Kandinsky: linear	0.900 (0.018)	0.863; 0.947; 0.884; 0.959; 0.862
	Marginal	0.905 (0.018)	0.909; 0.915; 0.905; 0.930; 0.795
Retina 128×128	Proposed	0.914 (0.015)	0.876; 0.966; 0.899; 0.964; 0.898
	Common threshold	0.976 (0.013)	0.984; 0.983; 0.974; 0.984; 0.909
	Kandinsky: groups	0.902 (0.016)	0.873; 0.948; 0.885; 0.954; 0.865
	Kandinsky: linear	0.901 (0.018)	0.863; 0.950; 0.888; 0.956; 0.860
	Marginal	0.904 (0.018)	0.907; 0.915; 0.904; 0.928; 0.788
ChestMNIST	Proposed	0.902 (0.015)	0.904; 0.901
	Common threshold	0.913 (0.011)	0.909; 0.914
	Kandinsky: groups	0.899 (0.014)	0.903; 0.898
	Kandinsky: linear	0.901 (0.014)	0.905; 0.900
	Marginal	0.905 (0.010)	0.902; 0.906
NLM report cohort	Proposed	0.880 (0.021)	0.912; 0.869
	Common threshold	0.891 (0.019)	0.920; 0.881
	Kandinsky: groups	0.873 (0.021)	0.908; 0.862
	Kandinsky: linear	0.877 (0.022)	0.910; 0.866
	Marginal	0.880 (0.019)	0.911; 0.869
ChestMNIST MAE	Proposed	0.911 (0.024)	0.848; 0.932
	Common threshold	0.925 (0.019)	0.866; 0.944
	Kandinsky: groups	0.903 (0.025)	0.840; 0.924
	Kandinsky: linear	0.903 (0.025)	0.861; 0.917
	Marginal	0.912 (0.019)	0.843; 0.935

S8.8 Regional radius and an accepted reconstruction

We examine whether a representative image can satisfy the observed coarse averages while keeping the worst-case regional error close to the smallest possible value. Proposition S1 answers this question for each nonempty prediction set. We evaluate its construction in the same 100 replications for each of the five medical cohorts, using identical images, predictor queries, scores, and group thresholds.

Table 24 separates three quantities. The first is the exact regional radius of the score set before imposing the observed coarse averages. Theorem S1 accounts for all 84 nested regional constraints in this calculation. The second is the radius after imposing the observation, as in Proposition 2. The third restricts the center to this latter set and gives the smallest worst-case regional error among accepted images. Within each replication, all three means use the same target images whose final

prediction sets are nonempty. We report the empty-set fraction separately and then average over the 100 replications. Numbers in parentheses represent standard deviations.

Table 24: Regional radii and reconstructions satisfying the observed means.

Cohort	Score set	With observed means	Accepted center	Empty fraction
Retina 64×64	0.127974 (0.004054)	0.127902 (0.004056)	0.128370 (0.004003)	0.000000
Retina 128×128	0.124478 (0.004061)	0.124441 (0.004060)	0.126053 (0.003943)	0.000000
ChestMNIST	0.274507 (0.006058)	0.274386 (0.006056)	0.274608 (0.006035)	0.001756
NLM report cohort	0.261836 (0.009024)	0.261770 (0.009021)	0.262277 (0.008977)	0.001675
ChestMNIST MAE	0.447526 (0.005685)	0.446709 (0.005673)	0.449614 (0.005959)	0.000000

The optimal accepted center increases the mean radius by 0.08% to 1.30% relative to the unrestricted center of the same set. It satisfies the score constraints and the observed coarse averages. The coordinatewise midpoint does not satisfy these averages in the evaluated nonempty sets. Thus the accepted center gives a reconstruction compatible with the observation at a small additional cost in the largest regional error over the set. When the set contains the true image, this radius bounds all regional errors of the accepted center simultaneously.

Imposing the observed coarse averages reduces the mean radius by 0.025% to 0.183% in these cohorts. This reduction is distinct from the approximately 20% retinal improvement over the common threshold in Table 2, where both methods already impose the same observation. That comparison measures the effect of group calibration. The present comparison measures the effect of the observation restriction and the cost of requiring an accepted center.

S9 Resolution and summary budget in normalized image units

The resolution experiment measures the number of dyadic summaries needed for a fixed bound on the full-image radius. The comparison uses the same root mean squared intensity units for every grid and does not infer a summary budget from the two medical resolutions, whose regional score is unchanged.

S9.1 Data-generating distributions

The regular and unstructured image models isolate the role of spatial restrictions in the summary budget.

The grid side length is $16, 32, 64$, or 128. The covariate is an observed binary group with source probability 0.3 and target probability 0.8 for group one. Each replication contains 400 independent calibration images and 400 independent target images. The predictor center is the known conditional mean, which isolates spatial approximation from predictor-estimation error.

At pixel coordinates $u, v \in [0, 1]$, the regular image is

$$Y(u, v) = \frac{1}{2} + 0.04G \sin(2\pi u) + \frac{a_G}{3} \{U_1 \sin(2\pi u) + U_2 \cos(2\pi v) + U_3 \sin(2\pi u) \sin(2\pi v)\}, \tag{S.81}$$

where $a_0 = 0.08$, $a_1 = 0.14$, and U_1, U_2, U_3 are independent uniform variables on $[-1, 1]$. This distribution is unchanged as the grid is refined. The same underlying function is evaluated at pixel centers. In the unstructured case, pixels are independent uniform variables on $[0, 1]$, and the center is the constant image $1/2$.

The regular functions in (S.81) have Lipschitz constant at most $2\pi\{(0.04+2a_1/3)^2+(2a_1/3)^2\}^{1/2} < 0.42\sqrt{6}$. Lemma S1 thus verifies (S.11) with $s = 1$ and $R = 0.42$ for every generated image and every resolution. The conditional mean satisfies the same bound. This verifies the model class used for the radius calculation; it is not a regularity condition estimated from the test responses.

S9.2 Calibration and radius measurement

For every dyadic level, we calibrate the projected norm score within groups using (4). Regular-image candidates are restricted to $\mathcal{Y}^{(V)}_{1,0.42}$, and their radius bound is the calibrated projected threshold plus $0.84\, 2^{-j}$. At full resolution the projection remainder is zero. Unstructured candidates range

over $[0,1]^V$. The constant center gives radius bound $1/2$ before the full image is included, and the full-score threshold gives the radius bound at $M = V$.

The reported budget is the smallest tested dyadic rank whose mean radius bound is at most 0.35. It concerns the stated radius bound and does not assert that the construction minimizes the exact image radius. For unstructured candidates, Theorem 3 also gives the lower bound $M \geq V\{1-(0.35/0.4)^2\}$ for any linear summary that admits the stated interior image. The center is accepted by every norm-score set in this experiment and lies in $[0.4,0.6]^V$.

Table 25: Normalized image-radius analysis on the 16×16 grid.

Image class	M	Coverage	Projected radius	Full radius bound	Omitted norm
Regular	4	0.904 (0.027)	0.027 (0.001)	0.447 (0.001)	0.019 (0.000)
Regular	16	0.903 (0.027)	0.033 (0.001)	0.243 (0.001)	0.012 (0.000)
Regular	64	0.903 (0.028)	0.036 (0.001)	0.141 (0.001)	0.006 (0.000)
Regular	256	0.903 (0.028)	0.036 (0.001)	0.036 (0.001)	0.000 (0.000)
Unstructured	4	0.905 (0.024)	0.051 (0.002)	0.500 (0.000)	0.286 (0.000)
Unstructured	16	0.903 (0.024)	0.087 (0.002)	0.500 (0.000)	0.279 (0.000)
Unstructured	64	0.902 (0.028)	0.159 (0.001)	0.500 (0.000)	0.250 (0.000)
Unstructured	256	0.902 (0.026)	0.299 (0.001)	0.299 (0.001)	0.000 (0.000)

Table 26: Normalized image-radius analysis on the 32×32 grid.

Image class	M	Coverage	Projected radius	Full radius bound	Omitted norm
Regular	4	0.904 (0.027)	0.027 (0.001)	0.447 (0.001)	0.019 (0.000)
Regular	16	0.903 (0.027)	0.032 (0.001)	0.242 (0.001)	0.012 (0.000)
Regular	64	0.903 (0.028)	0.035 (0.001)	0.140 (0.001)	0.006 (0.000)
Regular	256	0.903 (0.029)	0.036 (0.001)	0.089 (0.001)	0.003 (0.000)
Regular	1024	0.903 (0.028)	0.036 (0.001)	0.036 (0.001)	0.000 (0.000)
Unstructured	4	0.900 (0.027)	0.025 (0.001)	0.500 (0.000)	0.288 (0.000)
Unstructured	16	0.899 (0.025)	0.044 (0.001)	0.500 (0.000)	0.286 (0.000)
Unstructured	64	0.905 (0.024)	0.080 (0.001)	0.500 (0.000)	0.279 (0.000)
Unstructured	256	0.900 (0.030)	0.152 (0.001)	0.500 (0.000)	0.250 (0.000)
Unstructured	1024	0.901 (0.026)	0.294 (0.000)	0.294 (0.000)	0.000 (0.000)

Table 27: Normalized image-radius analysis on the 64×64 grid.

Image class	M	Coverage	Projected radius	Full radius bound	Omitted norm
Regular	4	0.904 (0.027)	0.027 (0.001)	0.447 (0.001)	0.019 (0.000)
Regular	16	0.903 (0.027)	0.032 (0.001)	0.242 (0.001)	0.012 (0.000)
Regular	64	0.903 (0.028)	0.035 (0.001)	0.140 (0.001)	0.006 (0.000)
Regular	256	0.903 (0.029)	0.036 (0.001)	0.089 (0.001)	0.003 (0.000)
Regular	1024	0.903 (0.029)	0.036 (0.001)	0.063 (0.001)	0.001 (0.000)
Regular	4096	0.903 (0.028)	0.036 (0.001)	0.036 (0.001)	0.000 (0.000)
Unstructured	4	0.906 (0.025)	0.013 (0.000)	0.500 (0.000)	0.289 (0.000)
Unstructured	16	0.900 (0.026)	0.022 (0.000)	0.500 (0.000)	0.288 (0.000)
Unstructured	64	0.905 (0.025)	0.040 (0.000)	0.500 (0.000)	0.286 (0.000)
Unstructured	256	0.909 (0.022)	0.076 (0.000)	0.500 (0.000)	0.280 (0.000)
Unstructured	1024	0.905 (0.027)	0.148 (0.000)	0.500 (0.000)	0.250 (0.000)
Unstructured	4096	0.908 (0.024)	0.291 (0.000)	0.291 (0.000)	0.000 (0.000)

Table 28: Normalized image-radius analysis on the 128×128 grid.

Image class	M	Coverage	Projected radius	Full radius bound	Omitted norm
Regular	4	0.904 (0.027)	0.027 (0.001)	0.447 (0.001)	0.019 (0.000)
Regular	16	0.903 (0.027)	0.032 (0.001)	0.242 (0.001)	0.012 (0.000)
Regular	64	0.903 (0.028)	0.035 (0.001)	0.140 (0.001)	0.006 (0.000)
Regular	256	0.903 (0.029)	0.036 (0.001)	0.089 (0.001)	0.003 (0.000)
Regular	1024	0.903 (0.028)	0.036 (0.001)	0.063 (0.001)	0.002 (0.000)
Regular	4096	0.903 (0.028)	0.036 (0.001)	0.050 (0.001)	0.001 (0.000)
Regular	16384	0.903 (0.028)	0.036 (0.001)	0.036 (0.001)	0.000 (0.000)
Unstructured	4	0.901 (0.024)	0.006 (0.000)	0.500 (0.000)	0.289 (0.000)
Unstructured	16	0.902 (0.024)	0.011 (0.000)	0.500 (0.000)	0.289 (0.000)
Unstructured	64	0.899 (0.027)	0.020 (0.000)	0.500 (0.000)	0.288 (0.000)
Unstructured	256	0.902 (0.026)	0.038 (0.000)	0.500 (0.000)	0.286 (0.000)
Unstructured	1024	0.903 (0.025)	0.074 (0.000)	0.500 (0.000)	0.280 (0.000)
Unstructured	4096	0.907 (0.022)	0.146 (0.000)	0.500 (0.000)	0.250 (0.000)
Unstructured	16384	0.898 (0.031)	0.290 (0.000)	0.290 (0.000)	0.000 (0.000)

The coverage entries in Tables 25–28 evaluate membership of the actual projected-score set, including the declared response class. The omitted-component norms are separate descriptive measurements. They are not substituted for prediction-set coverage. A set can include the true image with the required probability yet contain images that differ greatly in features omitted by the summary. Coverage alone does not bound those differences.

S9.3 A SHARPER BOUND FOR THE SAME REGULAR-IMAGE SETS

We evaluate (S.15) on the same 100 repetitions at every grid and dyadic level. The calibrated group thresholds and candidate sets are unchanged. For each target image, we apply the bound to its group threshold before averaging; applying the square root to an average threshold would give a different quantity. Table 29 reports the resulting bounds and their paired difference, defined as the orthogonal bound minus the additive bound. Monte Carlo standard errors equal the reported standard deviations divided by 10.

Table 29 shows that, at $M = 4$, the mean radius bound decreases from about 0.447 to 0.212 on every grid. The smallest tested budget with mean bound at most 0.35 is 4 for the orthogonal bound and 16 for the additive bound. Both bounds are equal at full resolution. Figure 4 shows that the sharper guarantee uses fewer spatial summaries at the same radius criterion. This comparison improves the bound for a fixed calibrated set; it does not change that set or its coverage and does not estimate a minimax summary budget.

Table 29: Additive and orthogonal radius bounds for the same regular-image sets. Numbers in parentheses represent standard deviations.

Grid	M	Coverage	Additive bound	Orthogonal bound	Paired difference
16 × 16	4	0.9041 (0.0274)	0.4468 (0.0006)	0.2118 (0.0001)	-0.2350 (0.0005)
16 × 16	16	0.9032 (0.0271)	0.2425 (0.0009)	0.1101 (0.0003)	-0.1324 (0.0006)
16 × 16	64	0.9028 (0.0285)	0.1406 (0.0010)	0.0637 (0.0006)	-0.0769 (0.0004)
16 × 16	256	0.9030 (0.0284)	0.0364 (0.0010)	0.0364 (0.0010)	0.0000 (0.0000)
32 × 32	4	0.9040 (0.0273)	0.4466 (0.0006)	0.2117 (0.0001)	-0.2349 (0.0005)
32 × 32	16	0.9032 (0.0273)	0.2424 (0.0009)	0.1100 (0.0003)	-0.1323 (0.0006)
32 × 32	64	0.9028 (0.0283)	0.1404 (0.0010)	0.0636 (0.0006)	-0.0768 (0.0004)
32 × 32	256	0.9028 (0.0286)	0.0887 (0.0010)	0.0450 (0.0008)	-0.0438 (0.0002)
32 × 32	1024	0.9030 (0.0284)	0.0364 (0.0010)	0.0364 (0.0010)	0.0000 (0.0000)
64 × 64	4	0.9039 (0.0272)	0.4466 (0.0006)	0.2117 (0.0001)	-0.2349 (0.0005)
64 × 64	16	0.9033 (0.0272)	0.2423 (0.0009)	0.1100 (0.0003)	-0.1323 (0.0006)
64 × 64	64	0.9028 (0.0282)	0.1404 (0.0010)	0.0636 (0.0006)	-0.0768 (0.0004)
64 × 64	256	0.9028 (0.0286)	0.0887 (0.0010)	0.0449 (0.0008)	-0.0438 (0.0002)
64 × 64	1024	0.9029 (0.0285)	0.0626 (0.0010)	0.0388 (0.0010)	-0.0239 (0.0001)
64 × 64	4096	0.9030 (0.0284)	0.0364 (0.0010)	0.0364 (0.0010)	0.0000 (0.0000)
128 × 128	4	0.9040 (0.0272)	0.4466 (0.0006)	0.2117 (0.0001)	-0.2349 (0.0005)
128 × 128	16	0.9033 (0.0272)	0.2423 (0.0009)	0.1100 (0.0003)	-0.1323 (0.0006)
128 × 128	64	0.9028 (0.0282)	0.1404 (0.0010)	0.0636 (0.0006)	-0.0768 (0.0004)
128 × 128	256	0.9028 (0.0286)	0.0887 (0.0010)	0.0449 (0.0008)	-0.0438 (0.0002)
128 × 128	1024	0.9029 (0.0285)	0.0626 (0.0010)	0.0388 (0.0010)	-0.0239 (0.0001)
128 × 128	4096	0.9030 (0.0284)	0.0495 (0.0010)	0.0370 (0.0010)	-0.0125 (0.0000)
128 × 128	16384	0.9030 (0.0284)	0.0364 (0.0010)	0.0364 (0.0010)	0.0000 (0.0000)

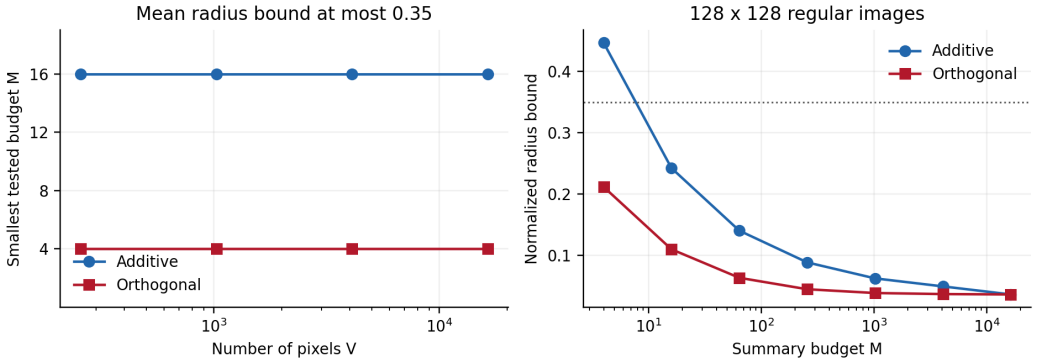


Figure 4: Summary budgets and radius bounds for the same regular-image sets.

S10 PREDICTOR QUERIES UNDER CONTINUOUS COVARIATES

Increasing the number of predictor queries reduces sampling error in its moments but does not remove a fixed error in the predictor distribution. We examine this distinction for the group rule in Theorem S2, using the exact regional radius of Proposition 2.

We use 32×32 images and the same 84 regional averages as in the medical analysis. Let W be constant on the 8×8 fine regions. Within each group of four fine regions sharing one observed coarse block, the two regions in the first row have value 1 and the other two have value -1. Every coarse average of W equals zero. Given $T \sim \mathrm{Unif}(0,1)$ and $G \in \{0,1\}$, define

$$m(T) = 0.45 + 0.10T, \qquad a(T,G) = \sigma_0(1+T)(1+0.5G)/2,$$
$$Y = m(T)\mathbf{1} + a(T,G)UW, \qquad \widetilde{Y}_k = m(T)\mathbf{1} + \{d + a(T,G)U_k\}W,$$

where $U, U_1, \ldots, U_K$ are independent uniform variables on $[-1,1]$. The observed coarse image equals $m(T)\mathbf{1}$, and the observed group is G. We use $\sigma_0 \in \{0, 0.001, 0.05\}$ and $d \in \{0, 0.01\}$. All source and predictor images have intensities in $[0,1]$ and preserve the observed coarse image. Their means belong to the same closed convex candidate class, including when $\sigma_0 = 0$. The source conditional mean and standard deviation of a regional average h are $h\{m(T)\mathbf{1}\}$ and $a(T,G)|h(W)|/\sqrt{3}$. The predictor population mean differs by $d\,h(W)$ and its variance is unchanged.

Each of the 100 independent repetitions has 200 calibration images and 400 prediction images per group. We set $\alpha = 0.1$, $\lambda = 0.05$, and $K \in \{8, 32, 128, 512\}$. Predictor query batches are independent across observations and independent of the responses. Within each repetition, the same covariates, source uniforms, and nested query batches are used for all parameter settings and query counts. The finite-query, predictor-population, and source-population scores use the same calibration responses and corrected group ranks. We evaluate target populations with $P_f(G=1) \in \{0.2, 0.5, 0.8\}$ and the same distribution of T within groups. The common-threshold comparison uses the largest group threshold and the same finite-query score.

For each target mixture, we first average the paired radius difference within each group, combine the two group means with the target weights, and then take the absolute value. The source difference compares the finite-query and source-population sets. The query difference compares the finite-query and predictor-population sets and separates sampling error from predictor bias. These are Monte Carlo estimates of differences between expected radii, as in Theorem S2; they are not expectations of pointwise absolute differences. We also compute the standard error of the signed target-mean difference from the independent prediction observations. This design does not use a common random map across covariates and does not test the uniform statement of Proposition S20.

Table 30 reports the equally weighted target mixture. All radius and difference entries in Tables 30 and 31 are multiplied by 10^3. For $\sigma_0 = 0$, the reported values are identical at all four query counts. With $d = 0$, increasing K from 8 to 512 decreases the mean source difference from 0.0000631 to 0.00000826 at $\sigma_0 = 0.001$ and from 0.00373 to 0.000330 at $\sigma_0 = 0.05$. The mean coverage in the positive-scale settings ranges from 0.9001 to 0.9053. The zero-scale settings have coverage one, as permitted by the conservative group guarantee.

Table 30: Predictor queries and exact regional radius under continuous covariates. Numbers in parentheses represent standard deviations.

σ_0	d	K	Coverage	Group radius	Source difference	Query difference
0	0	8–512	1.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
0	0.01	8–512	1.000 (0.000)	10.000 (0.000)	10.000 (0.000)	0.000 (0.000)
0.001	0	8	0.902 (0.018)	0.942 (0.034)	0.063 (0.029)	0.063 (0.029)
0.001	0	32	0.900 (0.018)	0.892 (0.032)	0.020 (0.014)	0.020 (0.014)
0.001	0	128	0.902 (0.019)	0.884 (0.029)	0.012 (0.010)	0.012 (0.010)
0.001	0	512	0.902 (0.019)	0.882 (0.029)	0.008 (0.006)	0.008 (0.006)
0.001	0.01	8	0.903 (0.017)	10.771 (0.035)	9.892 (0.039)	0.033 (0.024)
0.001	0.01	32	0.905 (0.018)	10.750 (0.036)	9.871 (0.034)	0.016 (0.013)
0.001	0.01	128	0.904 (0.019)	10.744 (0.035)	9.865 (0.035)	0.012 (0.009)
0.001	0.01	512	0.905 (0.018)	10.743 (0.034)	9.864 (0.035)	0.008 (0.006)
0.05	0	8	0.901 (0.018)	46.177 (1.628)	3.726 (1.473)	3.726 (1.473)
0.05	0	32	0.900 (0.018)	43.127 (1.241)	0.936 (0.658)	0.936 (0.658)
0.05	0	128	0.903 (0.020)	42.754 (1.106)	0.559 (0.344)	0.559 (0.344)
0.05	0	512	0.901 (0.020)	42.517 (1.132)	0.330 (0.239)	0.330 (0.239)
0.05	0.01	8	0.903 (0.018)	50.002 (1.678)	7.551 (1.555)	2.917 (1.376)
0.05	0.01	32	0.903 (0.020)	47.781 (1.427)	5.330 (1.319)	0.985 (0.741)
0.05	0.01	128	0.903 (0.020)	47.220 (1.360)	4.769 (1.275)	0.530 (0.426)
0.05	0.01	512	0.903 (0.018)	47.149 (1.237)	4.698 (1.259)	0.379 (0.248)

Figure 5 separates finite-query error from the error of the predictor population. The query difference decreases with K at both positive scales, including when $d = 0.01$. The source difference includes the fixed predictor error. At $\sigma_0 = 0.05$ and $d = 0.01$, it is 0.00470 at $K = 512$, close to the predictor-population difference 0.00465. At $\sigma_0 = 0$ and $d = 0.01$, both population and finite-query

radii equal 0.01 for every query count. These results support increasing the query budget to reduce sampling error while treating predictor bias separately.

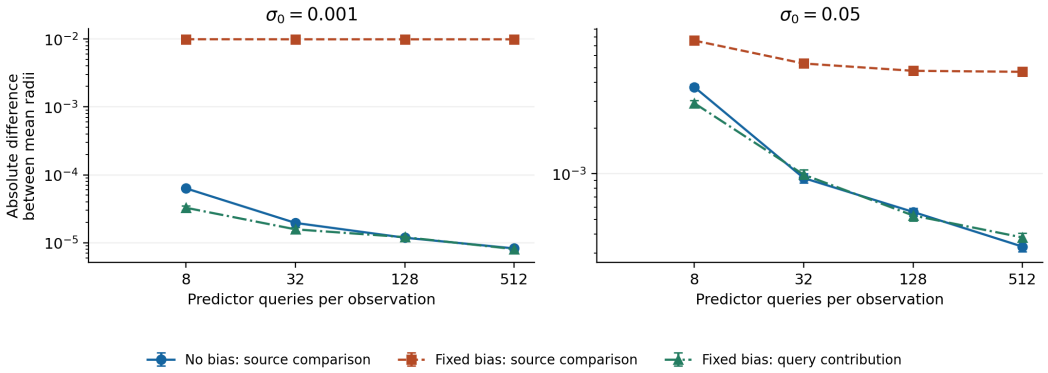


Figure 5: Predictor-query error and fixed predictor bias in the exact regional radius. Error bars represent Monte Carlo standard errors over 100 repetitions.

Table 31 shows that group thresholds reduce the exact radius relative to the largest group threshold across the three target mixtures in the positive-scale settings. At $K = 512$, $d = 0$, and equal group weights, the mean radius is 0.000882 compared with 0.001063 at $\sigma_0 = 0.001$, and 0.04252 compared with 0.04779 at $\sigma_0 = 0.05$. In these settings, the common threshold has higher coverage and a larger radius. All reported intersections are nonempty because their predictor centers satisfy the observed coarse constraints.

Table 31: Group thresholds and the largest group threshold at $K = 512$. Numbers in parentheses represent standard deviations.

σ_0	d	$P_f(G = 1)$	Group coverage	Common coverage	Group radius	Common radius
0	0	0.2	1.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.000 (0.000)
0	0	0.5	1.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.000 (0.000)
0	0	0.8	1.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.000 (0.000)
0	0.01	0.2	1.000 (0.000)	1.000 (0.000)	10.000 (0.000)	10.000 (0.000)
0	0.01	0.5	1.000 (0.000)	1.000 (0.000)	10.000 (0.000)	10.000 (0.000)
0	0.01	0.8	1.000 (0.000)	1.000 (0.000)	10.000 (0.000)	10.000 (0.000)
0.001	0	0.2	0.901 (0.021)	0.981 (0.006)	0.772 (0.024)	1.061 (0.050)
0.001	0	0.5	0.902 (0.019)	0.952 (0.015)	0.882 (0.029)	1.063 (0.050)
0.001	0	0.8	0.904 (0.023)	0.923 (0.023)	0.992 (0.041)	1.064 (0.051)
0.001	0.01	0.2	0.904 (0.022)	0.973 (0.011)	10.653 (0.031)	10.857 (0.056)
0.001	0.01	0.5	0.905 (0.018)	0.948 (0.015)	10.743 (0.034)	10.871 (0.056)
0.001	0.01	0.8	0.905 (0.020)	0.922 (0.020)	10.834 (0.046)	10.885 (0.056)
0.05	0	0.2	0.900 (0.023)	0.979 (0.009)	37.328 (1.071)	45.764 (1.693)
0.05	0	0.5	0.901 (0.020)	0.950 (0.016)	42.517 (1.132)	47.790 (1.766)
0.05	0	0.8	0.902 (0.024)	0.922 (0.024)	47.706 (1.534)	49.815 (1.842)
0.05	0.01	0.2	0.902 (0.022)	0.965 (0.013)	42.467 (1.281)	49.151 (1.836)
0.05	0.01	0.5	0.903 (0.018)	0.942 (0.017)	47.149 (1.237)	51.327 (1.922)
0.05	0.01	0.8	0.904 (0.020)	0.920 (0.021)	51.831 (1.657)	53.502 (2.011)

S11 Proofs for main and supporting results

Throughout the proofs, an expectation of the radius of a data-dependent prediction set is read as the outer expectation when the radius map is not measurable and as the ordinary expectation when it is measurable. The almost-sure inequality in Theorem S6 is read in the outer sense when the selected-fiber radial map is not measurable. This convention keeps every probability bound well defined without restricting the geometry of the prediction sets.

The proofs of the group-calibration and normalized-radius results appear first in Appendix S1. The proofs below establish the complete generic-class and additional theoretical statements.

S11.1 Auxiliary results for the main theory

Convex feasibility gives the candidate-augmented calibration rule. The score and threshold comparisons quantify approximation error for the generic discrepancy rule.

Proof of Proposition S13. The feasible convex combinations in (S.28) form $\mathrm{conv}\{V_{x,y}(q) : q \in \mathcal{Q}_{x,y}\}$, proving Proposition S13's equivalence. The sentinels satisfy $V_{x,y}\{q_-(x,y)\} = \bar{B}_x$ and $V_{x,y}\{q_+(x,y)\} = 0$. Hence

$$\alpha \bar{B}_x = \alpha V_{x,y}\{q_-(x,y)\} + (1-\alpha) V_{x,y}\{q_+(x,y)\},$$

which implies that (S.28) always holds. This completes the proof. □

Proof of Corollary S2. By Proposition S13, $\alpha \bar{B}_x$ lies in the convex hull of $V_{x,y}(q) \in \mathbb{R}^{d+1}$. Carathéodory's theorem gives a representation using at most $d+2$ vectors, hence a feasible $\pi_{x,y}$ with that support size. This completes the proof. □

Proof of Proposition S14. Recall that $B_j = B(X_j)$ for $j \in \mathcal{I}_{\mathrm{cal}} \cup \{\star\}$, with $X_\star = x$. By (S.28),

$$\frac{1}{N} \sum_{j \in \mathcal{I}_{\mathrm{cal}} \cup \{\star\}} B_j \{\widehat{a}_j(x,y) - \alpha\} = 0.$$

Every $f \in \mathcal{F}_d$ has the form $f(z) = \beta_f^\top B(z)$ for some $\beta_f \in \mathbb{R}^{d+1}$. Multiplying the above identity by $\beta_f^\top$ yields (S.42). This completes the proof. □

Proof of Proposition S16. Because all calibration scores are finite, the sentinels give $V_{\mathrm{cal}}(q_-) = \bar{B}_{\mathrm{cal}}$ and $V_{\mathrm{cal}}(q_+) = 0$, which implies that

$$\alpha \bar{B}_{\mathrm{cal}} = \alpha V_{\mathrm{cal}}(q_-) + (1-\alpha) V_{\mathrm{cal}}(q_+) \in \mathrm{conv}\{V_{\mathrm{cal}}(q) : q \in \mathcal{Q}_{\mathrm{cal}}\},$$

and the system (S.46) admits a solution.

Fix $f \in \mathcal{F}_d$ and write $f(z) = \beta_f^\top B(z)$. Multiplying (S.46) by $\beta_f^\top$ and using $\sum_q \widehat{\pi}(q) = 1$ yields

$$\sum_{q \in \mathcal{Q}_{\mathrm{cal}}} \widehat{\pi}(q) \frac{1}{n_{\mathrm{cal}}} \sum_{i \in \mathcal{I}_{\mathrm{cal}}} f(X_i)\{\mathbf{1}\{R_i > q\} - \alpha\} = 0. \tag{S.82}$$

Conditional on $\mathcal{D}_{\mathrm{cal}}$, draw $q^\star \sim \widehat{\pi}$ independently of the target observation $(X^*, Y^*) \sim P_f$. Then

$$P_f\{Y^* \notin \widehat{\mathcal{C}}_{\mathrm{mix}}(X^*; q^\star) \mid \mathcal{D}_{\mathrm{cal}}\} = \sum_{q \in \mathcal{Q}_{\mathrm{cal}}} \widehat{\pi}(q)\, P_f\{R_n(X,Y) > q\}.$$

On the event in (S.34), it holds uniformly over $f \in \mathcal{F}_d$ and $q \in \mathbb{R}$ that

$$\left| \frac{1}{n_{\mathrm{cal}}} \sum_{i \in \mathcal{I}_{\mathrm{cal}}} f(X_i)\{\mathbf{1}\{R_i > q\} - \alpha\} - \big(P_f\{R_n(X,Y) > q\} - \alpha\big) \right| \le \psi_n(\delta, \mathcal{F}_d).$$

Define the bracketed discrepancy

$$D_f(q) = P_f\{R_n(X,Y) > q\} - \alpha - \frac{1}{n_{\mathrm{cal}}} \sum_{i \in \mathcal{I}_{\mathrm{cal}}} f(X_i)\{\mathbf{1}\{R_i > q\} - \alpha\}.$$

Combining this bound with (S.82) gives, for every $f \in \mathcal{F}_d$,

$$\left| P_f\{Y^* \notin \widehat{\mathcal{C}}_{\mathrm{mix}}(X^*; q^\star) \mid \mathcal{D}_{\mathrm{cal}}\} - \alpha \right| = \left| \sum_{q \in \mathcal{Q}_{\mathrm{cal}}} \widehat{\pi}(q) D_f(q) \right| \le \psi_n(\delta, \mathcal{F}_d),$$

which is equivalent to (S.47). This completes the proof. □

Proof of Proposition S7. Put $M = \max\{0, \max_{i \in \mathcal{I}_{\mathrm{cal}}} R_i\}$. For every $q \ge M$, all strict-exceedance indicators $\mathbf{1}\{R_i > q\}$ are zero, which implies that

$$\Gamma_n^+(q; \mathcal{F}) = \sup_{f \in \mathcal{F}} \Big[-\alpha \frac{1}{n_{\mathrm{cal}}} \sum_{i \in \mathcal{I}_{\mathrm{cal}}} f(X_i) \Big] \le 0 \le \tau_n,$$

because $f \ge 0$. Thus the feasible set in (S.30) is nonempty and $\widehat{q} \le M < \infty$.

For attainment, let

$$\mathcal{G} = \{0, M\} \cup \{R_i : R_i \geq 0,\ i \in \mathcal{I}_{\text{cal}}\}.$$

For every $q \geq 0$, the indicator vector $\{\mathbf{1}\{R_i > q\} : i \in \mathcal{I}_{\text{cal}}\}$ equals its value at $g(q) = \max\{u \in \mathcal{G} : u \leq q\}$. Hence the smallest feasible point of $\mathcal{G}$ attains the infimum in (S.30).

Pointwise measurability of $\mathcal{F}$ makes $\Gamma_n^+(q; \mathcal{F})$ a countable supremum. The grid values are measurable functions of the scores, and finitely many measurable comparisons select the smallest feasible value. Thus $\widehat{q}$ has a measurable version. This completes the proof. □

Lemma S4. *If* $\sup_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} |R_n(x, y) - R^\circ(x, y)| \leq e$, *then*

$$P_f\{R^\circ(X, Y) \leq q - e\} \leq P_f\{R_n(X, Y) \leq q\} \leq P_f\{R^\circ(X, Y) \leq q + e\}$$

for every $f \in \mathcal{F}$ *and every* $q \in \mathbb{R}_+$.

Lemma S4 shows that a uniform error in the score changes coverage only through neighboring oracle thresholds.

Lemma S5. *Suppose that Assumption S1 holds. Let* $\widehat{q}$ *be defined by the one-sided rule* (S.30) *with tolerance* $\tau \geq 0$. *On the event where* $\sup_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} |R_n(x, y) - R^\circ(x, y)| \leq e$ *and*

$$\sup_{f \in \mathcal{F}} \sup_{q \in [q^\circ_{\alpha,\mathcal{F}} - r_0, q^\circ_{\alpha,\mathcal{F}} + r_0]} \left| \frac{1}{n_{\text{cal}}} \sum_{i \in \mathcal{I}_{\text{cal}}} f(X_i)\{\mathbf{1}\{R_n(X_i, Y_i) > q\} - \alpha\} - \big(P_f\{R_n(X, Y) > q\} - \alpha\big) \right| \leq \psi,$$

there exist constants $c_0, C > 0$, *depending only on* c_-, *such that, if*

$$\psi \leq \tau \qquad \text{and} \qquad \tau + \psi + e \leq c_0 r_0,$$

then $\left|\widehat{q} - q^\circ_{\alpha,\mathcal{F}}\right| \leq C(\tau + \psi + e)$.

Lemma S5 converts score and calibration errors into threshold error using the local population quantile condition.

Proposition S36. *Suppose that Proposition S11 holds. Then, it holds that any threshold bound* $\widehat{q} \leq q^\circ_{\alpha,\mathcal{F}} + t_n$ *implies* $\mathcal{S}_{\mathcal{F},\mathcal{H}}(\widehat{\mathcal{C}}) \leq A_{\mathcal{F}}(q^\circ_{\alpha,\mathcal{F}} + t_n)$.

Proposition S36 converts a threshold upper bound into an expected-radius upper bound. It does not identify the exact radius of the oracle set.

S11.2 Proof of Theorem S23

An unbounded image functional and coverage over every target weight imply an infinite expected radius when the covariate distribution is nonatomic.

Proof. Fix a source with nonatomic covariate marginal, and put $h = h_0$ and $T = h(Y)$. Define

$$\widehat{I}_h(x) = \{t \in \mathbb{R} : \exists y \in \widehat{\mathcal{C}}(x) \text{ with } h(y) = t\}.$$

Because $\mathrm{Gr}(\widehat{\mathcal{C}})$ is analytic and h is measurable, the set

$$G_h = \{(x, t) \in \mathcal{X} \times \mathbb{R} : \exists y \in \mathcal{Y} \text{ such that } (x, y) \in \mathrm{Gr}(\widehat{\mathcal{C}}) \text{ and } h(y) = t\}$$

is analytic. Define

$$\ell_h(x) = \inf\{t : (x, t) \in G_h\}, \qquad u_h(x) = \sup\{t : (x, t) \in G_h\},$$

with $\ell_h(x) = +\infty$ and $u_h(x) = -\infty$ on empty sections. For rational a, the sets $\{x : \ell_h(x) < a\}$ and $\{x : u_h(x) > a\}$ are analytic projections. Thus ℓ_h, u_h are universally measurable. Let

$$\widehat{J}_h(x) = \begin{cases} [\ell_h(x), u_h(x)], & \ell_h(x) \leq u_h(x), \\ \emptyset, & \ell_h(x) > u_h(x). \end{cases}$$

The interval $\widehat{J}_h(x)$ is universally measurable and contains $\widehat{I}_h(x)$ with the same extended diameter. Coverage of $\widehat{\mathcal{C}}$ gives $P_f\{T \in \widehat{J}_h(X)\} \geq 1 - \alpha$ for all $f \in \mathcal{F}_{\text{all}}$. For measurable A with $P_{\text{src},X}(A) > 0$, use $dP_f = f\, dP_{\text{src}}$ and $f(x) = \mathbf{1}\{x \in A\}/P_{\text{src},X}(A)$ to obtain

$$P_{\text{src}}\{T \in \widehat{J}_h(X) \mid X \in A\} \geq 1 - \alpha.$$

We apply the conditional-coverage argument of Barber et al. (2021, Proposition 2.2) directly to $h(\mathcal{Y})$. Fix $t \in h(\mathcal{Y})$ and $y_t \in \mathcal{Y}$ with $h(y_t) = t$. Let Q_A preserve $P_{\mathrm{src},X}$ and the response distribution off A, with $Y = y_t$ on A. The distribution-free premise applies to Q_A and its shift with weight $\mathbf{1}\{X \in A\}/P_{\mathrm{src},X}(A)$, under which the scalar response equals t.

Couple the N source observations under P_{src} and Q_A using the same covariates and auxiliary randomization. They differ with probability at most $NP_{\mathrm{src},X}(A)$. Coverage under Q_A gives

$$\mathbb{E}_{P_{\mathrm{src}}^N}\Big[P_{\mathrm{src},X}\{t \in \widehat{J}_h(X) \mid X \in A\}\Big] \geq 1 - \alpha - NP_{\mathrm{src},X}(A).$$

The expectation includes auxiliary randomization. Partition the nonatomic covariate distribution into finitely many cells of mass at most $\eta > 0$. Weighting the inequality by cell probabilities and summing gives $\mathbb{P}_{P_{\mathrm{src}}^N \times P_{\mathrm{src},X}}\{t \in \widehat{J}_h(X)\} \geq 1 - \alpha - N\eta$. Let $\eta \downarrow 0$.

Choose an integer $m > (1-\alpha)^{-1}$. For every $D > 0$, the unbounded range of h contains m points separated by at least D. An interval containing k of these points has length at least $D(k-1)_+$. Thus, with λ_1 denoting one-dimensional Lebesgue measure,

$$\mathbb{E}\Big[\lambda_1\{\widehat{J}_h(X)\}\Big] \geq D\{m(1-\alpha) - 1\}.$$

Letting $D \to \infty$ gives $\mathbb{E}\Big[\lambda_1\{\widehat{J}_h(X)\}\Big] = \infty$.

For an empty set, $\lambda_1\{\widehat{J}_h(x)\} = 0$; for infinite radius the required comparison is immediate. Fix x with $\widehat{\mathcal{C}}(x) \neq \emptyset$ and $\mathrm{rad}_{\mathcal{H}}\{\widehat{\mathcal{C}}(x)\} < \infty$. For each $\varepsilon > 0$, choose $y_{0,\varepsilon}(x) \in \mathcal{Y}$ such that

$$\sup_{y \in \widehat{\mathcal{C}}(x)} \sup_{g \in \mathcal{H}} |g(y) - g\{y_{0,\varepsilon}(x)\}| \leq \mathrm{rad}_{\mathcal{H}}\{\widehat{\mathcal{C}}(x)\} + \varepsilon.$$

Because $h \in \mathcal{H}$, the scalar image $\widehat{I}_h(x)$ lies in an interval of length at most $2\{\mathrm{rad}_{\mathcal{H}}\{\widehat{\mathcal{C}}(x)\} + \varepsilon\}$. Its hull has the same diameter. Letting $\varepsilon \downarrow 0$ gives $\lambda_1\{\widehat{J}_h(X)\} \leq 2\mathrm{rad}_{\mathcal{H}}\{\widehat{\mathcal{C}}(X)\}$ in the extended sense, hence $\mathbb{E}^*[\mathrm{rad}_{\mathcal{H}}\{\widehat{\mathcal{C}}(X)\}] = \infty$. This completes the proof. □

S11.3 Proof of Theorem S7

Permutation symmetry converts the candidate-augmented moment identity into exact target coverage.

Proof. Condition on the training fold and independent score-construction randomness. Let $N = n_{\mathrm{cal}} + 1$ and $Z_j = (X_j, Y_j)$, $j = 1, \ldots, N$, be i.i.d. from P_{src}. Apply the solver to the unordered augmented sample $Z_{1:N}$ and define

$$A_j = \sum_{q \in \mathcal{Q}_{Z_{1:N}}} \pi_{Z_{1:N}}(q)\mathbf{1}\{R_n(X_j, Y_j) > q\}, \qquad j = 1, \ldots, N,$$

where $\mathcal{Q}_{Z_{1:N}}$ is the augmented threshold grid and $\pi_{Z_{1:N}}$ the measurable permutation equivariant feasible selector. Thus $(A_1, \ldots, A_N)$ permutes with the observations.

Proposition S14 gives, for every $f \in \mathcal{F}_d$ and every augmented sample, $\frac{1}{N}\sum_{j=1}^N f(X_j)\{A_j - \alpha\} = 0$. Taking expectations and using exchangeability gives

$$\mathbb{E}_{P_{\mathrm{src}}}\{f(X_N)(A_N - \alpha) \mid \mathcal{D}_{\mathrm{tr}}\} = \mathbb{E}_{P_{\mathrm{src}}}\Big[\frac{1}{N}\sum_{j=1}^N f(X_j)\{A_j - \alpha\} \mid \mathcal{D}_{\mathrm{tr}}\Big] = 0.$$

Measurability of $\widehat{a}_{\mathrm{fd}}^+ : \mathcal{X} \times \mathcal{Y} \to [0,1]$ and $dP_f = f\, dP_{\mathrm{src}}$ give

$$\mathbb{E}_f\{\widehat{a}_{\mathrm{fd}}^+(X^*, Y^*) \mid \mathcal{D}_{\mathrm{tr}}\} = \mathbb{E}_{P_{\mathrm{src}}}\{f(X)\widehat{a}_{\mathrm{fd}}^+(X, Y) \mid \mathcal{D}_{\mathrm{tr}}\},$$

where $(X, Y) \sim P_{\mathrm{src}}$ is an independent draw. Identify it with (X_N, Y_N), for which $\widehat{a}_{\mathrm{fd}}^+(X_N, Y_N) = A_N$, and apply the exchangeability identity to obtain

$$\mathbb{E}_f\{\widehat{a}_{\mathrm{fd}}^+(X^*, Y^*) \mid \mathcal{D}_{\mathrm{tr}}\} = \mathbb{E}_{P_{\mathrm{src}}}\{f(X_N)A_N \mid \mathcal{D}_{\mathrm{tr}}\} = \alpha.$$

This proves the first identity in (S.32). Finally, Proposition S15 yields

$$P_f\{Y^* \notin \widehat{\mathcal{C}}_{\mathrm{fd}}^+(X^*; U_0) \mid \mathcal{D}_{\mathrm{tr}}\} = \mathbb{E}_f\{\widehat{a}_{\mathrm{fd}}^+(X^*, Y^*) \mid \mathcal{D}_{\mathrm{tr}}\} = \alpha,$$

which is equivalent to the second identity in (S.32). This completes the proof. □

Proof of Corollary S3. For each ℓ, the distribution of a source observation conditional on $X \in G_\ell$ is P_{f_ℓ} with $f_\ell(x) = \mathbf{1}\{x \in G_\ell\}/P_{\mathrm{src},X}(G_\ell)$. Because $f_\ell \in \mathcal{F}_d$ by assumption, Theorem S7 applied to $f = f_\ell$ yields (S.45). Multiplying (S.45) by $P_{\mathrm{src},X}(G_\ell)$ and taking complements yields (S.44). This completes the proof. □

S11.4 Proof of Theorem S8

Score error and calibration error determine how far the threshold can lie from the worst-case population quantile.

Proof. Put $e_{n,\delta} = e_n(\delta/2)$ and $\psi_{n,\delta} = \psi_n(\delta/2, \mathcal{F})$. Let

$$\mathcal{E}_{\mathrm{sc}} = \left\{ \sup_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} |R_n(x,y) - R^\circ(x,y)| \le e_{n,\delta} \right\},$$

and let $\mathcal{E}_{\mathrm{ep}}$ be the event in (S.34) with bound $\psi_{n,\delta}$. Equations (S.33) and (S.34) at levels $\delta/2$ give $\mathbb{P}(\mathcal{E}_{\mathrm{sc}} \cap \mathcal{E}_{\mathrm{ep}}) \ge 1 - \delta$. On this intersection, Lemma S5 gives $\left|\widehat{q} - q^\circ_{\alpha,\mathcal{F}}\right| \le C\{\tau_n + \psi_{n,\delta} + e_{n,\delta}\}$. This completes the proof. □

S11.5 Proof of Corollary S1

The common quantile and the upper local probability bound convert threshold localization into a two-sided coverage bound.

Proof. Put $\Delta_n = \tau_n + \psi_{n,\delta} + e_{n,\delta}$ and $q_\star = q^\circ_{\alpha,\mathcal{F}}$. The events in (S.33) and (S.34), each at level $\delta/2$, have intersection $\mathcal{E}$ with probability at least $1 - \delta$. On $\mathcal{E}$, Theorem S8 gives $|\widehat{q} - q_\star| \le C\Delta_n$, where C depends only on c_-, and $\sup_{x,y} |R_n(x,y) - R^\circ(x,y)| \le e_{n,\delta}$. Choose $c_0' \le c_0$ such that $(C+1)\Delta_n \le r_0$. For $f \in \mathcal{F}$, $q^\circ_{f,\alpha} = q_\star$. Letting $t \downarrow 0$ in the upper bound on $F^\circ_f(q_\star + t) - F^\circ_f(q_\star - t)$ excludes an atom, hence $F^\circ_f(q_\star) = 1 - \alpha$ by the quantile definition. Lemma S4 gives

$$F^\circ_f(\widehat{q} - e_{n,\delta}) \le P_f\{R_n(X,Y) \le \widehat{q}\} \le F^\circ_f(\widehat{q} + e_{n,\delta}).$$

Because $|\widehat{q} \pm e_{n,\delta} - q_\star| \le (C+1)\Delta_n \le r_0$, apply the local bound $\left|F^\circ_f(q_\star + u) - F^\circ_f(q_\star)\right| \le F^\circ_f(q_\star + |u|) - F^\circ_f(q_\star - |u|) \le c_+ |u|$ for $u \in [-r_0, r_0]$ at $u = \widehat{q} \pm e_{n,\delta} - q_\star$. With $P_f\{Y \in \widehat{\mathcal{C}}(X)\} = P_f\{R_n(X,Y) \le \widehat{q}\}$, this gives

$$\left| P_f\{Y \in \widehat{\mathcal{C}}(X)\} - (1-\alpha) \right| \le c_+(C+1)\Delta_n.$$

Taking the supremum over $f \in \mathcal{F}$ yields the claim with $C' = c_+(C+1)$. This completes the proof. □

S11.6 Proof of Theorem S9

Uniform calibration control bounds undercoverage directly at the selected threshold.

Proof. Put $\psi_{n,\delta} = \psi_n(\delta, \mathcal{F})$ and let $\mathcal{E}_{\mathrm{ep}}$ be the event in (S.34) with this bound. On $\mathcal{E}_{\mathrm{ep}}$, the definition in (S.30) and right-continuity of the strict-exceedance criterion give

$$\sup_{f \in \mathcal{F}} \frac{1}{n_{\mathrm{cal}}} \sum_{i \in \mathcal{I}_{\mathrm{cal}}} f(X_i)\{\mathbf{1}\{R_n(X_i,Y_i) > \widehat{q}\} - \alpha\} \le \tau_n.$$

Equation (S.34) at $\widehat{q}$ gives, uniformly over $f \in \mathcal{F}$, $P_f\{R_n(X,Y) > \widehat{q}\} - \alpha \le \tau_n + \psi_{n,\delta}$. Taking complements yields

$$(1-\alpha) - P_f\{Y \in \widehat{\mathcal{C}}(X)\} \le \tau_n + \psi_{n,\delta}.$$

Taking the positive part and then the supremum over $f \in \mathcal{F}$ yields (S.36). This completes the proof. □

S11.7 Proof of Theorem S10

A threshold just above the worst-case oracle quantile plus the score error is feasible and yields the radius upper bound.

Proof. Put $e_{n,\delta} = e_n(\delta/2)$ and $\psi_{n,\delta} = \psi_n(\delta/2, \mathcal{F})$. The events in (S.33) and (S.34) intersect with probability at least $1-\delta$. On this intersection, put $q_\star = q^\circ_{\alpha,\mathcal{F}}$ and, for $\eta > 0$, $q_{0,\eta} = q_\star + e_{n,\delta} + \eta$. For each $f \in \mathcal{F}$, $q_\star + \eta > q^\circ_{f,\alpha}$ gives $u_{f,\eta} < q_\star + \eta$ such that $P_f\{R^\circ(X,Y) \le u_{f,\eta}\} \ge 1-\alpha$. Lemma S4 gives

$$P_f\{R_n(X,Y) > q_{0,\eta}\} \le P_f\{R^\circ(X,Y) > q_\star + \eta\} \le P_f\{R^\circ(X,Y) > u_{f,\eta}\} \le \alpha.$$

Thus $\Gamma_n^+(q_{0,\eta}; \mathcal{F}) \le \psi_{n,\delta} \le \tau_n$, and $\widehat{q} \le q_{0,\eta}$. Let $\eta \downarrow 0$ and apply Proposition S36 to obtain $\mathcal{S}_{\mathcal{F},\mathcal{H}}(\widehat{\mathcal{C}}) \le A_{\mathcal{F}}(q^\circ_{\alpha,\mathcal{F}} + e_{n,\delta})$. This completes the proof. □

S11.8 Proof of Theorem S11

The score bound controls the paired radius difference directly; the calibration empirical process is needed for the coverage conclusion.

Proof. Condition on the queried support in the theorem. Put $\psi_q = \psi_{n,q}(\delta_q, \mathcal{F})$, $b_q = b_{q,K}(\delta_q)$, and $e_{n,q} = \sup_{x\in\mathcal{X}_q} \sup_{y\in\mathcal{Y}} \left|R_{n,K}(x,y) - R^\circ_q(x,y)\right|$. By (S.59) in Theorem S16, $\mathcal{E}_{\mathrm{T}} = \{e_{n,q} \le C_{\mathrm{sc}} b_q\}$ has probability at least $1-\delta_q$, where $\delta_q = \delta/3$. Proposition S2 compares the thresholds directly. Condition (C2) gives the empirical-process bounds and shared radial Lipschitz property.

Let $\mathcal{E}_{\mathrm{ep},K}$ denote the event

$$\sup_{f\in\mathcal{F}} \sup_{t\in\mathbb{R}} \left| \frac{1}{n_{\mathrm{cal}}} \sum_{i\in\mathcal{I}_{\mathrm{cal}}} f(X_i)\{\mathbf{1}\{R_{n,K}(X_i,Y_i) > t\} - \alpha\} - \{P_f(R_{n,K}(X,Y) > t) - \alpha\} \right| \le \psi_q,$$

and define $\mathcal{E}_{\mathrm{ep},\circ}$ analogously for R°_q. The screened empirical-process bounds give $\mathbb{P}(\mathcal{E}_{\mathrm{ep},K}), \mathbb{P}(\mathcal{E}_{\mathrm{ep},\circ}) \ge 1-\delta_q$. Hence $\mathcal{E} = \mathcal{E}_{\mathrm{T}} \cap \mathcal{E}_{\mathrm{ep},K} \cap \mathcal{E}_{\mathrm{ep},\circ}$ has probability at least $1-\delta$.

On $\mathcal{E}_{\mathrm{ep},K}$, the definition of $\widehat{q}_{q,K}$ gives

$$\sup_{f\in\mathcal{F}} \frac{1}{n_{\mathrm{cal}}} \sum_{i\in\mathcal{I}_{\mathrm{cal}}} f(X_i)\{\mathbf{1}\{R_{n,K}(X_i,Y_i) > \widehat{q}_{q,K}\} - \alpha\} \le 2\psi_q.$$

Applying the empirical-process event at the random threshold $\widehat{q}_{q,K}$ yields, for every $f \in \mathcal{F}$,

$$P_f\{R_{n,K}(X,Y) > \widehat{q}_{q,K}\} - \alpha \le 3\psi_q.$$

Taking complements, positive parts, and the supremum over $f \in \mathcal{F}$ gives $\Delta^-_{\mathcal{F}}(\widehat{\mathcal{C}}_{q,K}) \le 3\psi_q$, proving (S.38).

The thresholds use the same calibration sample, weights, and tolerance. On $\mathcal{E}_{\mathrm{T}}$, Proposition S2 gives $|\widehat{q}_{q,K} - \widehat{q}^\circ_q| \le e_{n,q}$. In particular,

$$\left|\widehat{q}_{q,K} - \widehat{q}^\circ_q\right| \le C\{\psi_q + e_{n,q}\}. \tag{S.83}$$

On $\mathcal{E}_{\mathrm{T}}$, for every $x \in \mathcal{X}_q$,

$$\mathcal{C}^{R_{n,K}}_{\widehat{q}_{q,K}}(x) \subseteq \mathcal{C}^{R^\circ_q}_{\widehat{q}_{q,K}+e_{n,q}}(x), \qquad \mathcal{C}^{R^\circ_q}_{\widehat{q}^\circ_q}(x) \subseteq \mathcal{C}^{R_{n,K}}_{\widehat{q}^\circ_q+e_{n,q}}(x).$$

Radius monotonicity and the shared radial Lipschitz bound give, for every $x \in \mathcal{X}_q$,

$$\left|\mathrm{rad}_{\mathcal{H}_q}\{\widehat{\mathcal{C}}_{q,K}(x)\} - \mathrm{rad}_{\mathcal{H}_q}\{\widehat{\mathcal{C}}^\circ_q(x)\}\right| \le A(x)\{\left|\widehat{q}_{q,K} - \widehat{q}^\circ_q\right| + e_{n,q}\}.$$

Take $\mathbb{E}_f$ and the supremum over f, using $\sup_{f\in\mathcal{F}} \mathbb{E}_f\{A(X)\} \le A_{\mathcal{F}}$, (S.83), and $e_{n,q} \le C_{\mathrm{sc}} b_q$, to obtain

$$\left|\mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\widehat{\mathcal{C}}_{q,K}) - \mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\widehat{\mathcal{C}}^\circ_q)\right| \le A_{\mathcal{F}} C\{\psi_q + b_q\}.$$

Equation (S.6) also gives $2A_{\mathcal{F}} e_{n,q} \le 2A_{\mathcal{F}} C_{\mathrm{sc}} b_q$. Because $\psi_q \ge 0$, this implies (S.39). This completes the proof. □

S11.9 Proof of Theorem S12

The upper bound controls omitted Haar details through coefficient decay; the lower bound uses an omitted detail in an accepted fiber.

Proof. For part (i), Theorem S11 with $\mathcal{H}_q = \mathcal{H}^{\mathrm{MR}}_{\le J'}$ gives, with probability at least $1-\delta$,

$$\mathcal{S}_{\mathcal{F},\mathcal{H}^{\mathrm{MR}}_{\le J'}}(\widehat{\mathcal{C}}_{q,K}) \le \mathcal{S}_{\mathcal{F},\mathcal{H}^{\mathrm{MR}}_{\le J'}}(\widehat{\mathcal{C}}^{\circ}_{q}) + A_{\mathcal{F}} C\{\psi_{n,q}(\delta_q,\mathcal{F}) + b_{q,K}(\delta_q)\}.$$

By (S.24), $\widehat{\mathcal{C}}_{q,K}(x) \subseteq \mathcal{Y}_R$. Lemma S3 gives $\mathrm{rad}_{\mathcal{H}^{\mathrm{MR}}}\{\widehat{\mathcal{C}}_{q,K}(x)\} \le \mathrm{rad}_{\mathcal{H}^{\mathrm{MR}}_{\le J'}}\{\widehat{\mathcal{C}}_{q,K}(x)\} + 2R2^{-sJ'}$. Taking $\mathbb{E}_f$ and the supremum gives $\mathcal{S}_{\mathcal{F},\mathcal{H}^{\mathrm{MR}}}(\widehat{\mathcal{C}}_{q,K}) \le \mathcal{S}_{\mathcal{F},\mathcal{H}^{\mathrm{MR}}_{\le J'}}(\widehat{\mathcal{C}}_{q,K}) + 2R2^{-sJ'}$. Use $2^{sJ'} \ge 2R/\epsilon$ to obtain (S.40). Definition S3 gives $M = 4^{J'}$. For $J' = 0$, $M = 1$; for $J' \ge 1$, minimality gives $2^{s(J'-1)} < 2R/\epsilon$, hence $M < 4(2R/\epsilon)^{2/s}$.

For part (ii), use Definition S5 and Theorem S6 with $S = (h)_{h\in\mathcal{H}_q}$, $\mathcal{H} = \mathcal{H}^{\mathrm{MR}}$, and $\mathcal{S}_M = \{S\}$. On the stated event, an accepted $y^\dagger \in [\rho_0, 1-\rho_0]^V$ gives $t = S(y^\dagger)$ and $\Delta_{\widehat{S},\widehat{A}}(X) \ge \Delta_S(t)$. Proposition S19 gives $\Delta_S(t) \ge c_\star\sqrt{V/M}$, where $c_\star = \rho_0/\sqrt{2}$. Thus Theorem S6 yields $\mathrm{rad}_{\mathcal{H}^{\mathrm{MR}}}\{\widehat{\mathcal{C}}^{(M)}(X)\} \ge \frac{1}{2}\Delta_{\widehat{S},\widehat{A}}(X) \ge \frac{c_\star}{2}\sqrt{V/M}$, proving (S.41). This completes the proof. □

S11.10 Proof of Theorem 4

The two weights have equal total mass, which gives the half-L^1 bound for the target coverage difference.

Proof. Condition on the data and randomization defining $\widehat{\mathcal{C}}$. Put $p_{\widehat{\mathcal{C}}}(x) = P_{\mathrm{src},Y|X=x}\{Y \in \widehat{\mathcal{C}}(x)\}$. Because r and $f \in \mathcal{F}$ integrate to one,

$$P_{\mathrm{tar}}\{Y \in \widehat{\mathcal{C}}(X)\} - P_f\{Y \in \widehat{\mathcal{C}}(X)\} = \mathbb{E}_{P_{\mathrm{src},X}}[(r-f)(p_{\widehat{\mathcal{C}}} - 1/2)] \ge -\frac{1}{2}\mathbb{E}_{P_{\mathrm{src},X}}|r-f|.$$

Choose $f_\varepsilon \in \mathcal{F}$ with $\frac{1}{2}\mathbb{E}_{P_{\mathrm{src},X}}|r-f_\varepsilon| \le \kappa(r,\mathcal{F}) + \varepsilon$. The definition gives $P_{f_\varepsilon}\{Y \in \widehat{\mathcal{C}}(X)\} \ge 1-\alpha-\Delta^-_{\mathcal{F}}(\widehat{\mathcal{C}})$. Hence $P_{\mathrm{tar}}\{Y \in \widehat{\mathcal{C}}(X)\} \ge 1-\alpha-\Delta^-_{\mathcal{F}}(\widehat{\mathcal{C}}) - \kappa(r,\mathcal{F}) - \varepsilon$. Let $\varepsilon \downarrow 0$ to obtain (12). This completes the proof. □

S11.11 Proofs of supporting results

Moment perturbations bound score error, and neighboring score events yield the threshold comparisons.

Proof of Proposition S10. For every (x,y,h),

$$\left|\frac{|h(y)-\widehat{\mu}_h(x)|}{\widehat{\sigma}_h(x)+\lambda} - \frac{|h(y)-\mu^\circ_h(x)|}{\sigma^\circ_h(x)+\lambda}\right| \le \frac{|\widehat{\mu}_h(x)-\mu^\circ_h(x)|}{\lambda} + \frac{|h(y)-\mu^\circ_h(x)|\,|\widehat{\sigma}_h(x)-\sigma^\circ_h(x)|}{\lambda^2}$$
$$\le \lambda^{-1}a + \lambda^{-2}M_{\mathcal{H}}b.$$

Taking the supremum over $h \in \mathcal{H}$ and using $|\sup_h u_h - \sup_h v_h| \le \sup_h |u_h - v_h|$ yields

$$\sup_{x\in\mathcal{X}}\sup_{y\in\mathcal{Y}}|R_n(x,y) - R^\circ(x,y)| \le \lambda^{-1}a + \lambda^{-2}M_{\mathcal{H}}b,$$

which is (S.35). This completes the proof. □

Proof of Lemma S4. If $\sup_{x\in\mathcal{X}}\sup_{y\in\mathcal{Y}}|R_n(x,y)-R^\circ(x,y)| \le e$, then $R^\circ(x,y)-e \le R_n(x,y) \le R^\circ(x,y)+e$ for every (x,y), which implies that

$$\{R^\circ(X,Y) \le q-e\} \subseteq \{R_n(X,Y) \le q\} \subseteq \{R^\circ(X,Y) \le q+e\},$$

and taking P_f-probabilities yields the result. This completes the proof. □

Proof of Lemma S5. Put $q_\star = q^\circ_{\alpha,\mathcal{F}}$ and $\Delta = \tau + \psi + e$. Choose $c_0 \le 1$. Then $\Delta \le c_0 r_0$ places $q = q_\star + e$ in the empirical-process interval. For every $f \in \mathcal{F}$, $q - e \ge q^\circ_{f,\alpha}$, and Lemma S4 gives $P_f\{R_n > q\} \le P_f\{R^\circ > q - e\} \le \alpha$. Hence $\Gamma^+_n(q;\mathcal{F}) \le \psi \le \tau$, and (S.30) gives $\widehat{q} \le q_\star + e$.

If $\Delta = 0$, the upper bound gives $\widehat{q} \le q_\star$; for $q_\star = 0$, equality follows from $\widehat{q} \ge 0$. For $q_\star > 0$, take $0 < s \le (r_0 \wedge q_\star)/2$, put $q_s = q_\star - s$, and choose $f_s \in \mathcal{F}$ with $q^\circ_{f_s,\alpha} \ge q_\star - s/4$. Lemma S4 and Assumption S1, with $e = \psi = \tau = 0$, give

$$P_{f_s}\{R_n(X,Y) > q_s\} \ge P_{f_s}\{R^\circ(X,Y) > q_s\} \ge \alpha + c_- s/2.$$

Thus $\Gamma^+_n(q_s;\mathcal{F}) > 0 = \tau$. Monotonicity excludes every $q \le q_s$, hence $\widehat{q} \ge q_\star - s$. Let $s \downarrow 0$ to obtain $\widehat{q} = q_\star$.

Suppose $\Delta > 0$. If $q_\star \le C_1\Delta$, then $\widehat{q} \ge 0 \ge q_\star - C_1\Delta$. For $q_\star > C_1\Delta$, take $0 < s \le (r_0 \wedge q_\star)/2$, put $q_s = q_\star - s \ge 0$, and choose $f_s \in \mathcal{F}$ with $q^\circ_{f_s,\alpha} \ge q_\star - s/4$. If $s \ge 4e$, then $q^\circ_{f_s,\alpha} \ge q_\star - r_0$ and $q^\circ_{f_s,\alpha} - (q_s + e) \in [s/2, r_0]$. Lemma S4 and Assumption S1 give

$$P_{f_s}\{R_n(X,Y) > q_s\} \ge P_{f_s}\{R^\circ(X,Y) > q_s + e\} \ge \alpha + c_- s/2.$$

The empirical-process bound at $q_s \in [q_\star - r_0, q_\star + r_0]$ gives $\Gamma^+_n(q_s;\mathcal{F}) \ge c_- s/2 - \psi$. For $s > 2(\tau + \psi)/c_-$ and $s \ge 4e$, monotonicity excludes all $q \le q_s$. Choose $C_2 > 4 \vee 2/c_-$, $s = C_2\Delta$, $C_1 \ge 2C_2$, and $c_0 \le (2C_2)^{-1}$. These choices ensure $s \le r_0/2$, $s \le q_\star/2$, $s \ge 4e$, and $s > 2(\tau + \psi)/c_-$. Thus $\widehat{q} \ge q_\star - C\Delta$, where C depends only on c_-. Combining the bounds completes the proof. □

Proof of Proposition S36. Proposition S11 implies that, for every x, $\mathrm{rad}_{\mathcal{H}}\{\widehat{\mathcal{C}}(x)\} = \mathrm{rad}_{\mathcal{H}}\{\mathcal{C}_{\widehat{q}}(x)\} \le A(x)\widehat{q} \le A(x)(q^\circ_{\alpha,\mathcal{F}} + t_n)$. Taking expectations under any $f \in \mathcal{F}$ and then the supremum over f yields $\mathcal{S}_{\mathcal{F},\mathcal{H}}(\widehat{\mathcal{C}}) \le \sup_{f\in\mathcal{F}} \mathbb{E}_f\{A(X)\}(q^\circ_{\alpha,\mathcal{F}} + t_n) \le A_{\mathcal{F}}(q^\circ_{\alpha,\mathcal{F}} + t_n)$. This completes the proof. □

Proof of Proposition S17. Let $M_n(\theta) = \widehat{\Lambda}(\theta) - n_{\mathrm{tar}}^{-1}\sum_{j=1}^{n_{\mathrm{tar}}} \theta^\top \phi(X^{\mathrm{tar}}_j)$, where $\widehat{\Lambda}(\theta) = \log\{n_{\mathrm{src}}^{-1}\sum_{i=1}^{n_{\mathrm{src}}} \exp(\theta^\top\phi(X_i))\}$. Let B_ϕ bound $\|\phi(X)\|$, and put $M_\Theta = \sup_{\theta\in\Theta_{\mathrm{tilt}}}\|\theta\|$ and $D_\Theta = \mathrm{diam}(\Theta_{\mathrm{tilt}})$. The denominator of $\nabla\widehat{\Lambda}$ is at least $e^{-M_\Theta B_\phi}$. The classes $g_{0,\theta}(x) = \exp\{\theta^\top\phi(x)\}$ and $g_{1,\theta,\ell}(x) = \phi_\ell(x)\exp\{\theta^\top\phi(x)\}$, $1 \le \ell \le d$, are bounded and Lipschitz in θ, hence Euclidean. Their constants depend only on $B_\phi, M_\Theta, D_\Theta, d$ and the fixed metric-entropy bound for Θ_{tilt}. The bounded Euclidean-class maximal inequality (see, e.g., van der Vaart & Wellner, 1996, Chapter 2.14), McDiarmid's inequality, and a union bound over coordinates give, with probability at least $1 - \delta$,

$$\sup_{\theta\in\Theta_{\mathrm{tilt}}}\left|n_{\mathrm{src}}^{-1}\sum_{i=1}^{n_{\mathrm{src}}} g_{0,\theta}(X_i) - \mathbb{E}_{P_{\mathrm{src},X}} g_{0,\theta}(X)\right| + \sup_{\theta\in\Theta_{\mathrm{tilt}}}\max_{1\le\ell\le d}\left|n_{\mathrm{src}}^{-1}\sum_{i=1}^{n_{\mathrm{src}}} g_{1,\theta,\ell}(X_i) - \mathbb{E}_{P_{\mathrm{src},X}} g_{1,\theta,\ell}(X)\right|$$
$$\le C_1\sqrt{\frac{1+\log(1/\delta)}{n_{\mathrm{src}}}},$$

and

$$\left\|n_{\mathrm{tar}}^{-1}\sum_{j=1}^{n_{\mathrm{tar}}}\phi(X^{\mathrm{tar}}_j) - \mathbb{E}_{P_{\mathrm{tar},X}}\phi(X)\right\| \le C_1\sqrt{\frac{1+\log(1/\delta)}{n_{\mathrm{tar}}}}.$$

The maps $u \mapsto \log u$ and $(v,u) \mapsto v/u$ are Lipschitz for $u \ge e^{-M_\Theta B_\phi}$ and bounded v. Thus

$$\sup_{\theta\in\Theta_{\mathrm{tilt}}}\|\nabla M_n(\theta) - \nabla M(\theta)\| \le C_1\Big\{\sqrt{\frac{1+\log(1/\delta)}{n_{\mathrm{src}}}} + \sqrt{\frac{1+\log(1/\delta)}{n_{\mathrm{tar}}}}\Big\},$$

and

$$\sup_{\theta\in\Theta_{\mathrm{tilt}}}|M_n(\theta) - M(\theta)| \le C_1\Big\{\sqrt{\frac{1+\log(1/\delta)}{n_{\mathrm{src}}}} + \sqrt{\frac{1+\log(1/\delta)}{n_{\mathrm{tar}}}}\Big\},$$

with C_1 depending only on the stated feature, parameter, and entropy bounds. Let Δ_n be the right side, and use $r_{\mathrm{loc}}, m_{\mathrm{loc}}, \eta_{\mathrm{loc}}$ from the proposition. If $\Delta_n \le \eta_{\mathrm{loc}}/3$, then $M_n(\widehat{\theta}) \le M_n(\theta_0)$ gives

$$M(\widehat{\theta}) \le M_n(\widehat{\theta}) + \Delta_n \le M_n(\theta_0) + \Delta_n \le M(\theta_0) + 2\Delta_n < M(\theta_0) + \eta_{\mathrm{loc}}.$$

Hence $\widehat{\theta} \in B(\theta_0, r_{\mathrm{loc}})$ is interior, and $\nabla M_n(\widehat{\theta}) = \nabla M(\theta_0) = 0$. Local strong convexity gives

$$m_{\mathrm{loc}}\|\widehat{\theta} - \theta_0\|^2 \le (\widehat{\theta} - \theta_0)^\top \{\nabla M(\widehat{\theta}) - \nabla M(\theta_0)\} \le \|\widehat{\theta} - \theta_0\| \sup_{\theta \in B(\theta_0, r_{\mathrm{loc}})} \|\nabla M_n(\theta) - \nabla M(\theta)\|.$$

Thus $\|\widehat{\theta} - \theta_0\| \le \Delta_n / m_{\mathrm{loc}}$. For $\Delta_n > \eta_{\mathrm{loc}}/3$, compactness gives $\|\widehat{\theta} - \theta_0\| \le D_\Theta \le (3D_\Theta/\eta_{\mathrm{loc}})\Delta_n$. Both cases give the stated rate.

If $\widehat{\theta}$ is an interior minimizer, the first-order condition yields $\nabla M_n(\widehat{\theta}) = 0$, which gives

$$\frac{\sum_{i=1}^{n_{\mathrm{src}}} \phi(X_i) e^{\widehat{\theta}^\top \phi(X_i)}}{\sum_{i=1}^{n_{\mathrm{src}}} e^{\widehat{\theta}^\top \phi(X_i)}} = \frac{1}{n_{\mathrm{tar}}} \sum_{j=1}^{n_{\mathrm{tar}}} \phi(X_j^{\mathrm{tar}}),$$

and rewriting the left-hand side as $n_{\mathrm{src}}^{-1} \sum_{i=1}^{n_{\mathrm{src}}} \widehat{r}(X_i)\phi(X_i)$ yields the moment equation. This completes the proof. □

S12 Remaining proofs

Query approximation and reused nuisance fits contribute different terms to calibration and radius error. The arguments establish these terms from the stated population conditions and prove the specified calibration and query lower bounds.

S12.1 Proof of Proposition S15

Averaging the auxiliary randomization gives the candidate-dependent exclusion probability.

Proof. By (S.29), exclusion is $\{U_0 < \widehat{a}_{\mathrm{fd}}^+(x, y)\}$ up to a null boundary. Conditional on $(\mathcal{D}_{\mathrm{tr}}, \mathcal{D}_{\mathrm{cal}}, x, y, \mathcal{G}_{\mathrm{sc}})$, the independent uniform variable U_0 gives this event probability $\widehat{a}_{\mathrm{fd}}^+(x, y)$. This proves (S.43) and completes the proof. □

The following lemma provides uniform concentration bounds for the query estimates. Let $\widetilde{\mu}_h(x) = \mathbb{E}_{Q_x^{\mathrm{T}}} h(Y)$ and $\widetilde{\sigma}_h(x) = \mathrm{Var}_{Q_x^{\mathrm{T}}}\{h(Y)\}^{1/2}$. Conditional on $\{Q_x^{\mathrm{T}} : x \in \mathcal{X}_q\}$, suppose that for each fixed $x \in \mathcal{X}_q$ the variables $\{\widetilde{Y}^{(k)}(x) : k = 1, \dots, K\}$ are i.i.d. with common distribution Q_x^{T} across k. Define $\widehat{\mu}_h(x) = K^{-1} \sum_{k=1}^K h\{\widetilde{Y}^{(k)}(x)\}$ and $\widehat{\sigma}_h^2(x) = K^{-1} \sum_{k=1}^K [h\{\widetilde{Y}^{(k)}(x)\} - \widehat{\mu}_h(x)]^2$. For $0 < \delta < 1$, put $L_{q,\delta} = 1 + \log(|\mathcal{X}_q||\mathcal{H}_q|/\delta)$. Here $\|Z\|_{\psi_2(P)} = \inf\{c > 0 : \mathbb{E}_P \exp(Z^2/c^2) \le 2\}$; its definition also implies that Z^2 is sub-exponential.

Lemma S6. *Suppose $\mathcal{X}_q, \mathcal{H}_q$ are finite. In the bounded case, assume $Q_x^{\mathrm{T}}\{|h(Y)| \le B_{\mathcal{H}}\} = 1$ for every x, h. Then, for C depending only on $B_{\mathcal{H}}$, with probability at least $1 - \delta$,*

$$\sup_{x \in \mathcal{X}_q,\, h \in \mathcal{H}_q} |\widehat{\mu}_h(x) - \widetilde{\mu}_h(x)| + \sup_{x \in \mathcal{X}_q,\, h \in \mathcal{H}_q} |\widehat{\sigma}_h(x) - \widetilde{\sigma}_h(x)| \le C\sqrt{\frac{L_{q,\delta}}{K}}.$$

If instead $\sup_{x,h} \|h(Y)\|_{\psi_2(Q_x^{\mathrm{T}})} \le B_{\mathcal{H}}$ and $\inf_{x,h} \widetilde{\sigma}_h(x) \ge \sigma_{\min} > 0$, then, for C also depending on $\sigma_{\min}$, with probability at least $1 - \delta$,

$$\sup_{x \in \mathcal{X}_q,\, h \in \mathcal{H}_q} |\widehat{\mu}_h(x) - \widetilde{\mu}_h(x)| + \sup_{x \in \mathcal{X}_q,\, h \in \mathcal{H}_q} |\widehat{\sigma}_h(x) - \widetilde{\sigma}_h(x)| \le C\left\{\sqrt{\frac{L_{q,\delta}}{K}} + \frac{L_{q,\delta}}{K}\right\}.$$

Lemma S6 permits zero population variance for bounded functionals. A variance-sensitive argument gives the standard-deviation rate directly. The positive scale is used only in the sub-Gaussian extension.

Proof. For bounded queries at one pair (x, h), put $Z_k = h\{\widetilde{Y}^{(k)}(x)\}$, $\mu = \mathbb{E}Z_k$, $\sigma^2 = \mathrm{Var}Z_k$, $M_2 = K^{-1}\sum_k (Z_k - \mu)^2$, and $v = B_{\mathcal{H}}\sqrt{t/K}$. Because $|Z_k - \mu| \le 2B_{\mathcal{H}}$,

$$\mathrm{Var}\{(Z_k - \mu)^2\} \le 4B_{\mathcal{H}}^2\sigma^2, \qquad |M_2 - \sigma^2| \le C_1\sigma v + C_2 v^2$$

with probability at least $1 - 2e^{-t}$ by Bernstein's inequality (Boucheron et al., 2013). For $\sigma \geq v$, divide by $\sqrt{M_2} + \sigma \geq \sigma$ to obtain $|\sqrt{M_2} - \sigma| \leq (C_1 + C_2)v$. For $\sigma < v$, $M_2 \leq (1 + C_1 + C_2)v^2$ gives the same bound with a larger constant. The distance-to-constants representation of empirical standard deviation, the reverse triangle inequality, and Hoeffding's inequality give

$$|\widehat{\sigma} - \sqrt{M_2}| \leq |\widehat{\mu} - \mu| \leq Cv$$

except on an event of probability at most $2e^{-t}$. Choose $t = \log(4|\mathcal{X}_q||\mathcal{H}_q|/\delta)$ and take a union bound. This proves the bounded assertion, including $\sigma = 0$.

For the second-moment calculation, put $r_K = \sqrt{L_{q,\delta}/K}$ and $s_K = L_{q,\delta}/K$. Bernstein's inequality and a union bound give $\sup_{x \in \mathcal{X}_q,\, h \in \mathcal{H}_q} |\widehat{\mu}_h(x) - \widetilde{\mu}_h(x)| \leq C_1 r_K$ with probability at least $1 - \delta/2$. Let $\widetilde{m}_{2,h}(x) = \mathbb{E}_{Q_x^{\mathrm{T}}}\{h(Y)^2\}$ and $\widehat{m}_{2,h}(x) = K^{-1}\sum_{k=1}^{K} h\{\widetilde{Y}^{(k)}(x)\}^2$. Applying Bernstein's inequality to the centered squares gives the uniform second-moment bound with probability at least $1 - \delta/2$. On the intersection,

$$\left|\widehat{\sigma}_h^2(x) - \widetilde{\sigma}_h^2(x)\right| \leq |\widehat{m}_{2,h}(x) - \widetilde{m}_{2,h}(x)| + \left|\widehat{\mu}_h(x)^2 - \widetilde{\mu}_h(x)^2\right|.$$

For bounded queries, the second-moment error is at most $C_2 r_K$ and $|\widehat{\mu}_h(x)| + |\widetilde{\mu}_h(x)| \leq 2B_{\mathcal{H}}$, hence $\left|\widehat{\sigma}_h^2(x) - \widetilde{\sigma}_h^2(x)\right| \leq C_3 r_K$ uniformly over (x, h). For sub-Gaussian queries, the squares are sub-exponential, giving error $C_4(r_K + s_K)$. Because $|\widetilde{\mu}_h(x)| \leq C_5 B_{\mathcal{H}}$ and $|\widehat{\mu}_h(x)| \leq C_5 B_{\mathcal{H}} + C_1 r_K$,

$$\left|\widehat{\mu}_h(x)^2 - \widetilde{\mu}_h(x)^2\right| \leq C_6(r_K + s_K).$$

Hence the sub-Gaussian variance error satisfies $\left|\widehat{\sigma}_h^2(x) - \widetilde{\sigma}_h^2(x)\right| \leq C_7(r_K + s_K)$ uniformly over (x, h). Use

$$|\widehat{\sigma}_h(x) - \widetilde{\sigma}_h(x)| \leq \frac{\left|\widehat{\sigma}_h^2(x) - \widetilde{\sigma}_h^2(x)\right|}{\widehat{\sigma}_h(x) + \widetilde{\sigma}_h(x)}$$

and $\widetilde{\sigma}_h(x) \geq \sigma_{\min}$ for the stated standard-deviation bound. This completes the proof of Lemma S6. □

Proof of Proposition S28. Condition on the nuisance fits and on the deterministic equal-size fold partition. For $k = 1, \ldots, K_{\mathrm{CV}}$, write

$$P_{n,k} g = |\mathcal{I}_{\mathrm{cal}}^{(k)}|^{-1} \sum_{i \in \mathcal{I}_{\mathrm{cal}}^{(k)}} g(X_i, Y_i), \qquad P g = \mathbb{E}_{P_{\mathrm{src}}} g(X, Y).$$

Because the folds have equal size, the left-hand side of (S.65) is bounded by the sum of

$$\sup_{f \in \mathcal{F}, q \in \mathbb{R}} \left| K_{\mathrm{CV}}^{-1} \sum_{k=1}^{K_{\mathrm{CV}}} (P_{n,k} - P) \left[\{f + \widehat{r}^{(-k)}\} \{\mathbf{1}\{R_n^{(-k)} > q\} - \alpha\} \right] \right|$$

and

$$|\overline{\widehat{r}}| \sup_{q \in \mathbb{R}} \left| K_{\mathrm{CV}}^{-1} \sum_{k=1}^{K_{\mathrm{CV}}} (P_{n,k} - P) \{\mathbf{1}\{R_n^{(-k)} > q\} - \alpha\} \right|.$$

The latter term is bounded by $B_{\mathrm{DR},0}$ times the same process with multiplier class $\{1\}$.

For the former term, define

$$\mathfrak{G} = \left\{ (k, x, y) \mapsto \{f(x) + \widehat{r}^{(-k)}(x)\} \{\mathbf{1}\{R_n^{(-k)}(x, y) > q\} - \alpha\} : f \in \mathcal{F},\ q \in \mathbb{R} \right\}.$$

Fix a fold k_0 and enlarge $\mathfrak{G}$ by allowing $f = w - \widehat{r}^{(-k_0)}$, $w \in \mathcal{W}_{\mathrm{DR},0}$. The original class is included because $f + \widehat{r}^{(-k_0)} \in \mathcal{W}_{\mathrm{DR},0}$. For fixed $f_0 \in \mathcal{F}$, $|\widehat{r}^{(-k)} - \widehat{r}^{(-k_0)}| = |(f_0 + \widehat{r}^{(-k)}) - (f_0 + \widehat{r}^{(-k_0)})| \leq 2B_{\mathrm{DR},0}$, giving multiplier bound $3B_{\mathrm{DR},0}$. Translate a countable dense subclass of $\mathcal{W}_{\mathrm{DR},0}$ and take rational thresholds decreasing to q. These give a countable pointwise dense subclass of the enlarged $\mathfrak{G}$.

For a probability measure Q on (k, x, y) with covariate marginal Q_X, the multiplier distance is $\|w - w'\|_{L_2(Q_X)}$. An $\eta B_{\mathrm{DR},0}/2$-net controls all folds with the same index. The nested threshold sets

admit an $\eta^2/36$-net of attainable tail probabilities with at most $1 + 36/\eta^2$ points by Lemma S12. The product error is at most $\eta B_{\mathrm{DR},0}/2 + 3B_{\mathrm{DR},0}\sqrt{\eta^2/36} = \eta B_{\mathrm{DR},0}$, also for signed multipliers. Thus

$$\sup_Q N\{\eta B_{\mathrm{DR},0}, \mathfrak{G}, L_2(Q)\} \le C K_{\mathrm{CV}} (2A_{\mathrm{DR}}/\eta)^{v_{\mathrm{DR}}} \eta^{-2}.$$

Let

$$Z_{\mathfrak{G}} = \sup_{g\in\mathfrak{G}} \left| \frac{1}{n_{\mathrm{cal}}} \sum_{i\in\mathcal{I}_{\mathrm{cal}}} \{g(k(i), X_i, Y_i) - \mathbb{E}_{P_{\mathrm{src}}} g(k(i), X, Y)\} \right|.$$

Conditional on the nuisance fits, the (X_i, Y_i) are independent source observations and the $k(i)$ are fixed. Symmetrization gives

$$\mathbb{E} Z_{\mathfrak{G}} \le 2\mathbb{E}\left[\sup_{g\in\mathfrak{G}} \left| \frac{1}{n_{\mathrm{cal}}} \sum_{i\in\mathcal{I}_{\mathrm{cal}}} \varepsilon_i g(k(i), X_i, Y_i) \right| \right],$$

where the ε_i are independent Rademacher variables, independent of calibration. An empirical L_2 net of radius $\eta B_{\mathrm{DR},0}$ has cardinality at most $C K_{\mathrm{CV}} (2A_{\mathrm{DR}}/\eta)^{v_{\mathrm{DR}}} \eta^{-2}$. The finite-net maximal inequality gives

$$\mathbb{E} Z_{\mathfrak{G}} \le C B_{\mathrm{DR},0} \sqrt{\frac{(v_{\mathrm{DR}} + 2)\log(A_{\mathrm{DR}} n_{\mathrm{cal}}) + \log K_{\mathrm{CV}}}{n_{\mathrm{cal}}}}.$$

Replacing one calibration observation changes $Z_{\mathfrak{G}}$ by at most $6B_{\mathrm{DR},0}/n_{\mathrm{cal}}$. McDiarmid's inequality gives, with conditional probability at least $1 - \delta/2$,

$$Z_{\mathfrak{G}} \le C B_{\mathrm{DR},0} \sqrt{\frac{(v_{\mathrm{DR}} + 2)\log(A_{\mathrm{DR}} n_{\mathrm{cal}}) + \log(2K_{\mathrm{CV}}/\delta)}{n_{\mathrm{cal}}}}.$$

Apply the same bound to the singleton scalar-centering class at failure probability $\delta/2$. A union bound gives conditional probability at least $1 - \delta$. Multiplication by $|\widehat{\bar{r}}| \le B_{\mathrm{DR},0}$ and addition give (S.65) with $v_{\mathrm{DR}} + 3$, absorbing the extra $\log 2$ into the universal constant. This completes the proof. □

Proof of Proposition S29. The equal fold sizes give

$$n_{\mathrm{cal}}^{-1} \sum_{i\in\mathcal{I}_{\mathrm{cal}}} a_i = K_{\mathrm{CV}}^{-1} \sum_{k=1}^{K_{\mathrm{CV}}} |\mathcal{I}_{\mathrm{cal}}^{(k)}|^{-1} \sum_{i\in\mathcal{I}_{\mathrm{cal}}^{(k)}} a_i.$$

Put $P_{n,k} = |\mathcal{I}_{\mathrm{cal}}^{(k)}|^{-1} \sum_{i\in\mathcal{I}_{\mathrm{cal}}^{(k)}} \delta_{(X_i,Y_i)}$ and $P = P_{\mathrm{src}}$. Conditional on $\mathcal{A}_k$, the fold observations are i.i.d. from P and $\widehat{r}^{(-k)}$, $R_n^{(-k)}$ are fixed. Define

$$\mathfrak{G}_k = \left\{ (x,y) \mapsto \{f(x) + \widehat{r}^{(-k)}(x)\}\{\mathbf{1}\{R_n^{(-k)}(x,y) > q\} - \alpha\} : f \in \mathcal{F},\ q \in \mathbb{R} \right\}$$

and

$$\mathfrak{H}_k = \left\{ (x,y) \mapsto \mathbf{1}\{R_n^{(-k)}(x,y) > q\} - \alpha : q \in \mathbb{R} \right\}.$$

The class $\mathfrak{G}_k$ is bounded by $B_{\mathrm{DR},K}$. Lemma S12's rational-threshold argument gives pointwise measurability of $\mathfrak{G}_k, \mathfrak{H}_k$. Applying the same lemma to $\mathcal{W}_{\mathrm{DR},k}$ gives

$$\sup_Q N\{\eta B_{\mathrm{DR},K}, \mathfrak{G}_k, L_2(Q)\} \le C(2A_{\mathrm{DR}}/\eta)^{v_{\mathrm{DR}}} \eta^{-2}.$$

For the singleton multiplier class $\{1\}$, the threshold net gives entropy $C\eta^{-2}$ for $\mathfrak{H}_k$. Apply Lemma S11 conditional on $\mathcal{A}_k$, at level $\delta/(2K_{\mathrm{CV}})$ for each class, to obtain

$$\sup_{g\in\mathfrak{G}_k} |(P_{n,k} - P)g| \le C B_{\mathrm{DR},K} \sqrt{\frac{(v_{\mathrm{DR}} + 3)\log(A_{\mathrm{DR}} n_{\mathrm{cal}}) + \log(K_{\mathrm{CV}}/\delta)}{n_{\mathrm{fold}}}}$$

and

$$\sup_{g\in\mathfrak{H}_k} |(P_{n,k} - P)g| \le C \sqrt{\frac{3\log n_{\mathrm{cal}} + \log(K_{\mathrm{CV}}/\delta)}{n_{\mathrm{fold}}}}$$

with conditional probability at least $1-\delta/K_{\mathrm{CV}}$. Integrating and taking a union bound over folds gives simultaneous probability at least $1-\delta$.

On this event, the left-hand side of (S.65) is bounded by

$$K_{\mathrm{CV}}^{-1}\sum_{k=1}^{K_{\mathrm{CV}}}\sup_{g\in\mathfrak{G}_k}|(P_{n,k}-P)g|+|\widehat{\bar{r}}|\,K_{\mathrm{CV}}^{-1}\sum_{k=1}^{K_{\mathrm{CV}}}\sup_{g\in\mathfrak{H}_k}|(P_{n,k}-P)g|.$$

Use $|\widehat{\bar{r}}|\le B_{\mathrm{DR},K}$, $B_{\mathrm{DR},K}\ge 1$, and $n_{\mathrm{fold}}=n_{\mathrm{cal}}/K_{\mathrm{CV}}$ to bound this by $\psi_{n,K}^{\mathrm{DR}}(\delta)$, enlarging the universal constant. This proves (S.65) and completes the proof. □

Proof of Proposition S30. For every $\theta\in\Theta_{\mathrm{tilt}}$ and every $x\in\mathcal{X}$,

$$|\theta^\top\phi(x)|\le M_\Theta B_\phi.$$

Both $\widehat{\Lambda}^{(-k)}(\theta)$ and $\Lambda(\theta)$ then belong to $[-M_\Theta B_\phi, M_\Theta B_\phi]$, which gives

$$0\le\widehat{r}^{(-k)}(x)\le B_{\mathrm{tilt}}\qquad\text{for every } x\in\mathcal{X},\ k=1,\dots,K_{\mathrm{CV}}.$$

Both translated classes are bounded by $B_F+B_{\mathrm{tilt}}\le B_{\mathrm{DR}}^{\mathrm{tilt}}$, and averaging gives $|\widehat{\bar{r}}|\le B_{\mathrm{DR}}^{\mathrm{tilt}}$.

Fix Q and $0<\eta\le 1$. An ηB_F-net $\mathcal{F}_\eta$ of $\mathcal{F}$ in $L_2(Q)$ has cardinality at most $(A_F/\eta)^{v_F}$. Because $B_{\mathrm{DR}}^{\mathrm{tilt}}\ge B_F$, each translate $\mathcal{F}_\eta+\widehat{r}^{(-k)}$ covers its corresponding class at radius $\eta B_{\mathrm{DR}}^{\mathrm{tilt}}$. Taking the union over folds gives

$$N\{\eta B_{\mathrm{DR}}^{\mathrm{tilt}},\mathcal{W}_{\mathrm{DR},0}^{\mathrm{tilt}},L_2(Q)\}\le K_{\mathrm{CV}}(A_F/\eta)^{v_F}\le(A_{\mathrm{DR}}^{\mathrm{tilt}}/\eta)^{v_{\mathrm{DR}}^{\mathrm{tilt}}},$$

where $A_{\mathrm{DR}}^{\mathrm{tilt}}=2K_{\mathrm{CV}}(A_F\vee 1)$ and $\eta\le 1$ give the last inequality. A single translate gives the bound for $\mathcal{W}_{\mathrm{DR},k}^{\mathrm{tilt}}$. Measurability of ϕ and a finite union of translated countable dense subclasses of $\mathcal{F}$ give pointwise measurability. This completes the proof. □

Proof of Corollary S5. By Proposition S29, the empirical-process event (S.65) holds with bound $\psi_{n,K}^{\mathrm{DR}}(\delta)$ and probability at least $1-\delta$. On this event, the deterministic implication proved in Theorem S19 gives the stated bound. This completes the proof. □

Proof of Proposition S31. For every $q\in\mathbb{R}$,

$$\begin{aligned}&\mathbb{E}_{P_{\mathrm{src}}}\Big|r(X)\bar{M}(q)-K_{\mathrm{CV}}^{-1}\sum_{k=1}^{K_{\mathrm{CV}}}\{f_{\mathrm{aug}}(X)+\widehat{r}^{(-k)}(X)-\widehat{\bar{r}}\}M_k(q)\Big|\\&\le K_{\mathrm{CV}}^{-1}\sum_{k=1}^{K_{\mathrm{CV}}}\mathbb{E}_{P_{\mathrm{src},X}}\Big|r(X)-f_{\mathrm{aug}}(X)-\widehat{r}^{(-k)}(X)+\widehat{\bar{r}}\Big|,\end{aligned}$$

where we have used $0\le M_k(q)\le 1$. Taking the supremum over q and then evaluating the infimum in (S.64) at f_{aug} gives

$$\rho_n^{\mathrm{aug}}\le\eta_n^{\mathrm{aug}}+\alpha\Big|\widehat{\bar{r}}-\bar{r}_n^P\Big|\le\eta_n^{\mathrm{aug}}+\alpha\eta_n^{\mathrm{cent}}.$$

This completes the proof. □

Proof of Proposition S32. The ratios satisfy $r(x),\widehat{r}^{(-k)}(x)\le B_{\mathrm{tilt}}=\exp(2M_\Theta B_\phi)$, where $M_\Theta=\sup_{\theta\in\Theta_{\mathrm{tilt}}}\|\theta\|$. Proposition S17 and a union bound give, with probability at least $1-\delta/3$,

$$\|\widehat{\theta}^{(-k)}-\theta_0\|\le C\left\{\sqrt{\frac{1+\log(6K_{\mathrm{CV}}/\delta)}{n_{\mathrm{src},k}}}+\sqrt{\frac{1+\log(6K_{\mathrm{CV}}/\delta)}{n_{\mathrm{tar},k}}}\right\}$$

for all folds k. The Euclidean-class maximal inequality in Proposition S17's proof and another union bound give, with probability at least $1-\delta/3$,

$$\sup_{\theta\in\Theta_{\mathrm{tilt}}}\Big|\widehat{\Lambda}^{(-k)}(\theta)-\Lambda(\theta)\Big|\le C\sqrt{\frac{1+\log(6K_{\mathrm{CV}}/\delta)}{n_{\mathrm{src},k}}}$$

for every $k = 1, \dots, K_{\mathrm{CV}}$. On the intersection of these two events,

$$\begin{aligned}\sup_{x\in\mathcal{X}} \left|\log \widehat{r}^{(-k)}(x) - \log r(x)\right| &\le B_\phi \|\widehat{\theta}^{(-k)} - \theta_0\| + \left|\widehat{\Lambda}^{(-k)}(\widehat{\theta}^{(-k)}) - \Lambda(\widehat{\theta}^{(-k)})\right| \\ &\quad + \left|\Lambda(\widehat{\theta}^{(-k)}) - \Lambda(\theta_0)\right| \\ &\le C\left\{\sqrt{\frac{1+\log(6K_{\mathrm{CV}}/\delta)}{n_{\mathrm{src},k}}} + \sqrt{\frac{1+\log(6K_{\mathrm{CV}}/\delta)}{n_{\mathrm{tar},k}}}\right\},\end{aligned}$$

because Λ is B_ϕ-Lipschitz. The exponential map is Lipschitz on $[-2M_\Theta B_\phi, 2M_\Theta B_\phi]$, hence

$$\mathbb{E}_{P_{\mathrm{src},X}}\left|\widehat{r}^{(-k)}(X) - r(X)\right| \le C\left\{\sqrt{\frac{1+\log(6K_{\mathrm{CV}}/\delta)}{n_{\mathrm{src},k}}} + \sqrt{\frac{1+\log(6K_{\mathrm{CV}}/\delta)}{n_{\mathrm{tar},k}}}\right\}.$$

Next, define

$$\overline{r}_k = n_{\mathrm{fold}}^{-1} \sum_{i\in\mathcal{I}_{\mathrm{cal}}^{(k)}} \widehat{r}^{(-k)}(X_i).$$

Conditional on the data defining $\widehat{r}^{(-k)}$, its fold evaluations are i.i.d. and bounded by B_{tilt}. Hoeffding's inequality and a union bound give, with probability at least $1 - \delta/3$,

$$\left|\overline{r}_k - \mathbb{E}_{P_{\mathrm{src},X}}\widehat{r}^{(-k)}(X)\right| \le C\sqrt{\frac{\log(6K_{\mathrm{CV}}/\delta)}{n_{\mathrm{fold}}}}$$

for every $k = 1, \dots, K_{\mathrm{CV}}$. Because $\overline{\widehat{r}} = K_{\mathrm{CV}}^{-1}\sum_{k=1}^{K_{\mathrm{CV}}} \overline{r}_k$, this implies

$$\left|\overline{\widehat{r}} - \bar{r}_n^P\right| \le C\sqrt{\frac{K_{\mathrm{CV}}\log(6K_{\mathrm{CV}}/\delta)}{n_{\mathrm{cal}}}}.$$

Finally, evaluating Proposition S31 at $f_{\mathrm{aug}} \equiv 1$ and using $\mathbb{E}_{P_{\mathrm{src},X}} r(X) = 1$, we obtain that

$$\begin{aligned}&K_{\mathrm{CV}}^{-1}\sum_{k=1}^{K_{\mathrm{CV}}} \mathbb{E}_{P_{\mathrm{src},X}}\left|r(X) - 1 - \widehat{r}^{(-k)}(X) + \overline{\widehat{r}}\right| \\ &\le K_{\mathrm{CV}}^{-1}\sum_{k=1}^{K_{\mathrm{CV}}} \mathbb{E}_{P_{\mathrm{src},X}}\left|r(X) - \widehat{r}^{(-k)}(X)\right| + \left|\overline{\widehat{r}} - 1\right| \\ &\le 2K_{\mathrm{CV}}^{-1}\sum_{k=1}^{K_{\mathrm{CV}}} \mathbb{E}_{P_{\mathrm{src},X}}\left|r(X) - \widehat{r}^{(-k)}(X)\right| + \left|\overline{\widehat{r}} - \bar{r}_n^P\right|.\end{aligned}$$

Combining the preceding bounds with Proposition S31 yields

$$\rho_n^{\mathrm{aug}} \le C\zeta_n^{\mathrm{tilt}}(\delta),$$

after enlarging C. This completes the proof. □

The following lemma establishes a minimax lower bound for the threshold estimation error. Let $\mathcal{Q}_K$ be the class of threshold estimators measurable with respect to the K teacher samples.

Lemma S7. *Suppose the bounded two-point query submodel in Theorem S20 holds. Then there exists a constant $c > 0$ such that $\inf_{\widetilde{q}\in\mathcal{Q}_K}\sup_{P\in\mathcal{P}_{\mathrm{bin}}}\mathbb{E}_P|\widetilde{q} - q^\circ_{\alpha,\mathcal{F},q}| \ge cB_{\mathcal{H}}K^{-1/2}$.*

The information restriction to predictor samples is part of this lower bound. The statement does not cover estimators with additional source observations.

Proof. Take $\mathcal{X} = \{x_0\}$, $\mathcal{F} = \{1\}$, and $\mathcal{H}_q = \{h_0\}$. Give y_+, y_- source probabilities $p, 1-p$. Then $h_0(Y)$ has mean $B_{\mathcal{H}}(2p-1)$ and standard deviation $2B_{\mathcal{H}}\sqrt{p(1-p)}$. For $\max\{1/2, 1-\alpha\} < p < 1$, the oracle quantile is $g(p) = 2B_{\mathcal{H}}/\{2B_{\mathcal{H}}\sqrt{p/(1-p)} + \lambda/(1-p)\}$. Both denominator terms increase,

making g continuous and strictly decreasing. Fix an interior value q_0. For small $|\vartheta|$, assign source probability $g^{-1}(q_0+\vartheta)$ and teacher probability $1/2+\vartheta/(2B_{\mathcal{H}})$ to y_+. This realizes $q^{\circ}_{\alpha,\mathcal{F},q} = q_0+\vartheta$ on the fixed support $\{y_-, y_+\}$ without the upper bound's moment-matching condition.

For any stated two-point instance, put $\vartheta_\pm = \pm c_1 B_{\mathcal{H}} K^{-1/2}$ and $p_\pm = 1/2 + \vartheta_\pm/(2B_{\mathcal{H}})$. Keep the source distributions realizing $q_0+\vartheta_\pm$ and replace their teachers by $p_\pm\delta_{y_+}+(1-p_\pm)\delta_{y_-}$. Both models remain in $\mathcal{P}_{\mathrm{bin}}$. Because $h_0(y_-) \neq h_0(y_+)$, full-image queries on this fixed support are equivalent to Bernoulli labels. Their K-sample Kullback–Leibler divergence is at most $CK(p_+-p_-)^2 \le Cc_1^2$. Le Cam's method (Tsybakov, 2009) gives $\inf_{\widetilde{\vartheta}} \sup_{\vartheta\in\{\vartheta_-,\vartheta_+\}} \mathbb{E}_\vartheta|\widetilde{\vartheta}-\vartheta| \ge cB_{\mathcal{H}}K^{-1/2}$. The identity $\widetilde{q}-q^{\circ}_{\alpha,\mathcal{F},q} = (\widetilde{q}-q_0)-\vartheta$ transfers this bound to threshold estimation, first on the pair and then on $\mathcal{P}_{\mathrm{bin}}$. This completes the proof. □

S12.2 Proof of Theorem S19

The augmented criterion controls coverage through calibration error and the additive nuisance remainder.

Proof. Fix the nuisance fits and calibration data on (S.65). For each fold k and threshold q, let

$$M_k(q) = \mathbf{1}\{R_n^{(-k)}(X,Y) > q\}$$

for a fresh source observation, and write $\bar{M}(q) = K_{\mathrm{CV}}^{-1}\sum_{k=1}^{K_{\mathrm{CV}}} M_k(q)$. By the definition of $\widehat{q}_{\mathrm{DR}}$ and (S.65), on the event in the theorem,

$$K_{\mathrm{CV}}^{-1}\sum_{k=1}^{K_{\mathrm{CV}}} \mathbb{E}_{P_{\mathrm{src}}}\big[(f+\widehat{r}^{(-k)}-\overline{\widehat{r}})\{M_k(\widehat{q}_{\mathrm{DR}})-\alpha\}\big] \le \tau_n + \psi_n^{\mathrm{DR}}(\delta)$$

for every $f\in\mathcal{F}$. Let

$$\bar{r}_K(x) = K_{\mathrm{CV}}^{-1}\sum_{k=1}^{K_{\mathrm{CV}}} \widehat{r}^{(-k)}(x).$$

Because $K^\star$ is uniform and independent, we have

$$\mathbb{E}_{K^\star}\Big[P_{\mathrm{tar}}\{Y\notin\widehat{\mathcal{C}}_{\mathrm{DR}}(X)\mid K^\star,\mathcal{D}_{\mathrm{cal}}\}\Big]-\alpha = \mathbb{E}_{P_{\mathrm{src}}}\{r(X)(\bar{M}(\widehat{q}_{\mathrm{DR}})-\alpha)\}.$$

Fix $\eta>0$ and choose $f_\eta\in\mathcal{F}$ such that

$$\sup_{q\in\mathbb{R}}\mathbb{E}_{P_{\mathrm{src}}}\left|r(X)\bar{M}(q)-K_{\mathrm{CV}}^{-1}\sum_{k=1}^{K_{\mathrm{CV}}}\{f_\eta(X)+\widehat{r}^{(-k)}(X)-\overline{\widehat{r}}\}M_k(q)\right|+\alpha\left|\overline{\widehat{r}}-\bar{r}_n^P\right| \le \rho_n^{\mathrm{aug}}+\eta.$$

Applying the preceding population inequality with $f=f_\eta$, and writing $\bar{M}=\bar{M}(\widehat{q}_{\mathrm{DR}})$ and $M_k = M_k(\widehat{q}_{\mathrm{DR}})$, we obtain that

$$\begin{aligned}
&\mathbb{E}_{P_{\mathrm{src}}}\{r(\bar{M}-\alpha)\}-K_{\mathrm{CV}}^{-1}\sum_{k=1}^{K_{\mathrm{CV}}}\mathbb{E}_{P_{\mathrm{src}}}\big[(f_\eta+\widehat{r}^{(-k)}-\overline{\widehat{r}})\{M_k-\alpha\}\big]\\
&\quad=\mathbb{E}_{P_{\mathrm{src}}}\left[r\bar{M}-K_{\mathrm{CV}}^{-1}\sum_{k=1}^{K_{\mathrm{CV}}}\{f_\eta+\widehat{r}^{(-k)}-\overline{\widehat{r}}\}M_k\right]+\alpha(\bar{r}_n^P-\overline{\widehat{r}})\\
&\quad\le\mathbb{E}_{P_{\mathrm{src}}}\left|r\bar{M}-K_{\mathrm{CV}}^{-1}\sum_{k=1}^{K_{\mathrm{CV}}}\{f_\eta+\widehat{r}^{(-k)}-\overline{\widehat{r}}\}M_k\right|+\alpha\left|\overline{\widehat{r}}-\bar{r}_n^P\right|\\
&\quad\le\rho_n^{\mathrm{aug}}+\eta,
\end{aligned}$$

by $\mathbb{E}_{P_{\mathrm{src}}}r(X)=\mathbb{E}_{P_{\mathrm{src}}}f_\eta(X)=1$, the definition of $\bar{r}_n^P$, and (S.64). The event was fixed before choosing r. Each normalized nonnegative ratio has its own f_η and remainder; an infinite remainder gives an immediate bound. Thus the argument holds for all such ratios on the same event. Applying (S.65) at $\widehat{q}_{\mathrm{DR}}$ gives

$$\mathbb{E}_{K^\star}\Big[P_{\mathrm{tar}}\{Y\notin\widehat{\mathcal{C}}_{\mathrm{DR}}(X)\mid K^\star,\mathcal{D}_{\mathrm{cal}}\}\Big]-\alpha\le\tau_n+\psi_n^{\mathrm{DR}}(\delta)+\rho_n^{\mathrm{aug}}+\eta.$$

Letting $\eta\downarrow 0$ yields (S.66). Propositions S28 and S29 give the stated probability bounds for (S.65), with conditioning on all nuisance fits only in the fully sample-split case. This completes the proof. □

S12.3 Proof of Theorem S20

Finitely many returned images create a moment-estimation error. Its effect on paired radii is distinct from its effect on a population threshold.

Proof. Put $\delta_q = \delta/3$. Moment matching gives $\widetilde{\sigma}_h(x) = \sigma_h^\circ(x)$. The bounded assertion of Lemma S6, at level δ_q, proves (S.67) without a variance lower bound. Proposition S10 applied to $(R_{n,K}, R_q^\circ)$ gives (S.68). For the population-quantile conclusion, use the additional local-quantile and smallness conditions. Choose C_{mc} at least the constant in (S.68), depending only on $B_{\mathcal{H}}, \lambda$, to ensure $e_{n,q} + 3\psi_{n,q}(\delta_q, \mathcal{F}) \le c_0 r_{0,q}$. On this event and the empirical-process events for both scores, each of probability at least $1 - \delta_q$, Theorem S8's proof gives, with C_2 depending only on $B_{\mathcal{H}}, \lambda, c_{-,q}$,

$$\left|\widehat{q}_K - q^\circ_{\alpha,\mathcal{F},q}\right| \le C_2\left\{\psi_{n,q}(\delta_q,\mathcal{F}) + e_{n,q}\right\}, \qquad \left|\widehat{q}^\circ - q^\circ_{\alpha,\mathcal{F},q}\right| \le C_2\psi_{n,q}(\delta_q,\mathcal{F}),$$

which gives

$$|\widehat{q}_K - \widehat{q}^\circ| \le C_2\left\{\psi_{n,q}(\delta_q,\mathcal{F}) + e_{n,q}\right\}.$$

Enlarging C_2 if necessary and using (S.68) proves (S.69). On the event in (S.68), we have

$$\mathcal{C}^{R_{n,K}}_{\widehat{q}_K}(x) \subseteq \mathcal{C}^{R_q^\circ}_{\widehat{q}_K + e_{n,q}}(x), \qquad \mathcal{C}^{R_q^\circ}_{\widehat{q}^\circ}(x) \subseteq \mathcal{C}^{R_{n,K}}_{\widehat{q}^\circ + e_{n,q}}(x).$$

These inclusions and the condition for the score pair give

$$\left|\mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\widehat{\mathcal{C}}_K) - \mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\widehat{\mathcal{C}}^\circ)\right| \le A_{\mathcal{F}}\left\{|\widehat{q}_K - \widehat{q}^\circ| + e_{n,q}\right\},$$

which, after enlarging C_2 if necessary and using (S.68), yields (S.70). Lemma S7 gives (S.71) on the specified two-point experiment.

Under only the paired radial conditions, Proposition S2 on (S.68) gives $2A_{\mathcal{F}}C_1\sqrt{L_{q,K,\delta}/K}$. Adding the nonnegative calibration term gives (S.70) without a population-quantile condition. This completes the proof. □

S12.4 Proof of Proposition S34

An independent selection sample estimates each candidate radius without reusing responses from its construction.

Proof. Condition on the candidate sets, fixing $\rho_1, \dots, \rho_J$. Put $P_X = P_{\mathrm{src},X}$ and $P^X_{n_{\mathrm{sel}}} = n_{\mathrm{sel}}^{-1}\sum_i \delta_{X_i^{\mathrm{sel}}}$. For measurable g, write $P_X g = \mathbb{E}_{P_{\mathrm{src},X}} g(X)$ and $P^X_{n_{\mathrm{sel}}} g = n_{\mathrm{sel}}^{-1}\sum_i g(X_i^{\mathrm{sel}})$. Define

$$\mathcal{G}_j = \{x \mapsto f(x)\rho_j(x) : f \in \mathcal{F}\}.$$

The class $\mathcal{G}_j$ is bounded by $B_F B_\rho$. For fixed nonnegative measurable ρ_j, the subgraph of $g_f = f\rho_j$ is

$$\{(x,t) : g_f(x) > t\} = \{(x,t) : \rho_j(x) > 0,\ f(x) > t/\rho_j(x)\} \cup \{(x,t) : \rho_j(x) = 0,\ t < 0\}.$$

This is a measurable pullback of the subgraph class of $\mathcal{F}$ and a fixed half-space, giving VC-subgraph dimension at most $C(v_{\mathcal{F}} + 1)$. Multiplying a countable dense subclass of $\mathcal{F}$ by ρ_j gives pointwise measurability. Lemma S11, conditional on the candidates, gives

$$\sup_{g\in\mathcal{G}_j}\left|\frac{1}{n_{\mathrm{sel}}}\sum_{i=1}^{n_{\mathrm{sel}}} g(X_i^{\mathrm{sel}}) - \mathbb{E}_{P_{\mathrm{src},X}} g(X)\right| \le C B_F B_\rho\sqrt{\frac{1 + (v_{\mathcal{F}}+1)\log(e n_{\mathrm{sel}}) + \tau}{n_{\mathrm{sel}}}}$$

with conditional probability at least $1 - e^{-\tau}$.

Because $S_j = \sup_{g\in\mathcal{G}_j} P_X g$ and $\widehat{S}_j = \sup_{g\in\mathcal{G}_j} P^X_{n_{\mathrm{sel}}} g$, we have

$$\left|\widehat{S}_j - S_j\right| \le \sup_{g\in\mathcal{G}_j}\left|P^X_{n_{\mathrm{sel}}} g - P_X g\right|.$$

Choosing $\tau = \log(J/\delta)$ and taking a union bound over $j = 1, \dots, J$ yields (S.74). The final assertion follows from the pointwise bound $\rho_j(x) \le A_j(x)\widehat{q}_j \le B_A \max_{1\le j\le J}\widehat{q}_j$. This completes the proof. □

S12.5 Proof of Theorem S21

Simultaneous coverage and radius-estimation bounds remain valid for the selected candidate.

Lemma S8. *Suppose that the candidate sets $\widehat{\mathcal{C}}_1, \ldots, \widehat{\mathcal{C}}_J$ satisfy $\Delta_{\mathcal{F}}^{-}(\widehat{\mathcal{C}}_j) \le \varepsilon_n$ for every $j = 1, \ldots, J$. Then, any selected candidate $\widehat{\mathcal{C}}_{\widehat{j}}$ whose index $\widehat{j}$ is measurable with respect to the sigma-field generated by the candidate construction and the selection sample satisfies $\Delta_{\mathcal{F}}^{-}(\widehat{\mathcal{C}}_{\widehat{j}}) \le \varepsilon_n$.*

Lemma S8 uses simultaneous population coverage of the candidate sets. Its premise is not high empirical coverage in a reused evaluation sample.

Proof. Conditional on construction and selection data, $\widehat{j}$ is fixed and $P_f\{Y \in \widehat{\mathcal{C}}_{\widehat{j}}(X)\} = \sum_{j=1}^{J} \mathbf{1}\{\widehat{j} = j\} P_f\{Y \in \widehat{\mathcal{C}}_j(X)\}$. Thus $\Delta_{\mathcal{F}}^{-}(\widehat{\mathcal{C}}_j) \le \varepsilon_n$ for all j implies $\Delta_{\mathcal{F}}^{-}(\widehat{\mathcal{C}}_{\widehat{j}}) \le \varepsilon_n$. This completes the proof. □

Lemma S9. *Suppose that the event in (S.75) holds. Then, the selected index $\widehat{j} \in \operatorname{argmin}_{1 \le j \le J}\{\widehat{S}_j + \mathrm{pen}_j\}$ satisfies $\mathcal{S}_{\mathcal{F},\mathcal{H}_{\widehat{j}}}(\widehat{\mathcal{C}}_{\widehat{j}}) \le \inf_{1 \le j \le J}\{\mathcal{S}_{\mathcal{F},\mathcal{H}_j}(\widehat{\mathcal{C}}_j) + 2\mathrm{pen}_j\}$.*

Lemma S9 shows that the selection penalty accounts for error in estimated set size when comparing candidates. The conclusion requires simultaneous error bounds for all candidates.

Proof. On (S.75), $\left|\widehat{S}_j - S_j\right| \le \mathrm{pen}_j$ for every j. The definition of $\widehat{j}$ gives

$$\widehat{S}_{\widehat{j}} + \mathrm{pen}_{\widehat{j}} \le \widehat{S}_j + \mathrm{pen}_j \qquad \text{for all } j = 1, \ldots, J.$$

This implies that, for every $j = 1, \ldots, J$,

$$S_{\widehat{j}} \le \widehat{S}_{\widehat{j}} + \mathrm{pen}_{\widehat{j}} \le \widehat{S}_j + \mathrm{pen}_j \le S_j + 2\mathrm{pen}_j.$$

Taking the infimum over $j = 1, \ldots, J$ yields the claim. This completes the proof. □

Proof of Theorem S21. On the intersection of simultaneous candidate validity and (S.75), Lemmas S8 and S9 give (S.76) and (S.77). A union bound gives the stated probability. This completes the proof. □

S12.6 Proof of Theorem S22

The two calibrated scores have comparable radii when score approximation and the stated radial continuity hold.

Proof. Put $e_{n,\delta} = e_n(\delta/3)$ and $\psi_{n,\delta} = \psi_n(\delta/3, \mathcal{F})$. Let $\mathcal{E}_{\mathrm{sc}}$ be the event in (S.33), and let $\mathcal{E}_{\mathrm{ep}}^{n}, \mathcal{E}_{\mathrm{ep}}^{\circ}$ be the events in (S.34) for R_n, R° with bound $\psi_{n,\delta}$. Their intersection has probability at least $1 - \delta$. Theorem S9 on each empirical-process event gives $\Delta_{\mathcal{F}}^{-}(\widehat{\mathcal{C}}^{\circ}), \Delta_{\mathcal{F}}^{-}(\widehat{\mathcal{C}}) \le C\psi_{n,\delta}$, proving (S.78).

On $\mathcal{E}_{\mathrm{sc}}$, Proposition S2 gives $|\widehat{q} - \widehat{q}^{\circ}| \le e_{n,\delta}$ without either empirical-process event. Also,

$$\mathcal{C}_{\widehat{q}}^{R_n}(x) \subseteq \mathcal{C}_{\widehat{q}+e_{n,\delta}}^{R^{\circ}}(x), \qquad \mathcal{C}_{\widehat{q}^{\circ}}^{R^{\circ}}(x) \subseteq \mathcal{C}_{\widehat{q}^{\circ}+e_{n,\delta}}^{R_n}(x).$$

The assumption in Theorem S22 implies that

$$\left|\mathcal{S}_{\mathcal{F},\mathcal{H}}(\widehat{\mathcal{C}}) - \mathcal{S}_{\mathcal{F},\mathcal{H}}(\widehat{\mathcal{C}}^{\circ})\right| \le A_{\mathcal{F}}\left\{|\widehat{q} - \widehat{q}^{\circ}| + e_{n,\delta}\right\},$$

which proves (S.79). This completes the proof. □

Proof of Lemma S2. Fix $S \in \mathcal{R}$, $x \in \mathcal{X}$, and $t \ge 0$. The identities $\mathcal{C}_t^S(x) = \mathbb{B}_{\mathcal{H}}\{y_S(x), a_S(x)t\}$ and $\mathrm{rad}_{\mathcal{H}}\{\mathbb{B}_{\mathcal{H}}(y_0, r)\} = r$ give

$$\mathrm{rad}_{\mathcal{H}}\{\mathcal{C}_t^S(x)\} = \mathrm{rad}_{\mathcal{H}}[\mathbb{B}_{\mathcal{H}}\{y_S(x), a_S(x)t\}] = a_S(x)t \le A(x)t.$$

This equality implies that, for every $t_1, t_2 \geq 0$,

$$\left|\mathrm{rad}_{\mathcal{H}}\{\mathcal{C}^S_{t_2}(x)\} - \mathrm{rad}_{\mathcal{H}}\{\mathcal{C}^S_{t_1}(x)\}\right| = a_S(x)\,|t_2 - t_1| \leq A(x)\,|t_2 - t_1|\,.$$

These bounds hold for every $S \in \mathcal{R}$. Taking $\mathcal{R} = \{R_n\}$ gives Propositions S11 and S12; taking $\mathcal{R} = \{R_n, R^\circ\}$ gives the common bound in Theorem S22. This completes the proof. □

Proof of Proposition S18. Fix x and put $y_c = y_c(x)$, $m_j = m_j(x)$, $a_\vee = a_\vee(x)$, and $d_q(y, y') = \max_j |h_j(y) - h_j(y')|$. Because $T_q(y_c) = (m_1, \ldots, m_J)$, $y_c \in \mathcal{C}_t^{S_q}(x)$ for all $t \geq 0$. Using this center gives $\mathrm{rad}_{\mathcal{H}_q}\{\mathcal{C}_t^{S_q}(x)\} \leq \sup_{y \in \mathcal{C}_t^{S_q}(x)} d_q(y, y_c) \leq a_\vee t$.

By symmetry, take $0 \leq t_1 \leq t_2$. Set inclusion gives $\mathrm{rad}_{\mathcal{H}_q}\{\mathcal{C}_{t_1}^{S_q}(x)\} \leq \mathrm{rad}_{\mathcal{H}_q}\{\mathcal{C}_{t_2}^{S_q}(x)\}$. If $t_2 = 0$, both radii are equal. For $t_2 > 0$, put $s = t_1/t_2$ and choose, for $\varepsilon > 0$, a center $z_1 \in \mathcal{Y}$ such that

$$\sup_{v \in \mathcal{C}_{t_1}^{S_q}(x)} d_q(v, z_1) \leq \mathrm{rad}_{\mathcal{H}_q}\{\mathcal{C}_{t_1}^{S_q}(x)\} + \varepsilon.$$

For $y \in \mathcal{C}_{t_2}^{S_q}(x)$, Proposition S18's condition gives $y_s \in \mathcal{Y}$ satisfying

$$T_q(y_s) = T_q(y_c) + s\{T_q(y) - T_q(y_c)\}.$$

Then $|h_j(y_s) - m_j| = s|h_j(y) - m_j| \leq a_j(x)t_1$ for all j, hence $y_s \in \mathcal{C}_{t_1}^{S_q}(x)$ and

$$d_q(y, y_s) = (1 - s) \max_{1 \leq j \leq J} |h_j(y) - m_j| \leq a_\vee(t_2 - t_1).$$

The triangle inequality gives

$$d_q(y, z_1) \leq d_q(y, y_s) + d_q(y_s, z_1) \leq a_\vee(t_2 - t_1) + \mathrm{rad}_{\mathcal{H}_q}\{\mathcal{C}_{t_1}^{S_q}(x)\} + \varepsilon.$$

Take the supremum over $y \in \mathcal{C}_{t_2}^{S_q}(x)$ and let $\varepsilon \downarrow 0$ to obtain

$$\mathrm{rad}_{\mathcal{H}_q}\{\mathcal{C}_{t_2}^{S_q}(x)\} - \mathrm{rad}_{\mathcal{H}_q}\{\mathcal{C}_{t_1}^{S_q}(x)\} \leq a_\vee(t_2 - t_1).$$

Radius monotonicity gives the absolute difference bound. Taking $\mathbb{E}_f$ and the supremum over $f \in \mathcal{F}$ proves the final assertion. This completes the proof. □

Proof of Proposition S25. Fix $f \in \mathcal{F}$ satisfying $q^\circ_{f,\alpha,T} \geq q^\circ_{\alpha,\mathcal{F},T} - r_{0,T}$. If $0 \leq t \leq r_{0,T}$, then the absolute continuity and the assumed lower bound on the density yield

$$F_{f,T}(q^\circ_{f,\alpha,T}) - F_{f,T}(q^\circ_{f,\alpha,T} - t) \geq c_{-,T}t.$$

This and $F_{f,T}(q^\circ_{f,\alpha,T}) = 1 - \alpha$ yield the assertion. This completes the proof. □

Proof of Proposition S26. By symmetry, take $t \geq s$. Finiteness of $\mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\mathcal{C}_s^T)$ permits $f_\varepsilon \in \mathcal{F}$ with $\mathbb{E}_{P_{f_\varepsilon}} \mathrm{rad}_{\mathcal{H}_q}\{\mathcal{C}_s^T(X)\} \geq \mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\mathcal{C}_s^T) - \varepsilon$. Proposition S26's condition gives

$$\begin{aligned}\mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\mathcal{C}_t^T) &\geq \mathbb{E}_{P_{f_\varepsilon}} \mathrm{rad}_{\mathcal{H}_q}\{\mathcal{C}_t^T(X)\} \\ &\geq \mathbb{E}_{P_{f_\varepsilon}} \mathrm{rad}_{\mathcal{H}_q}\{\mathcal{C}_s^T(X)\} + (t - s)\mathbb{E}_{P_{f_\varepsilon}} a(X) \geq \mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\mathcal{C}_s^T) - \varepsilon + a_0(t - s).\end{aligned}$$

Letting $\varepsilon \downarrow 0$ yields $\mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\mathcal{C}_t^T) - \mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\mathcal{C}_s^T) \geq a_0(t - s)$. The case $s > t$ follows by interchanging s and t. This completes the proof. □

S12.7 Proof of Proposition S21

The dimension of the linear shift class determines its calibration error through the covering-number bound.

Proof. Condition on R_n. Under the stated conditional-i.i.d. assumption, the calibration empirical-process class in (S.34) is

$$\mathcal{G}_{R_n} = \{(x,y) \mapsto f(x)\{\mathbf{1}\{R_n(x,y) > q\} - \alpha\} : f \in \mathcal{F}_d,\ q \in \mathbb{R}\}.$$

Let $\Theta_d^\star$ denote the feasible parameter set

$$\Theta_d^\star = \{\theta \in \Theta_{\text{aff}} : \theta^\top B(x) \geq 0 \text{ for all } x,\ \mathbb{E}_{P_{\text{src},X}}\{\theta^\top B(X)\} = 1\}.$$

Choose a countable Euclidean-dense subset $\Theta_{d,0}^\star$ of $\Theta_d^\star \subseteq \mathbb{R}^{d+1}$. If $\theta_m \to \theta$, then $\theta_m^\top B(x) \to \theta^\top B(x)$ for every x, giving pointwise measurability of $\mathcal{F}_d$. For fixed Q, identify functions equal Q-almost surely. The span of $\mathcal{F}_d$ has dimension $r \leq d+1$ in $L_2(Q)$, and its members have norm at most B_d. For $r = 0$, one ball suffices. For $r > 0$, balls of radius $\eta B_d/2$ centered at an ηB_d-separated subset are disjoint and lie in the ball of radius $B_d + \eta B_d/2$. Volume comparison bounds their number by $(1 + 2/\eta)^r \leq (3/\eta)^{d+1}$. A maximal separated subset is an ηB_d-net, uniformly in Q. Lemma S12 gives product exponent $d+3$. Lemma S11, conditional on R_n, gives

$$\sup_{f \in \mathcal{F}_d}\sup_{q \in \mathbb{R}} |(P_n - P)\,[f(X)\{\mathbf{1}\{R_n(X,Y) > q\} - \alpha\}]| \leq C B_d \sqrt{\frac{1 + (d+3)\log(A_d e n_{\text{cal}}) + \log(1/\delta)}{n_{\text{cal}}}}$$

with conditional probability at least $1 - \delta$. This completes the proof. □

S12.8 Proof of Proposition S22

A bounded VC shift class has controlled complexity after multiplication by a thresholded score.

Proof. Condition on R_n. A uniformly bounded VC-subgraph class with dimension v and uniform bound B_F satisfies the standard uniform entropy bound

$$\sup_Q N\{\eta B_F, \mathcal{F}, L_2(Q)\} \leq (A_v/\eta)^{C_v v}, \qquad 0 < \eta \leq 1,$$

where A_v, C_v depend only on the VC-subgraph convention (van der Vaart & Wellner, 1996). Apply Lemma S12 to

$$\mathcal{G}_{R_n} = \{(x,y) \mapsto f(x)\{\mathbf{1}\{R_n(x,y) > q\} - \alpha\} : f \in \mathcal{F},\ q \in \mathbb{R}\}.$$

The additional threshold factor gives product exponent at most $C(v+2)$. Lemma S11, conditional on R_n, gives

$$\sup_{f \in \mathcal{F}}\sup_{q \in \mathbb{R}} |(P_n - P)\,[f(X)\{\mathbf{1}\{R_n(X,Y) > q\} - \alpha\}]| \leq C B_F \sqrt{\frac{1 + (v+2)\log(e n_{\text{cal}}) + \log(1/\delta)}{n_{\text{cal}}}}$$

with conditional probability at least $1 - \delta$, after enlarging C to absorb universal entropy constants. This completes the proof. □

S12.9 Proof of Proposition S23

Ordering the scores reduces the threshold supremum to partial sums of kernel features.

Proof. Condition on R_n, leaving i.i.d. source calibration observations. Let ε_i be independent Rademacher variables, independent of calibration. Put $W_i = (X_i, Y_i)$, $K_x = k(x, \cdot)$, and $a_q(W_i) = \mathbf{1}\{R_n(W_i) > q\} - \alpha$. For the unit RKHS ball,

$$\sup_{\|f\|_{\mathcal{H}_k} \leq 1} \left| \frac{1}{n_{\text{cal}}} \sum_{i \in \mathcal{I}_{\text{cal}}} \varepsilon_i f(X_i) a_q(W_i) \right| = \frac{1}{n_{\text{cal}}} \left\| \sum_{i \in \mathcal{I}_{\text{cal}}} \varepsilon_i a_q(W_i) K_{X_i} \right\|_{\mathcal{H}_k}.$$

Order the scores as $R_n(W_{\pi_1}) \geq \cdots \geq R_n(W_{\pi_{n_{\text{cal}}}})$. Each set $\{i : R_n(W_i) > q\}$ is a prefix, hence

$$\sup_q \left\| \sum_i \varepsilon_i a_q(W_i) K_{X_i} \right\|_{\mathcal{H}_k} \leq \alpha \left\| \sum_i \varepsilon_i K_{X_i} \right\|_{\mathcal{H}_k} + \sup_{0 \leq t \leq n_{\text{cal}}} \left\| \sum_{s=1}^{t} \varepsilon_{\pi_s} K_{X_{\pi_s}} \right\|_{\mathcal{H}_k},$$

where π is the score ordering. The bound $\|K_x\|_{\mathcal{H}_k} \le \kappa$ gives conditional expectation at most $\kappa\sqrt{n_{\mathrm{cal}}}$ for the first term. Doob's Hilbert-space maximal inequality gives $2\kappa\sqrt{n_{\mathrm{cal}}}$ for the second. Symmetrization yields

$$\mathbb{E} \sup_{f\in\mathcal{F},q\in\mathbb{R}} \left|(P_n - P)\{f(X)a_q(W)\}\right| \le C\kappa n_{\mathrm{cal}}^{-1/2}.$$

The class is bounded by κ, and replacing one observation changes the supremum by at most $2\kappa/n_{\mathrm{cal}}$. McDiarmid's inequality gives

$$\sup_{f\in\mathcal{F},q\in\mathbb{R}} \left|(P_n - P)\{f(X)a_q(W)\}\right| \le C\kappa\sqrt{\frac{1+\log(1/\delta)}{n_{\mathrm{cal}}}}$$

with probability at least $1-\delta$, after enlarging C. This completes the proof. □

S12.10 Proofs of Theorems S13 and S14

Uniform estimation of predictor moments gives the score bound, while the empirical-process bound controls calibration.

Lemma S10. *Let $\mathcal{A}$ be pointwise measurable, let $|a| \le B$ for every $a \in \mathcal{A}$, and suppose $N_\infty(\eta B, \mathcal{A}) \le (A_0/\eta)^v$ for every $0 < \eta \le 1$. If $W_1, \dots, W_K$ are i.i.d. with distribution P and $P_K = K^{-1}\sum_{k=1}^K \delta_{W_k}$, then, for every $0 < \delta < 1$, a universal constant $C > 0$ satisfies*

$$\mathbb{P}\left\{\sup_{a\in\mathcal{A}} |(P_K - P)a| > CB\sqrt{\frac{1 + v\log(A_0K) + \log(1/\delta)}{K}}\right\} \le \delta.$$

Lemma S10 controls predictor-moment estimation uniformly over the specified query and functional indices through a finite approximation of the class.

Proof of Lemma S10. Replace $\mathcal{A}$ by a countable pointwise dense subclass. For $0 < \eta \le 1$, choose a uniform ηB-net $\mathcal{A}_\eta$ and $\Pi_\eta(a) \in \mathcal{A}_\eta$ with $\|a - \Pi_\eta(a)\|_\infty \le \eta B$. Then $|(P_K - P)\{a - \Pi_\eta(a)\}| \le 2\eta B$. For $g \in \mathcal{A}_\eta$, Hoeffding gives $\mathbb{P}\{|(P_K - P)g| > t\} \le 2\exp\{-Kt^2/(2B^2)\}$. A union bound yields

$$\mathbb{P}\left\{\max_{g\in\mathcal{A}_\eta} |(P_K - P)g| > B\sqrt{\frac{2\log(2|\mathcal{A}_\eta|/\delta)}{K}}\right\} \le \delta.$$

Use $|\mathcal{A}_\eta| \le (A_0/\eta)^v$ and $\eta = K^{-1}$. The additive 1 absorbs the net constant and discretization error for all $\delta \in (0,1)$. This completes the proof. □

Lemma S11. *Let $\mathcal{G}$ be pointwise measurable, let $|g| \le B$ for every $g \in \mathcal{G}$, and suppose $\sup_Q N\{\eta B, \mathcal{G}, L_2(Q)\} \le (A_0/\eta)^v$ for every $0 < \eta \le 1$. If $W_1, \dots, W_n$ are i.i.d. with distribution P and $P_n = n^{-1}\sum_{i=1}^n \delta_{W_i}$, then, for every $0 < \delta < 1$, a universal constant $C > 0$ satisfies*

$$\mathbb{P}\left\{\sup_{g\in\mathcal{G}} |(P_n - P)g| > CB\sqrt{\frac{1 + v\log(A_0n) + \log(1/\delta)}{n}}\right\} \le \delta.$$

Lemma S11 bounds the difference between sample averages and population expectations uniformly over the function class. Boundedness and the stated approximation count determine the bound.

Proof of Lemma S11. Let $Z = \sup_{g\in\mathcal{G}} |(P_n - P)g|$. A ghost-sample argument and Rademacher symmetrization give

$$\mathbb{E}Z \le 2\mathbb{E}\left[\sup_{g\in\mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(W_i)\right|\right],$$

where the ε_i are independent Rademacher variables. Conditional on the observations, take an ηB-net $\mathcal{G}_\eta$ in $L_2(P_n)$ with $|\mathcal{G}_\eta| \le (A_0/\eta)^v$ and representatives $\Pi_\eta(g)$. Then

$$\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\{g(W_i) - \Pi_\eta(g)(W_i)\}\right| \le \{P_n(g - \Pi_\eta(g))^2\}^{1/2} \le \eta B.$$

For the finite net,

$$\mathbb{E}_\varepsilon \max_{g\in\mathcal{G}_\eta}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(W_i)\right| \le B\sqrt{\frac{2\log(2|\mathcal{G}_\eta|)}{n}}.$$

Taking $\eta = n^{-1}$ gives $\mathbb{E}Z \le CB\sqrt{\{1+v\log(A_0 n)\}/n}$. Changing one observation alters Z by at most $2B/n$, which implies, by McDiarmid's inequality, that

$$Z \le \mathbb{E}Z + B\sqrt{\frac{2\log(1/\delta)}{n}}$$

with probability at least $1-\delta$. The expectation bound completes the proof. □

The following lemma controls products of a bounded multiplier and a threshold indicator. In this lemma, $\mathcal{F}$ denotes a pointwise measurable class of real-valued functions on $\mathcal{X}$.

Lemma S12. *Suppose $|f| \le B_F$ for every $f \in \mathcal{F}$ and the covering-number bound in Assumption S3 holds. Then, for every measurable $R : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$, the class*

$$\mathcal{G}_R = \{(x,y) \mapsto f(x)\{\mathbf{1}\{R(x,y) > q\} - \alpha\} : f \in \mathcal{F},\ q \in \mathbb{R}\}$$

is pointwise measurable, is bounded by B_F, and satisfies

$$\sup_Q N\{\eta B_F, \mathcal{G}_R, L_2(Q)\} \le C(2A_F/\eta)^{v_F}\eta^{-2}$$

for every $0 < \eta \le 1$ and a universal constant $C > 0$.

Lemma S12 shows that varying the scalar threshold adds a controlled entropy term to the multiplier class. The complexity does not increase with the dimension of the response once its score is fixed. The lemma allows signed, unnormalized multipliers, as used in Propositions S28 and S29. The threshold comparison in Proposition S2 still requires nonnegative weights.

Proof of Lemma S12. Let $\mathcal{F}_0 \subset \mathcal{F}$ be a countable pointwise dense subclass. The class

$$\mathcal{G}_{R,0} = \{(x,y) \mapsto f(x)\{\mathbf{1}\{R(x,y) > q\} - \alpha\} : f \in \mathcal{F}_0,\ q \in \mathbb{Q}\}$$

is countable. For $f \in \mathcal{F}$ and $q \in \mathbb{R}$, choose $f_m \in \mathcal{F}_0$ converging pointwise to f and rational $q_m \downarrow q$ with $q_m > q$. Then $\mathbf{1}\{R(x,y) > q_m\} \to \mathbf{1}\{R(x,y) > q\}$ at every (x,y). The corresponding products converge pointwise, proving pointwise measurability of $\mathcal{G}_R$.

Fix Q on $\mathcal{X}\times\mathcal{Y}$. The covering-number bound gives an $\eta B_F/2$-net $\mathcal{F}_\eta$ in $L_2(Q_X)$ of size at most $(2A_F/\eta)^{v_F}$. Nested thresholds have squared $L_2(Q)$ distance $|Q(R > q) - Q(R > t)|$. A maximal $\eta^2/4$-separated subset of $\{Q(R > q) : q \in \mathbb{R}\} \subseteq [0,1]$ has at most $1 + 4/\eta^2$ points, each represented by an actual threshold. Maximality gives, for every q, a representative t_j with $Q\{\mathbf{1}\{R > q\} \ne \mathbf{1}\{R > t_j\}\} \le \eta^2/4$, including distributions with atoms. Hence

$$\begin{aligned}&\|f\{\mathbf{1}\{R > q\} - \alpha\} - f'\{\mathbf{1}\{R > t_j\} - \alpha\}\|_{L_2(Q)}\\ &\quad\le \|f - f'\|_{L_2(Q_X)} + B_F Q\{\mathbf{1}\{R > q\} \ne \mathbf{1}\{R > t_j\}\}^{1/2} \le \eta B_F.\end{aligned}$$

Taking the product of the two nets yields the claimed covering number. This completes the proof. □

Proof of Theorem S13. The countable $\|\cdot\|_\infty$-dense subclass in Assumption S2, the inequalities

$$\sup_x |\mu_h^{\mathrm{T}}(x) - \mu_{h'}^{\mathrm{T}}(x)| \le \|h - h'\|_\infty, \qquad \sup_x |m_{2,h}^{\mathrm{T}}(x) - m_{2,h'}^{\mathrm{T}}(x)| \le 2B_{\mathcal{H}}\|h - h'\|_\infty,$$

and their plug-in counterparts make R_n, R° countable suprema of measurable functions. Apply Lemma S10 to $\mathcal{A}_{\mu,n}, \mathcal{A}_{2,n}$, each at level $\delta/2$, to obtain, with probability at least $1-\delta$,

$$\sup_{x,h} |\widehat{\mu}_h(x) - \mu_h^{\mathrm{T}}(x)| \le \eta_{\mu,n}(\delta), \qquad \sup_{x,h} |\widehat{m}_{2,h}(x) - m_{2,h}^{\mathrm{T}}(x)| \le \eta_{2,n}(\delta),$$

where $\widehat{m}_{2,h}(x) = K^{-1}\sum_{k=1}^{K} h\{\widetilde{Z}_k(x)\}^2$. Combining these bounds with (S.52) gives

$$\sup_{x,h} |\widehat{\mu}_h(x) - \mu_h^{\circ}(x)| \le b_{\mu,n} + \eta_{\mu,n}(\delta), \qquad \sup_{x,h} |\widehat{m}_{2,h}(x) - m_{2,h}^{\circ}(x)| \le b_{2,n} + \eta_{2,n}(\delta).$$

Because $|\mu_h^\circ(x)|, |\widehat{\mu}_h(x)| \le B_{\mathcal{H}}$, putting $\varepsilon_{\mu,n} = b_{\mu,n} + \eta_{\mu,n}(\delta)$ gives

$$\sup_{x,h} \left|\widehat{\mu}_h(x)^2 - \{\mu_h^\circ(x)\}^2\right| \le 2B_{\mathcal{H}}\varepsilon_{\mu,n}.$$

Combining this with the second-moment bound gives

$$\sup_{x,h} |\widehat{\sigma}_h^2(x) - \{\sigma_h^\circ(x)\}^2| \le b_{2,n} + \eta_{2,n}(\delta) + 2B_{\mathcal{H}}\varepsilon_{\mu,n}.$$

The inequality $|\sqrt{u} - \sqrt{v}| \le |u - v|^{1/2}$ for $u, v \ge 0$ then gives $\sup_{x,h} |\widehat{\sigma}_h(x) - \sigma_h^\circ(x)| \le \eta_{\sigma,n}(\delta)$. Proposition S10 then yields (S.54). This completes the proof. □

Proof of Theorem S14. Condition on R, leaving i.i.d. source calibration observations. Lemma S12 gives bound B_F and entropy exponent $v_F + 2$ for $\mathcal{G}_R$. Lemma S11 yields

$$\mathbb{P}\left\{\sup_{g\in\mathcal{G}_R} |(P_n - P)g| > \psi_n(\delta, \mathcal{F}) \mid R\right\} \le \delta$$

after enlarging the constant in (S.53). Integrating over score-construction randomness proves (S.56). Substitute $R = R_n$ from Theorem S13 for the final assertion. This completes the proof. □

S12.11 Proof of Theorem S6

Images with the same accepted summary value all belong to the prediction set. Their spatial differences give a deterministic radius bound at fixed $\widehat{S}$, which we then minimize over $\mathcal{S}_M$.

Lemma S13. *Let $S \in \mathcal{S}_M$ be fixed, let $\mathcal{C} = \{y \in \mathcal{Y} : S(y) \in A\}$ for a measurable $A \subseteq \mathbb{R}^M$, and define $\Delta_{S,A} = \sup_{t\in A\cap S(\mathcal{Y})} \Delta_S(t)$, where the supremum over the empty set is zero. If $\mathcal{C} \neq \emptyset$, then* $\mathrm{rad}_{\mathcal{H}}(\mathcal{C}) \ge \Delta_{S,A}/2$.

Proof of Lemma S13. For $\Delta_{S,A} = 0$, the claim is immediate. Suppose $0 < \Delta_{S,A} < \infty$. Choose $t_\varepsilon \in A \cap S(\mathcal{Y})$ with $\Delta_S(t_\varepsilon) \ge \Delta_{S,A} - \varepsilon$. Its nonempty fiber lies in $\mathcal{C}$ and contains $y_\varepsilon, y'_\varepsilon$ satisfying

$$\sup_{h\in\mathcal{H}} |h(y_\varepsilon) - h(y'_\varepsilon)| \ge \Delta_{S,A} - 2\varepsilon.$$

Both points lie in $\mathcal{C}$. For every $\eta > 0$, choose $y_{0,\eta} \in \mathcal{Y}$ such that

$$\sup_{y\in\mathcal{C}} \sup_{h\in\mathcal{H}} |h(y) - h(y_{0,\eta})| \le \mathrm{rad}_{\mathcal{H}}(\mathcal{C}) + \eta.$$

The triangle inequality gives $\Delta_{S,A} - 2\varepsilon \le 2\{\mathrm{rad}_{\mathcal{H}}(\mathcal{C}) + \eta\}$. Letting $\varepsilon, \eta \downarrow 0$ proves the finite case. For $\Delta_{S,A} = \infty$ and any finite M_0, choose $t_{M_0} \in A \cap S(\mathcal{Y})$ with $\Delta_S(t_{M_0}) \ge M_0$ and $y_{M_0,\varepsilon}, y'_{M_0,\varepsilon} \in S^{-1}(t_{M_0}) \subseteq \mathcal{C}$ such that

$$\sup_{h\in\mathcal{H}} |h(y_{M_0,\varepsilon}) - h(y'_{M_0,\varepsilon})| \ge M_0 - \varepsilon.$$

If the radius were finite, an η-optimal center would give $M_0 - \varepsilon \le 2\{\mathrm{rad}_{\mathcal{H}}(\mathcal{C}) + \eta\}$. Let $\varepsilon, \eta \downarrow 0$ and $M_0 \to \infty$ to obtain $\mathrm{rad}_{\mathcal{H}}(\mathcal{C}) = \infty$. This completes the proof. □

Proof of Theorem S6. Condition on $(\widehat{S}, \widehat{A})$. Definition S5 gives $\widehat{S} \in \mathcal{S}_M$ almost surely and membership determined by $\widehat{S}(Y)$. On $\widehat{\mathcal{C}}^{(M)}(X) \neq \emptyset$, apply Lemma S13 with $S = \widehat{S}(\omega)$ and $A = \widehat{A}(X)(\omega)$:

$$\mathrm{rad}_{\mathcal{H}}\{\widehat{\mathcal{C}}^{(M)}(X)(\omega)\} \ge \frac{1}{2}\Delta_{\widehat{S},\widehat{A}}(X) \ge \Delta^\star_{\mathcal{S}_M}/2,$$

where the second inequality uses $\widehat{S}(\omega) \in \mathcal{S}_M$. This pathwise bound proves (S.31). This completes the proof. □

S12.12 Proof of Theorem S15

A family of subgroup probabilities reduces calibration error to simultaneous estimation across cells.

Proof. Let X be uniform on $\{1, \dots, d\}$. Given $X = j$, take a constant image of intensity $0, 1/2, 1$ with probabilities $p_j/2, 1 - p_j, p_j/2$. Its mean-intensity functional has center $1/2$ and standard deviation $\sqrt{p_j}/2$. Put $q_0 = \{2\sqrt{1-\alpha/2} + 4\lambda\}^{-1}$. The score in (S.25) is zero at $1/2$ and $(\sqrt{p_j} + 2\lambda)^{-1} > q_0$ elsewhere. Thus $Z = \mathbf{1}\{R^\circ > q_0\} \sim \text{Bernoulli}(p_j)$. Given $Z = 1$, the sign is uninformative about p_j, giving the Bernoulli divergence below for the full data. Also $\mathbb{E}_{P_{\text{src},X}} f(X) = 1$ and $0 \le f \le 2$ for $f \in \mathcal{F}_{\text{sub}}$. For $u \in \mathbb{R}^d$ and even d,

$$\sup_{A:|A|=d/2} \Big| \frac{2}{d} \sum_{j \in A} u_j \Big| \ge \frac{1}{2d} \sum_{j=1}^{d} |u_j|. \tag{S.84}$$

We define $u_j^+ = \max\{u_j, 0\}$ and $u_j^- = \max\{-u_j, 0\}$. Put $P_+ = \sum_j u_j^+$ and $P_- = \sum_j u_j^-$. If $P_+ \ge P_-$ and at least $d/2$ coordinates are nonnegative, their largest $d/2$ sum to at least $P_+/2 \ge (P_+ + P_-)/4$. If fewer are nonnegative, include all of them and the negative coordinates of smallest magnitude. The added magnitude is at most $P_-/2$, giving a sum at least $P_+ - P_-/2 \ge (P_+ + P_-)/4$. Replace u by $-u$ when $P_- > P_+$. Hence

$$\sup_{A:|A|=d/2} \Big| \sum_{j \in A} u_j \Big| \ge \frac{1}{4} \sum_{j=1}^{d} |u_j|,$$

which implies (S.84).

For an Assouad cube, put $p_0 = 1/2$ and, for $\omega \in \{-1, +1\}^d$, set

$$p_j^\omega = p_0 + \varepsilon \omega_j, \qquad \varepsilon = a_\alpha \sqrt{d/n},$$

where $a_\alpha > 0$ is chosen such that $p_j^\omega \in [\alpha/2, 1 - \alpha/2]$ for all j and all $n \ge d$. If ω and ω' differ in exactly one coordinate, the one-observation Kullback–Leibler divergence is

$$D_{\text{KL}}(P_\omega \| P_{\omega'}) = \frac{1}{d} D_{\text{KL}}\{\text{Bernoulli}(p_0 + \varepsilon) \| \text{Bernoulli}(p_0 - \varepsilon)\} \le \frac{C_\alpha \varepsilon^2}{d}.$$

The n-observation divergence is at most $C_\alpha n \varepsilon^2 / d = C_\alpha a_\alpha^2$. For small $a_\alpha > 0$, neighboring experiments have Hellinger affinity bounded below by a positive absolute constant. Assouad's lemma (see, e.g., Tsybakov, 2009) gives

$$\inf_{\widehat{p}} \sup_{\omega} \mathbb{E}_\omega \sum_{j=1}^{d} |\widehat{p}_j - p_j^\omega| \ge c_\alpha d \varepsilon.$$

Combining this bound with (S.84) yields

$$\inf_{\widehat{p}} \sup_{\omega} \mathbb{E}_\omega \sup_{f \in \mathcal{F}_{\text{sub}}} \Big| \frac{1}{d} \sum_{j=1}^{d} f(j)\{\widehat{p}_j - p_j^\omega\} \Big| \ge c_\alpha \varepsilon = c_\alpha \sqrt{d/n}.$$

The cube lies in $[\alpha/2, 1 - \alpha/2]^d$, hence the bound implies (S.57). This completes the proof. □

S12.13 Proof of Proposition S24

The Gaussian margin converts threshold estimation error into conditional coverage error for the specified prediction sets.

Proof. Put $z_\alpha = \Phi^{-1}(1 - \alpha)$ and $c_\alpha = \phi(|z_\alpha| + 1)$, where Φ and ϕ denote the standard normal distribution and density functions. For every $u \in \mathbb{R}$, the mean value theorem on $\{|u| \le 1\}$ and the monotonicity of Φ on $\{|u| > 1\}$ give

$$|\Phi(z_\alpha + u) - \Phi(z_\alpha)| \ge c_\alpha \min(|u|, 1). \tag{S.85}$$

For a threshold vector $\widehat{q}$, the coverage in cell j equals $\mathbb{P}_\theta\{Y \le \widehat{q}_j \mid X = j\} = \Phi(\widehat{q}_j - \theta_j)$. Writing $u_j = \Phi(\widehat{q}_j - \theta_j) - (1-\alpha)$ and applying (S.84),

$$\sup_{f \in \mathcal{F}_{\mathrm{sub}}} \left| P_{f,\theta}\{Y \in \widehat{\mathcal{C}}(X)\} - (1-\alpha) \right| = \sup_{A : |A| = d/2} \left| \frac{2}{d} \sum_{j \in A} u_j \right| \ge \frac{1}{2d} \sum_{j=1}^{d} |u_j| \,.$$

Put $\widetilde{\theta}_j = \widehat{q}_j - z_\alpha$. Equation (S.85) gives $|u_j| \ge c_\alpha \min(\left|\widetilde{\theta}_j - \theta_j\right|, 1)$. For $\omega \in \{-1,+1\}^d$, set $\theta_j^\omega = \varepsilon \omega_j$, where $\varepsilon = a\sqrt{d/n} \le a \le 1/2$ and $a > 0$ is small. Then $\theta^\omega \in \Theta$. Neighbors have one-observation divergence $d^{-1} D_{\mathrm{KL}}\{N(\varepsilon,1) \,\|\, N(-\varepsilon,1)\} = 2\varepsilon^2/d$ and n-observation divergence at most $2a^2$. The inequalities $\min(x,1)+\min(y,1) \ge \min(x+y,1)$ for $x, y \ge 0$ and $|t-\varepsilon|+|t+\varepsilon| \ge 2\varepsilon$ show that truncated absolute loss separates the two values at scale $2\varepsilon \le 1$. Assouad's lemma (see, e.g., Tsybakov, 2009) gives

$$\inf_{\widetilde{\theta}} \max_{\omega} \mathbb{E}_\omega \sum_{j=1}^{d} \min(\left|\widetilde{\theta}_j - \theta_j^\omega\right|, 1) \ge c\, d\varepsilon$$

for a constant $c > 0$. Combining the three preceding bounds yields

$$\inf_{\widehat{q}} \max_{\omega} \mathbb{E}_\omega \sup_{f \in \mathcal{F}_{\mathrm{sub}}} \left| P_{f,\theta^\omega}\{Y \in \widehat{\mathcal{C}}(X)\} - (1-\alpha) \right| \ge \frac{c_\alpha}{2d}\, c\, d\varepsilon = c'_\alpha \sqrt{\frac{d}{n}}.$$

The cube lies in Θ, proving (S.58). This completes the proof. □

S12.14 Proof of Theorem S16

Two auxiliary lemmas control the moments of $h(Y)$ under Wasserstein and Kullback–Leibler perturbations of the teacher distribution, and Proposition S10 then converts them into the score perturbation bound.

Lemma S14. *Under Assumption S4 on a class $\mathcal{G} \subseteq \mathcal{H}$, for every $h \in \mathcal{G}$ and $x \in \mathcal{X}$, $|\mathbb{E}_{Q_x^{\mathrm{T}}} h(Y) - \mathbb{E}_{P_{Y|X=x}} h(Y)| \le L_{\mathcal{G}} W_1(Q_x^{\mathrm{T}}, P_{Y|X=x})$ and $|\mathrm{Var}_{Q_x^{\mathrm{T}}} h(Y) - \mathrm{Var}_{P_{Y|X=x}} h(Y)| \le 4 L_{\mathcal{G}} B_{\mathcal{G}} W_1(Q_x^{\mathrm{T}}, P_{Y|X=x})$. This implies that, whenever $\min\{\widetilde{\sigma}_h(x), \sigma_h^\circ(x)\} > 0$,*

$$|\widetilde{\sigma}_h(x) - \sigma_h^\circ(x)| \le \frac{2 L_{\mathcal{G}} B_{\mathcal{G}} W_1(Q_x^{\mathrm{T}}, P_{Y|X=x})}{\min\{\widetilde{\sigma}_h(x), \sigma_h^\circ(x)\}}.$$

Lemma S14 also gives $|\widetilde{\sigma}_h(x) - \sigma_h^\circ(x)| \le 4 L_{\mathcal{G}} B_{\mathcal{G}} W_1(Q_x^{\mathrm{T}}, P_{Y|X=x}) / \{\widetilde{\sigma}_h(x) + \sigma_h^\circ(x)\}$ whenever the denominator is positive. Thus a positive lower bound on the sum suffices for a linear comparison; either standard deviation may equal zero.

Proof of Lemma S14. Kantorovich duality (Villani, 2008) gives $|\mathbb{E}_{Q_x^{\mathrm{T}}} h - \mathbb{E}_{P_{Y|X=x}} h| \le L_{\mathcal{G}} W_1(Q_x^{\mathrm{T}}, P_{Y|X=x})$. For the variance, write

$$\begin{aligned} \mathrm{Var}_{Q_x^{\mathrm{T}}} h - \mathrm{Var}_{P_{Y|X=x}} h &= \mathbb{E}_{Q_x^{\mathrm{T}}} h^2 - (\mathbb{E}_{Q_x^{\mathrm{T}}} h)^2 - \mathbb{E}_{P_{Y|X=x}} h^2 + (\mathbb{E}_{P_{Y|X=x}} h)^2 \\ &= (\mathbb{E}_{Q_x^{\mathrm{T}}} h^2 - \mathbb{E}_{P_{Y|X=x}} h^2) - \left[(\mathbb{E}_{Q_x^{\mathrm{T}}} h)^2 - (\mathbb{E}_{P_{Y|X=x}} h)^2\right]. \end{aligned}$$

Because h^2 is $2B_{\mathcal{G}} L_{\mathcal{G}}$-Lipschitz, the first term is at most $2B_{\mathcal{G}} L_{\mathcal{G}} W_1$. The second is bounded by $|a^2 - b^2| \le 2B_{\mathcal{G}}|a-b| \le 2B_{\mathcal{G}} L_{\mathcal{G}} W_1$. Hence $|\mathrm{Var}_{Q_x^{\mathrm{T}}} h - \mathrm{Var}_{P_{Y|X=x}} h| \le 4 L_{\mathcal{G}} B_{\mathcal{G}} W_1$. Divide by the positive sum $\widetilde{\sigma}_h(x) + \sigma_h^\circ(x)$. When both scales are positive, the sum is at least $2\min\{\widetilde{\sigma}_h(x), \sigma_h^\circ(x)\}$, giving the stated bound. This completes the proof. □

Lemma S15. *Under Assumption S4 on a class $\mathcal{G} \subseteq \mathcal{H}$, for every $h \in \mathcal{G}$ and $x \in \mathcal{X}$, $|\mathbb{E}_{Q_x^{\mathrm{T}}} h(Y) - \mathbb{E}_{P_{Y|X=x}} h(Y)| \le 2B_{\mathcal{G}} \sqrt{D_{\mathrm{KL}}(P_{Y|X=x} \,\|\, Q_x^{\mathrm{T}})/2}$, and the absolute difference between the variances is bounded by $8B_{\mathcal{G}}^2 \sqrt{D_{\mathrm{KL}}(P_{Y|X=x} \,\|\, Q_x^{\mathrm{T}})/2}$. In addition,*

$$|\widetilde{\sigma}_h(x) - \sigma_h^\circ(x)| \le 2B_{\mathcal{G}} \sqrt{D_{\mathrm{KL}}(P_{Y|X=x} \| Q_x^{\mathrm{T}})}. \tag{S.86}$$

Lemma S15 bounds moment changes by conditional distribution error for bounded functionals. It gives the Kullback–Leibler term in the score comparison.

Proof of Lemma S15. Put $D_x = D_{\mathrm{KL}}(P_{Y|X=x} \,\|\, Q_x^{\mathrm{T}})$. Pinsker's inequality gives $\|Q_x^{\mathrm{T}} - P_{Y|X=x}\|_{\mathrm{TV}} \le \sqrt{D_x/2}$. Because $|h| \le B_{\mathcal{G}}$, the mean difference is at most $2B_{\mathcal{G}}\|Q_x^{\mathrm{T}} - P_{Y|X=x}\|_{\mathrm{TV}}$. Lemma S14's decomposition with $|h^2| \le B_{\mathcal{G}}^2$ gives the variance bound. For (S.86), define $P = P_{Y|X=x}$ and $Q = Q_x^{\mathrm{T}}$, and let p, q be their densities relative to $P + Q$. Here $\|\cdot\|_2$ denotes the $L_2(P+Q)$ norm, and all integrals use $P+Q$. Define $\mu_Q = \mathbb{E}_Q h$, $\sigma_P = \mathrm{Var}_P(h)^{1/2}$, $\sigma_Q = \mathrm{Var}_Q(h)^{1/2}$, and $H = \|\sqrt{p} - \sqrt{q}\|_2$. The distance-to-constants representation gives

$$\sigma_P = \inf_{c\in\mathbb{R}} \|(h-c)\sqrt{p}\|_2 \le \|(h-\mu_Q)\sqrt{p}\|_2 \le \sigma_Q + 2B_{\mathcal{G}}H.$$

Interchanging P, Q bounds the absolute difference. If $D_x < \infty$, Jensen's inequality gives

$$D_x \ge -2\log\int\sqrt{pq} \ge 2\left(1 - \int\sqrt{pq}\right) = H^2.$$

The bound also holds when $D_x = \infty$. This completes the proof. □

Proof of Theorem S16, upper bound. Put $\delta_q = \delta/3$, $\widetilde{\mu}_h(x) = \mathbb{E}_{Q_x^{\mathrm{T}}} h(Y)$, and $\widetilde{\sigma}_h(x) = \mathrm{Var}_{Q_x^{\mathrm{T}}}\{h(Y)\}^{1/2}$. Apply Lemma S15 and, under Assumption S4, Lemma S14 with $\mathcal{G} = \mathcal{H}_q$. An unavailable Wasserstein bound is $+\infty$. Uniformly over $(x,h) \in \mathcal{X}_q \times \mathcal{H}_q$,

$$|\widetilde{\mu}_h(x) - \mu_h^\circ(x)| \le \min\{L_{\mathcal{H}}\varepsilon_W,\, 2B_{\mathcal{H}}\sqrt{\varepsilon_K/2}\}$$

and, when the positive lower bound on the sum of the two scales is imposed,

$$|\widetilde{\sigma}_h(x) - \sigma_h^\circ(x)| \le C_1 \min\left\{\frac{L_{\mathcal{H}}B_{\mathcal{H}}}{\sigma_0}\varepsilon_W,\; \frac{B_{\mathcal{H}}^2}{\sigma_0}\sqrt{\varepsilon_K/2}\right\}.$$

Without the scale lower bound, (S.86) gives $|\widetilde{\sigma}_h(x) - \sigma_h^\circ(x)| \le 2B_{\mathcal{H}}\sqrt{\varepsilon_K}$ independently of σ_0. Lemma S6, conditional on $\{Q_x^{\mathrm{T}}\}_{x\in\mathcal{X}_q}$, gives for $\widehat{\mu}_h(x) = K^{-1}\sum_k h(\widetilde{Y}^{(k)}(x))$ and $\widehat{\sigma}_h(x)$,

$$\sup_{x\in\mathcal{X}_q, h\in\mathcal{H}_q} |\widehat{\mu}_h(x) - \widetilde{\mu}_h(x)| + \sup_{x\in\mathcal{X}_q, h\in\mathcal{H}_q} |\widehat{\sigma}_h(x) - \widetilde{\sigma}_h(x)| \le C_2\sqrt{\frac{1 + \log(|\mathcal{X}_q||\mathcal{H}_q|/\delta_q)}{K}}$$

with probability at least $1 - \delta_q$. Put

$$a_n^{(q)} = \sup_{x\in\mathcal{X}_q, h\in\mathcal{H}_q} |\widehat{\mu}_h(x) - \mu_h^\circ(x)|, \qquad b_n^{(q)} = \sup_{x\in\mathcal{X}_q, h\in\mathcal{H}_q} |\widehat{\sigma}_h(x) - \sigma_h^\circ(x)|.$$

Combining these bounds gives, with C depending on $B_{\mathcal{H}}$ and λ, and on σ_0 only for the linear Wasserstein branch,

$$a_n^{(q)} + b_n^{(q)} \le C\left\{\min\{L_{\mathcal{H}}\varepsilon_W,\, B_{\mathcal{H}}\sqrt{\varepsilon_K}\} + \sqrt{\frac{1 + \log(|\mathcal{X}_q||\mathcal{H}_q|/\delta_q)}{K}}\right\}.$$

Proposition S10 restricted to $\mathcal{X}_q \times \mathcal{H}_q$ proves (S.59). Proposition S2 gives $|\widehat{q}_{q,K} - \widehat{q}_q^\circ| \le e_{n,q}$. If the local-quantile and two empirical-process conditions of Theorem S8 hold and $e_{n,q} + 3\psi_{n,q}(\delta_q, \mathcal{F}) \le c_0 r_{0,q}$, its proof also gives

$$\left|\widehat{q}_{q,K} - q^\circ_{\alpha,\mathcal{F},q}\right| \le C\{e_{n,q} + \psi_{n,q}(\delta_q,\mathcal{F})\}, \qquad \left|\widehat{q}_q^\circ - q^\circ_{\alpha,\mathcal{F},q}\right| \le C\psi_{n,q}(\delta_q,\mathcal{F}),$$

The paired comparison, without these additional conditions, gives $\left|\widehat{q}_{q,K} - \widehat{q}_q^\circ\right| \le C\{e_{n,q} + \psi_{n,q}(\delta_q,\mathcal{F})\}$. On (S.59),

$$\mathcal{C}^{R_{n,K}}_{\widehat{q}_{q,K}}(x) \subseteq \mathcal{C}^{R_q^\circ}_{\widehat{q}_{q,K}+e_{n,q}}(x), \qquad \mathcal{C}^{R_q^\circ}_{\widehat{q}_q^\circ}(x) \subseteq \mathcal{C}^{R_{n,K}}_{\widehat{q}_q^\circ+e_{n,q}}(x).$$

This implies that

$$\left|\mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\widehat{\mathcal{C}}_{q,K}) - \mathcal{S}_{\mathcal{F},\mathcal{H}_q}(\widehat{\mathcal{C}}_q^\circ)\right| \le A_{\mathcal{F}}\left\{\left|\widehat{q}_{q,K} - \widehat{q}_q^\circ\right| + e_{n,q}\right\},$$

which establishes (S.60). This completes the proof of the upper bound. □

Proof of Theorem S16, lower bound. Take $\mathcal{X} = \{x_0\}$, $\mathcal{Y} = [0,1]$, $d_{\mathcal{Y}}(y,y') = |y-y'|$, $\mathcal{F} = \{1\}$, and $\mathcal{H} = \{h_0\}$ with $h_0(y) = y$, hence $L_{\mathcal{H}} = B_{\mathcal{H}} = 1$. Let $P_{Y|X=x_0} = \delta_0$ and, for $\epsilon \in (0, 1/2]$, $Q^{\mathrm{T}}_{x_0,\epsilon} = \delta_\epsilon$. Then $W_1(Q^{\mathrm{T}}_{x_0,\epsilon}, P_{Y|X=x_0}) = \epsilon = \varepsilon_W$. For $K = \infty$, the two scores are

$$R^\circ(x_0,y) = \frac{|y|}{\lambda}, \qquad R_n(x_0,y) = \frac{|y-\epsilon|}{\lambda}.$$

All calibration responses are 0, giving oracle scores 0 and teacher scores ϵ/λ. For $\tau_n < 1-\alpha$, (S.30) gives

$$\widehat{q}^\circ_{\tau_n} = 0, \qquad \widehat{q}_{\tau_n} = \frac{\epsilon}{\lambda}$$

almost surely. This implies that

$$\widehat{\mathcal{C}}^\circ_{\tau_n}(x_0) = \left\{ y \in [0,1] : \frac{|y|}{\lambda} \le 0 \right\} = \{0\}, \qquad \widehat{\mathcal{C}}_{\tau_n}(x_0) = \left\{ y \in [0,1] : \frac{|y-\epsilon|}{\lambda} \le \frac{\epsilon}{\lambda} \right\} = [0, 2\epsilon].$$

By the definition (S.23), $\mathrm{rad}_{\mathcal{H}}\{\widehat{\mathcal{C}}^\circ_{\tau_n}(x_0)\} = 0$ and $\mathrm{rad}_{\mathcal{H}}\{\widehat{\mathcal{C}}_{\tau_n}(x_0)\} = \epsilon$. This implies that

$$\mathcal{S}_{\mathcal{F},\mathcal{H}}(\widehat{\mathcal{C}}_{\tau_n}) - \mathcal{S}_{\mathcal{F},\mathcal{H}}(\widehat{\mathcal{C}}^\circ_{\tau_n}) = \epsilon = \varepsilon_W,$$

which establishes (S.61). When only the condition on the Wasserstein distance is imposed, the lower bound is $L_{\mathcal{H}}\varepsilon_W$. This completes the proof. □

S12.15 Proofs of Theorems S17 and S18

A fixed calibrated set satisfies the same coverage bound under each protected target weight. Approximation error controls changes outside that class.

Proof of Theorem S17. Let $\psi_{n,\delta} = \psi_n(\delta, \mathcal{F})$. Condition on $\mathcal{G}_n$, which fixes $\widehat{\mathcal{C}}$. Under the event specified in Theorem S9,

$$\Delta^-_{\mathcal{F}}(\widehat{\mathcal{C}}) = \sup_{f \in \mathcal{F}} \big[(1-\alpha) - P_f\{Y \in \widehat{\mathcal{C}}(X)\}\big]_+ \le \tau_n + \psi_{n,\delta}.$$

For $f_t^\star \in \mathcal{F}$, the same supremum bounds conditional undercoverage. Average over $t = 1, \dots, T$. The event is $\mathcal{G}_n$-measurable and needs no union bound over time. This completes the proof. □

Proof of Theorem S18. Let $\psi_{n,\delta} = \psi_n(\delta, \mathcal{F})$. For each t, Theorem 4 applied with $r = f_t^\star$ yields

$$P_{f_t^\star}\{Y_{n+t} \in \widehat{\mathcal{C}}(X_{n+t}) \mid \mathcal{G}_n\} \ge 1 - \alpha - \Delta^-_{\mathcal{F}}(\widehat{\mathcal{C}}) - \kappa_t.$$

This implies that

$$\big[(1-\alpha) - P_{f_t^\star}\{Y_{n+t} \in \widehat{\mathcal{C}}(X_{n+t}) \mid \mathcal{G}_n\}\big]_+ \le \Delta^-_{\mathcal{F}}(\widehat{\mathcal{C}}) + \kappa_t.$$

Averaging over t and applying the bound $\Delta^-_{\mathcal{F}}(\widehat{\mathcal{C}}) \le \tau_n + \psi_{n,\delta}$ from Theorem S9 yields the desired result. This completes the proof. □

Supplementary References

Anastasios N. Angelopoulos, Amit Pal Kohli, Stephen Bates, Michael I. Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 717–730, 2022.

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

Felipe Areces, Chen Cheng, John Duchi, and Rohith Kuditipudi. Two fundamental limits for uncertainty quantification in predictive inference. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 186–218. PMLR, 2024. URL https://proceedings.mlr.press/v247/areces24a.html.

Konstantina Bairaktari, Jiayun Wu, and Steven Wu. Kandinsky conformal prediction: Beyond class- and covariate-conditional coverage. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 2581–2602, 2025. URL https://proceedings.mlr.press/v267/bairaktari25a.html.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021. doi: 10.1093/imaiai/iaaa017.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Albert Cohen, Wolfgang Dahmen, Ingrid Daubechies, and Ronald A. DeVore. Harmonic analysis of the space BV. *Revista Matemática Iberoamericana*, 19(1):235–263, 2003. doi: 10.4171/RMI/345.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. doi: 10.1093/jamia/ocv080.

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

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C. S. Valentim, Huiying Liang, Sally L. Baxter, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, 2018. doi: 10.1016/j.cell.2018.02.010.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Ruhan Liu, Xiangning Wang, Qiang Wu, Ling Dai, Xi Fang, Tao Yan, Jaemin Son, Shiqi Tang, Jiang Li, Zijian Gao, Adrian Galdran, J. M. Poorneshwaran, Hao Liu, Jie Wang, Yerui Chen, Prasanna Porwal, Gavin Siew Wei Tan, Xiaokang Yang, Chao Dai, Haitao Song, Mingang Chen, Huating Li, Weiping Jia, Dinggang Shen, Bin Sheng, and Ping Zhang. DeepDRiD: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns*, 3(6):100512, 2022.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990. doi: 10.1214/aop/1176990746.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://www.cs.mcgill.ca/~colt2009/papers/012.pdf.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

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

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

Jacopo Teneggi, Matthew Tivnan, Web Stayman, and Jeremias Sulam. How to trust your diffusion model: A convex optimization approach to conformal risk control. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33940–33960. PMLR, 2023. URL https://proceedings.mlr.press/v202/teneggi23a.html.

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

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2008.

Weiran Wang and Canyi Lu. Projection onto the capped simplex, 2015. URL https://arxiv.org/abs/1503.01002.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3462–3471, 2017. doi: 10.1109/CVPR.2017.369.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2 – a large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Yachong Yang, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Doubly robust calibration of prediction sets under covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(4):943–965, 2024. doi: 10.1093/jrsssb/qkae009.