When Average Calibration Fails: Site-Conditional Federated Conformal Risk Control

Nafis Fuad Shahid

Dhaka, Bangladesh
nafisfuadshahid@gmail.com

Abstract. Conformal risk control (CRC) provides distribution-free segmentation guarantees by calibrating a prediction-set threshold on held-out data. In federated deployments, the standard approach pools calibration scores into a single threshold. We quantify, on real multi-institutional brain tumor data (FeTS-2022, 1,251 subjects, 20 institutions), a critical failure: *naive pooled* CRC protects the average hospital but violates coverage at 40% of individual institutions, with the worst site exceeding the target false-negative rate by 7.8 percentage points. We trace this failure to a hidden design choice: the aggregation weights implicitly determine *whose* coverage is protected. Sample-size weighting optimizes patient-level validity but can sacrifice institution-level reliability; equal-site weighting improves institution-level reliability on this benchmark at comparable efficiency, using only a single scalar per site. We propose risk-curve shrinkage as a principled mechanism: each site transmits its empirical risk curve (G scalars) and a single hyperparameter n_0 smoothly interpolates between site-specific local calibration and sample-size-weighted pooled calibration. Leave-one-site-out sensitivity analysis identifies n_0=19, achieving 2.7/20 violations at 2.0× stretch. Direct Lagrangian budget optimization *fails* by concentrating risk on vulnerable hospitals; the finite-sample correction term is essential: removing it triples violations. No patient-level images, masks, or per-volume scores leave any site.

Keywords: Federated Learning · Conformal Risk Control · Segmentation · Uncertainty Quantification

1 Introduction

Deploying segmentation models across hospitals requires calibrated uncertainty: clinicians must know when a model's output can be trusted and when it cannot. Conformal risk control (CRC) [1] provides distribution-free, finite-sample guarantees of the form $\mathbb{E}[\ell(C_\lambda(X), Y)] \leq \alpha$, where ℓ is a monotone loss and λ is calibrated on held-out data. CRC has been applied to centralized medical segmentation, including morphological dilation families [10], conditional CRC [8], and semantic CRC [14], but never in the federated setting where calibration data is distributed across heterogeneous institutions that cannot share patient data.

In federated learning (FL) [9], the natural approach pools calibration scores from all sites to compute a single global threshold, inheriting federated conformal prediction machinery [7]. We show this *naive pooled CRC* harbors a critical failure mode: it satisfies the marginal coverage guarantee while violating coverage at 40% of individual institutions. On FeTS-2022 brain tumor data [12] with 20 institutions, the worst site reaches FNR=0.178; such site-specific errors could translate into missed tumor regions in downstream clinical workflows, motivating institution-level evaluation. This marginal-vs-conditional gap is well studied in CP theory [16,4], but its magnitude in federated medical segmentation has not been quantified.

Crucially, we show this failure is not inherent to scalar communication: an *unweighted* scalar-threshold average achieves comparable reliability to our full method. The root cause is the *aggregation weighting*: sample-size weights optimize patient-level validity but can sacrifice institution-level reliability. This reveals federated calibration as an *objective-selection* problem, not merely an estimation problem.

Contributions. (1) We quantify the marginal-conditional coverage gap for federated CRC on 1,251 real multi-institutional brain tumor volumes: $8/20$ institutions fail while the average appears calibrated. (2) We identify aggregation weighting as a hidden determinant of federated coverage: sample-size weighting targets patient-level validity, equal-site weighting targets institution-level reliability, and these diverge on FeTS-2022. (3) We propose risk-curve shrinkage as a controllable mechanism, with n_0 interpolating between local and pooled calibration, validated via sensitivity analysis. (4) Direct budget optimization fails; the finite-sample correction is essential: removing it triples violations.

2 Related Work

CRC for segmentation. Angelopoulos et al. [1] introduced CRC for monotone losses. Recent centralized extensions include morphological dilation families [10], conditional CRC [8], and semantic CRC [14]. All assume centralized calibration data.

Federated conformal prediction. Lu et al. [7] prove federated CP coverage under partial exchangeability for classification. Extensions address group-conditional [17], Byzantine-robust [6], privacy-preserving [13], and weighted-quantile [11] settings. To our knowledge, these methods do not address pixel-level segmentation CRC. The cell *federated* $\times$ *CRC* $\times$ *pixel-level segmentation* is empty. Exact conditional coverage is impossible without structural assumptions [16,4]; our per-site coverage is empirical, viewed as approximate group-conditional coverage where each "group" is a hospital.

Federated segmentation with uncertainty. FUNAvg [15] aggregates MC-dropout uncertainty; FedEvi [3] uses evidential learning. Neither provides formal coverage guarantees. FedStein [5] applies James–Stein estimation to federated batch-normalization statistics; we apply shrinkage to empirical risk curves for post-hoc calibration.

3 Method

3.1 Setting

Consider K sites with n_k calibration volumes (X_i^k, Y_i^k) each. A pre-trained model f produces predicted probabilities via sigmoid activation. We define nested prediction sets $C_\lambda(X) = \{v : f(X)_v \geq 1-\lambda\}$, growing with λ. The per-volume false-negative rate $\ell(C_\lambda, Y) = 1 - |C_\lambda \cap Y|/|Y|$ is non-increasing in λ and bounded in $[0, 1]$, so the CRC loss bound $B = \sup \ell = 1$ [1]. Goal: find $\hat{\lambda}$ such that $\mathbb{E}[\ell(C_{\hat{\lambda}}(X_{\text{test}}), Y_{\text{test}})] \leq \alpha$.

3.2 Pooled CRC and Its Hidden Objective

Centralized CRC [1] sets $\hat{\lambda} = \inf\{\lambda : \hat{R}(\lambda) + B/(n+1) \leq \alpha\}$ where $\hat{R}(\lambda) = \frac{1}{n}\sum_i \ell_i(\lambda)$. *Naive pooled federated CRC* pools all $N = \sum_k n_k$ scores, implicitly optimizing the **patient-uniform** objective $R_{\text{micro}}(\lambda) = \sum_k (n_k/N) R_k(\lambda)$: large hospitals dominate. Under partial exchangeability [7] this controls marginal risk but *not* site-conditional risk $\mathbb{E}[\ell \mid H{=}k]$.
An alternative **site-uniform** objective treats each institution equally: $R_{\text{macro}}(\lambda) = \frac{1}{K}\sum_k R_k(\lambda)$. The gap between these objectives is:

$$R_{\text{micro}}(\lambda) - R_{\text{macro}}(\lambda) = \frac{\text{Cov}_k(n_k, R_k(\lambda))}{\bar{n}}, \tag{1}$$

which is nonzero whenever hospital size correlates with site risk. On FeTS-2022, the contrast between weighted and unweighted threshold aggregation in Table 1 shows that the weights assigned to local operating points materially affect site-level reliability.

3.3 Calibration Poverty at Small Sites

Each site can independently set $\hat{\lambda}_k = \inf\{\lambda : \hat{R}_k(\lambda) + B/(n_k+1) \leq \alpha\}$. This guarantees site-conditional coverage, but for small n_k the finite-sample correction $B/(n_k+1)$ is prohibitively large: with $n_k{=}5$ and $B{=}1$ it alone exceeds $\alpha{=}0.10$, making nontrivial calibration impossible and forcing $\hat{\lambda}_k = \lambda_{\max}$ with stretch above $80\times$. We term this *calibration poverty*: small hospitals cannot calibrate locally without enormous prediction sets, motivating information borrowing from the federation.

3.4 Shrinkage-Based Federated CRC

We propose an approach that borrows information at the *risk-curve* level (Fig. 1). Each site transmits its empirical risk curve $\hat{R}_k(\lambda)$ on a grid Λ of G points. The server computes $\hat{R}_{\text{global}}(\lambda) = \frac{1}{N}\sum_k n_k \hat{R}_k(\lambda)$ and the **shrinkage risk curve**:

$$\hat{R}_k^{\text{shrink}}(\lambda) = w_k \cdot \hat{R}_k(\lambda) + (1 - w_k) \cdot \hat{R}_{\text{global}}(\lambda) + \text{corr}_k, \tag{2}$$

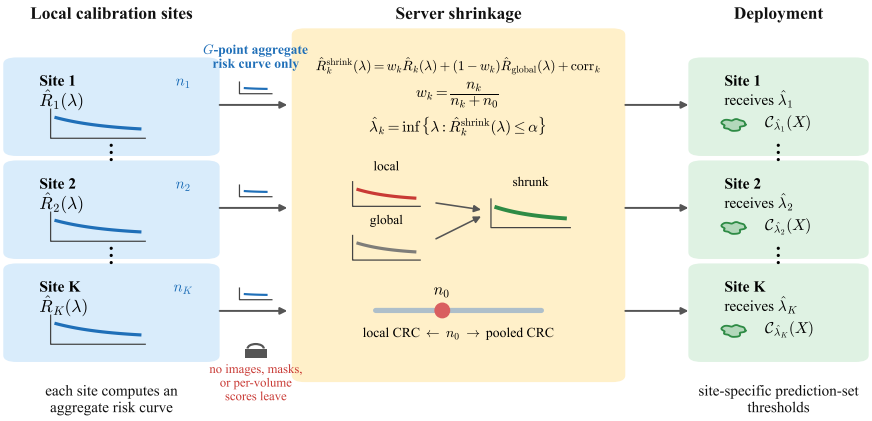

Fig. 1. Protocol overview. Each site transmits only its empirical risk curve (G scalars, ∼0.8 KB for G=200) to the server, which computes shrinkage-regularized per-site thresholds and broadcasts them back. No patient-level images, masks, or per-volume scores leave any site.

where $w_k = n_k/(n_k + n_0)$ is the shrinkage weight, the standard empirical Bayes form where n_0 acts as prior precision [5], and $\text{corr}_k = w_k \cdot B/(n_k{+}1) + (1{-}w_k) \cdot B/(N{+}1)$ is a heuristic interpolation between per-site and pooled finite-sample corrections, motivated by but not formally derived from CRC theory. The threshold is $\hat{\lambda}_k^{\text{shrink}} = \inf\{\lambda \in \Lambda : \hat{R}_k^{\text{shrink}}(\lambda) \le \alpha\}$. As $n_0 \to 0$ this recovers per-site local CRC; as $n_0 \to \infty$ it recovers pooled CRC. The hyperparameter n_0 thus provides a dial between local, site-specific calibration and pooled, patient-weighted calibration.

3.5 Marginal Safeguard

Proposition 1 (Marginal monotonicity safeguard). *Assume that the pooled threshold* $\hat{\lambda}_{\text{pool}}$ *satisfies* $\mathbb{E}[\ell(C_{\hat{\lambda}_{\text{pool}}}(X_{\text{test}}), Y_{\text{test}})] \le \alpha$ *under the target test mixture. For arbitrary site-specific thresholds* $\{\hat{\lambda}_k\}$*, define* $\hat{\lambda}_{\text{fed}} = \max\{\hat{\lambda}_{\text{pool}}, \max_k \hat{\lambda}_k\}$*. Then* $\mathbb{E}[\ell(C_{\hat{\lambda}_{\text{fed}}}(X_{\text{test}}), Y_{\text{test}})] \le \alpha$*.*

Proof. $\hat{\lambda}_{\text{fed}} \ge \hat{\lambda}_{\text{pool}}$ and ℓ non-increasing in λ give $\ell(C_{\hat{\lambda}_{\text{fed}}}) \le \ell(C_{\hat{\lambda}_{\text{pool}}})$ pointwise, so the expectation bound follows.

Remark. This safeguard holds for *any* choice of per-site thresholds, as long as the deployed threshold includes $\hat{\lambda}_{\text{pool}}$ in the maximum. Deploying $\hat{\lambda}_{\text{fed}}$ globally achieves 0 violations at n_0=9 but at 67× stretch (Table 2); per-site thresholds are therefore used in practice and evaluated empirically. Exact conditional coverage is impossible without structural assumptions [16,4].

4 Experiments

4.1 Setup

Data. FeTS-2022 training set [12]: 1,251 multi-modal brain MRI volumes from 23 institutions. After excluding institutions with fewer than six subjects, our analysis retains 20 institutions (calibration sizes: 3–255). Per-site 50/50 cal/test splits with seeds $\{42, 1337, 2024\}$.

Model. Pre-trained SegResNet from the MONAI model zoo [2], trained on BraTS-2021.

Baselines. B3 (Naive Pooled): pools all scores (patient-uniform weighting). **B2 (Per-site Local):** independent per-site thresholds. To test whether the failure is caused by scalar communication or by weighting, we include two scalar-threshold baselines: each site computes its local CRC threshold $\hat{\lambda}_k$ and the server broadcasts either the **sample-size-weighted average** $\hat{\lambda} = \sum_k (n_k/N)\hat{\lambda}_k$, following FedWQ-CP [11], or the **equal-site average** $\hat{\lambda} = \frac{1}{K}\sum_k \hat{\lambda}_k$.

CRC details. Loss: per-volume pixel-FNR. Prediction sets: $C_\lambda = \{v : f(X)_v \geq 1-\lambda\}$ on G=200 uniformly spaced $\lambda \in [0, 1]$, monotonicity enforced via cumulative minimum over increasing λ.
Target $\alpha = 0.10$, swept over $\{0.05, 0.10, 0.15, 0.20\}$.

Metrics. (i) Violations: sites with mean test FNR $> \alpha$ (mean±std over seeds). (ii) Worst-site FNR (mean over seeds). (iii) Stretch: $|C_{\hat{\lambda}}|/|Y|$ (mean over seeds).

4.2 Main Results

Table 1 presents results across three seeds. B3 violates coverage at 8.0 ± 2.4 of 20 sites, with worst FNR = 0.178, nearly double the target. B2 reduces violations to 1.3 but inflates stretch to 83× due to calibration poverty at small sites.

The weighting, not the representation, drives the failure. Weighted scalar aggregation (8.3 ± 2.9 violations) performs comparably to naive pooling: finite-sample correction pushes many calibration-poor sites to the conservative boundary $\hat{\lambda}_k$=$\lambda_{\max}$, and sample-size weighting suppresses their influence. Unweighted scalar aggregation preserves the influence of these conservative local thresholds, achieving 1.7 ± 0.9 violations at 2.3× stretch. Both scalar baselines require only 4 bytes/site, compared with ∼800 bytes/site for risk-curve transmission.

Risk-curve shrinkage provides a *controllable* trade-off. At n_0=9: 1.3 violations, 29× stretch. At n_0=15: 2.7 violations, 4.4× stretch. LOSO sensitivity analysis (Sec. 4.3) identifies n_0=19: 2.7 violations at 2.0× stretch, a 97.6% reduction in average stretch relative to B2. The failure is not a small-sample artifact: Fig. 2 shows that Institution 18 (382 patients, 30% of the dataset) is miscovered at FNR=0.150 under pooling, and B2's stretch explodes to 100−250× at small sites, while ours reduces average stretch to 4.4×.

Table 1. Results on FeTS-2022 (20 institutions, α=0.10, 3 seeds).

Method	**Violations** (↓)	**Worst FNR** (↓)	**Stretch** (↓)
B3: Naive Pooled	8.0 ± 2.4	0.178	1.5×
B2: Per-site Local	1.3 ± 1.2	0.111	83.2×
Weighted λ Agg.	8.3 ± 2.9	0.186	1.4×
Unweighted λ Agg.	1.7 ± 0.9	0.113	2.3×
Budget Alloc. (uncapped)	12.3 ± 0.9	0.351	1.4×
Budget Alloc. (δ=0.03)	2.7 ± 1.2	0.142	71.6×
Ours (n_0=9)	1.3 ± 1.2	0.112	28.8×
Ours (n_0=15)	2.7 ± 1.7	0.119	4.4×
Ours (n_0=19, LOSO)	2.7 ± 1.7	0.125	**2.0**×

4.3 The n_0 Dial and LOSO Sensitivity Analysis

Fig. 3 shows how n_0 controls the coverage-efficiency frontier. Small n_0 (≤ 9): few violations, high stretch (local regime). Large n_0 (≥ 30): low stretch, many violations (pooled regime). The knee at $n_0 \in [10, 20]$ offers the best trade-off.

As a post-hoc sensitivity analysis, we perform leave-one-site-out evaluation: for each candidate n_0, hold out one site, recompute the global curve from the remaining $K-1$ sites, compute the shrinkage threshold for the held-out site using its own calibration curve, and evaluate on the held-out site's test partition. LOSO identifies n_0=19 as the lowest-stretch operating point with $\leq$3 mean violations.

4.4 Why Direct Optimization Fails

A budget-allocation baseline minimizes total stretch subject to marginal risk $\leq \alpha$ (binary search on Lagrange multiplier μ, with $L_k(\lambda) = \hat{R}_k(\lambda) + B/(n_k{+}1)$ and $p_k = (n_k{+}1)/(N{+}K)$). Uncapped, this achieves 1.4× stretch but fails 12.3/20 sites (worst FNR=0.351): the optimizer concentrates risk on small, hard institutions. Adding a per-site cap (δ=0.03) reduces violations to 2.7 but inflates stretch to 71.6×, worse than shrinkage on *both* dimensions.

4.5 Ablations

Table 2 probes three axes.

The correction corr_k is essential, making this the most consequential ablation. Removing corr_k: violations jump from 1.3–2.7 to 8.0–9.3 across all n_0, comparable to naive pooling. corr_k interpolates between local and pooled corrections; its formal analysis is future work, but its empirical necessity is unambiguous.

Conservative $\hat{\lambda}_{\mathrm{fed}}$. The global threshold of Proposition 1 achieves 0 violations at n_0=9 but 67× stretch, confirming per-site deployment is necessary for clinical utility.

Grid G. G=200 is stable; G=50 inflates stretch to 35.7× (too coarse).

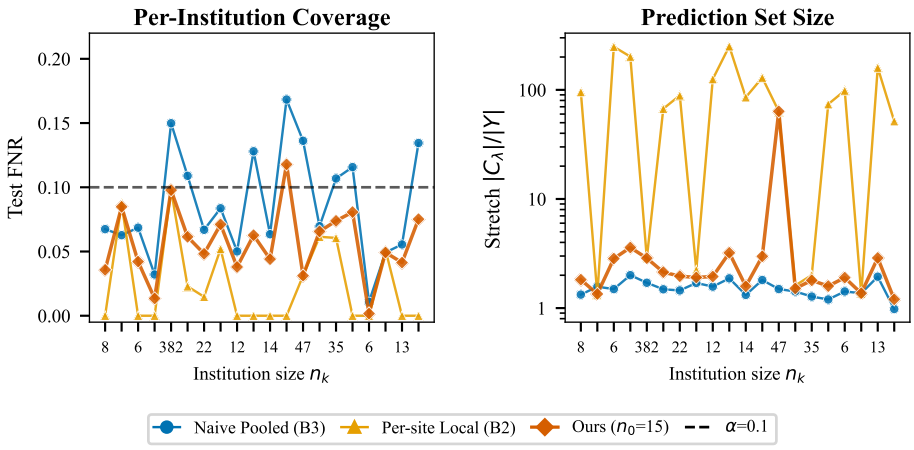

Fig. 2. Per-institution FNR (left) and prediction-set stretch (right) on FeTS-2022. B3 (blue) violates α=0.10 at 8/20 sites; B2 (orange) largely restores coverage at 83× stretch. Ours (red, n_0=15): practical stretch with 2–3 borderline violations. Scalar-aggregation baselines are reported in Table 1.

Table 2. Ablation results (α=0.10, mean over 3 seeds).

Ablation	n_0	Viol.	W-FNR	Stretch
No corr$_k$	9	9.3 ± 2.6	0.181	1.5×
No corr$_k$	15	8.3 ± 2.6	0.173	1.5×
$\hat{\lambda}_{\text{fed}}$ (Prop. 1)	9	0.0 ± 0.0	0.050	67.3×
$\hat{\lambda}_{\text{fed}}$ (Prop. 1)	15	0.3 ± 0.5	0.071	46.7×
Grid G=50	15	2.3 ± 1.9	0.116	35.7×
Grid G=500	15	3.3 ± 1.9	0.122	3.0×

5 Discussion

Weighting, not communication, drives the failure. The contrast between weighted and unweighted scalar aggregation (Table 1) reveals that the aggregation weights silently determine *whose* coverage is protected. The mechanism on FeTS-2022 is calibration poverty: finite-sample correction pushes many small sites to the conservative boundary, and sample-size weighting suppresses their influence while equal-site weighting preserves it. A 2× stretch means the prediction set contains roughly twice the ground-truth tumor volume, far below the 83× inflation of local CRC and plausibly reviewable in clinical workflows. On FeTS-2022 the scalar unweighted average is competitive; risk-curve transmission becomes advantageous when sites need different points on the coverage-efficiency frontier, or when downstream losses require the full risk curve rather than a single threshold. Designing federated calibration methods that explicitly separate

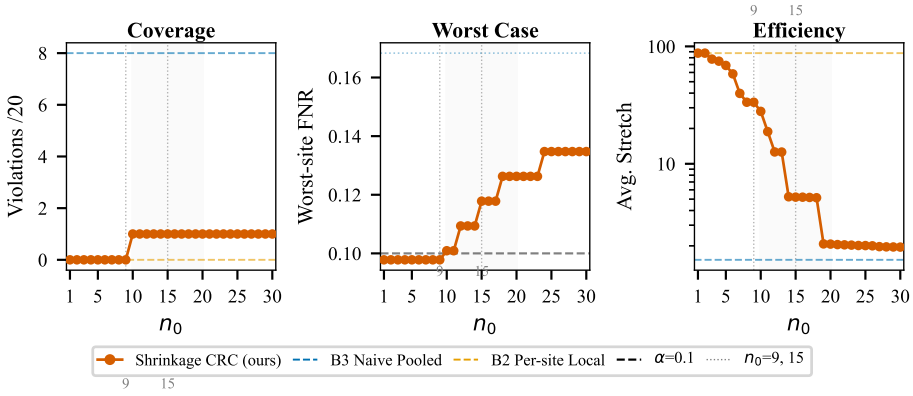

Fig. 3. Effect of n_0 on violations (left), worst-site FNR (center), and stretch (right). Dashed: B3 (blue) and B2 (orange). Shaded: sweet spot $n_0 \in [10, 20]$.

target weights (whose risk is controlled) from borrowing weights (whose information is used) is an important direction for future work.

Limitations. Proposition 1 is conditional on pooled-threshold validity; a finite-sample CRC correction under partial exchangeability remains open. We use a single pre-trained model on one dataset; validation with additional anatomies, backbones, and models trained federatively is needed. Adapting group-conditional FCP [17] to pixel-level CRC remains future work.

6 Conclusion

We quantified a failure mode of naive pooled federated calibration: on real multi-institutional brain tumor data, 40% of hospitals exceed the target false-negative rate while the average looks fine. This failure is driven by aggregation weights, not scalar communication: equal-site weighting improves reliability at comparable efficiency. Risk-curve shrinkage provides a controllable family of operating points between local and pooled calibration, with n_0 validated via sensitivity analysis. We encourage the community to evaluate calibration guarantees *per site*, not just on average.

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

References

1. Angelopoulos, A.N., Bates, S., Fisch, A., Lei, L., Schuster, T.: Conformal risk control. In: ICLR (2024)
2. Cardoso, M.J., et al.: MONAI: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)

3. Chen, J., Ma, B., Cui, H., Xia, Y.: FedEvi: Improving federated medical image segmentation via evidential weight aggregation. In: MICCAI (2024)
4. Foygel Barber, R., Candès, E.J., Ramdas, A., Tibshirani, R.J.: The limits of distribution-free conditional predictive inference. Information and Inference **10**(2), 455–482 (2021)
5. Gupta, S., Jangid, N., Sethi, A.: FedStein: Enhancing multi-domain federated learning through James-Stein estimator. In: AAAI Bridge Workshop (2025)
6. Kang, M., Lin, Z., Sun, J., Xiao, C., Li, B.: Certifiably Byzantine-robust federated conformal prediction. In: ICML (2024)
7. Lu, C., Yu, Y., Karimireddy, S.P., Jordan, M.I., Raskar, R.: Federated conformal predictors for distributed uncertainty quantification. In: ICML (2023)
8. Luo, R., Zhou, Z.: Conditional conformal risk adaptation. arXiv preprint arXiv:2504.07611 (2025)
9. McMahan, B., et al.: Communication-efficient learning of deep networks from decentralized data. In: AISTATS (2017)
10. Mossina, L., Friedrich, C.: Conformal prediction for image segmentation using morphological prediction sets. In: MICCAI (2025)
11. Nguyen, Q.H., Wang, J., Ku, W.S.: Conformalized neural networks for federated uncertainty quantification under dual heterogeneity. arXiv preprint arXiv:2602.23296 (2026)
12. Pati, S., et al.: Federated learning enables big data for rare cancer boundary detection. Nature Communications **13**, 7346 (2022)
13. Plassier, V., Makni, M., Rubashevskii, A., Moulines, E., Panov, M.: Conformal prediction for federated uncertainty quantification under label shift. In: ICML (2023)
14. Teneggi, J., Stayman, J.W., Sulam, J.: Conformal risk control for semantic uncertainty quantification in computed tomography. In: MICCAI (2025)
15. Tölle, M., Navarro, F., Eble, S., Wolf, I., Menze, B., Engelhardt, S.: FUNAvg: Federated uncertainty weighted averaging for datasets with diverse labels. In: MICCAI (2024)
16. Vovk, V.: Conditional validity of inductive conformal predictors. Machine Learning **92**(2–3), 349–376 (2012)
17. Wen, H., Simeone, O., Xing, H.: Efficient federated conformal prediction with group-conditional guarantees. arXiv preprint arXiv:2603.14198 (2026)