

Bound-Aware Per-Organ Recall Risk Control for Multi-Organ CT Segmentation under Clinical Domain Shift

Souraj Adhikary¹, Negar Chabi¹, and Andre Mastmeyer¹

Department of Engineering Sciences, Jade Hochschule Wilhelmshaven, Germany

souraj.adhikary@student.jade-hs.de

{negar.chabi, andre.mastmeyer}@jade-hs.de

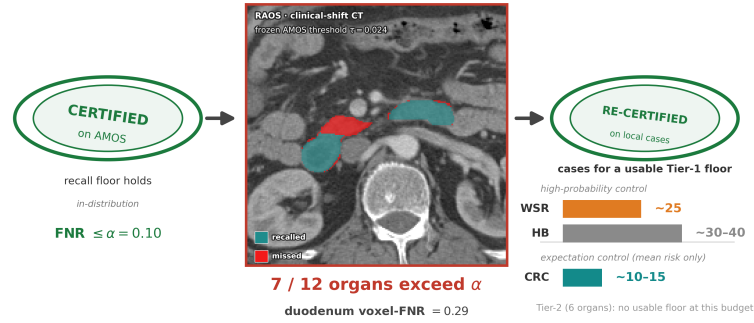


Fig. 1. Frozen AMOS thresholds often fail on RAOS. *Centre:* one RAOS case at the duodenum threshold $\tau=0.024$; red marks missed ground-truth voxels, teal recalled ones (case FNR 0.29, cohort mean 0.30). With pooled AMOS out-of-fold calibration, 7/12 organs exceed $\alpha=0.10$ on RAOS; only three of the other five have a non-vacuous threshold. Here, re-certification varies with the bound and budget (Section 5.3).

Abstract. Distribution-free risk control adds organ-specific recall guarantees to frozen segmentation. We calibrate per-organ thresholds for an AMOS-trained nnU-Net, audit transfer to RAOS, and estimate local re-certification cost using case-level voxel false-negative rate (FNR). The AMOS control passes, but 7/12 organs exceed $\alpha=0.10$ after transfer; smaller calibration sets can mask exceedances with conservative or vacuous thresholds. Risk-Controlling Prediction Sets (RCPS) give high-probability control of population-mean risk, whereas Conformal Risk Control (CRC) gives weaker expectation control. Both require exchangeability; fixed and global thresholds give no per-organ guarantee. The Waudby-Smith-Ramdas (WSR) betting bound re-certifies six Tier-1 organs with 25 local cases, versus 30–40 for Hoeffding-Bentkus (HB). CRC needs 10–15 but has a heavier individual-case tail. No Tier-2 organ meets our illustrative precision criterion with 25 cases.

Keywords: Conformal risk control, distribution shift, recall guarantees, multi-organ segmentation

1 Introduction

A high mean Dice score cannot show whether every organ-at-risk is recalled well enough for radiotherapy contouring or surgical planning [1,22]. Softmax scores can also be poorly calibrated [23]. Building on conformal prediction [11,12,13], distribution-free risk control sets post-hoc thresholds with finite-sample guarantees on a chosen risk. We study Risk-Controlling Prediction Sets (RCPS) [5] and Conformal Risk Control (CRC) [6]. Applied to each organ’s voxel false-negative rate ($\text{FNR} = 1 - \text{recall}$), these thresholds define a *recall floor*: the guaranteed quantity is the population-mean fraction of ground-truth organ voxels missed, not every patient’s loss.

These guarantees require exchangeable calibration and test cases. Differences in disease, scanner, or imaging protocol can break this assumption, so a research-cohort threshold may fail in clinical deployment. We ask: *how many labelled local cases are needed to re-certify the guarantee, and how much over-segmentation does the threshold produce?*

We study an AMOS-trained [2] nnU-Net on RAOS [3], an extension of WORD [4] that labels organ presence and absence across a graded surgical-shift ladder. We contribute: (1) a per-organ recall-risk formulation and in-distribution positive control; (2) a transfer audit that separates the effects of cohort shift and calibration size; (3) evidence that fixed and global thresholds provide no per-organ guarantee; and (4) a local re-certification analysis that keeps high-probability and expectation-level control separate.

2 Related Work

Distribution-free risk control. RCPS [5] calibrates a threshold so a bounded risk is controlled with high probability via a concentration bound; CRC [6] controls it in expectation; Learn-then-Test [7] handles multiplicity across a grid. Our third rule is the variance-adaptive WSR betting bound [8]. **Conformal under shift.** Covariate-shift weighting [9] and beyond-exchangeability bounds [10] relax the usual assumptions. They do not directly solve our setting because we cannot assume that $P(Y|X)$ is unchanged; correcting only the input distribution $P(X)$ may be insufficient. We instead re-certify on labelled local data, which may already exist where clinics contour these organs for radiotherapy. Prior work has likewise shown that conformal guarantees can degrade by subgroup under shift [19].

Conformal/risk control for segmentation. Recent work builds conformal prediction sets for segmentation masks [16], predicts Dice ranges for quality control [17], and applies CRC to FNR/FDR-style risks [18].

The two closest works. sem-CRC [14] controls an expectation-level semantic risk in CT within one distribution. We instead choose a softmax threshold

for each organ, compare high-probability RCPS with expectation-level CRC, and test both after exchangeability fails. COMPASS [15] uses representation-based importance weights to build conformal intervals for mask-derived quantities such as organ area under covariate shift. Its interval summarizes a fixed mask; our threshold changes the mask to meet a recall-risk target. Its weighting targets $P(X)$ and may not cover changes in $P(Y|X)$. Representation-based weighting could still complement local re-certification, but we do not test that combination.

Segmentation quality control and hallucination. RAOS [3] shows that AMOS-trained networks can fail on shifted clinical scans, and frozen nnU-Net models can fail silently out of distribution [21]. HALOS [20] retrain a model to suppress hallucinated organs after resection. We do not retrain; we certify recall only for organs known to be present and study whether that guarantee transfers. Standard Dice and HD95 do not express this guarantee [24].

3 Method

3.1 Per-organ recall risk

For organ k in a case where that organ is present, we measure the fraction of its ground-truth voxels missed at threshold τ :

$$L_k(\tau) = 1 - \frac{|\{v \in G_k : p_k(v) \geq \tau\}|}{|G_k|}, \quad R_k(\tau) = \mathbb{E}_{\text{case}}[L_k(\tau)], \quad (1)$$

Here $p_k(v)$ is the softmax score for organ k at voxel v ; measuring only inside G_k prevents background dilution. L_k is one case’s loss, whereas R_k is the population mean across cases. Every guarantee below constrains R_k , not each patient’s L_k . Cases, rather than voxels, are the exchangeable units. Lowering τ can only recall more voxels, so $L_k \in [0, 1]$ is non-decreasing in τ . This floor applies only to organs known to be present and does not control absent-organ hallucinations.

Unlike softmax argmax, which assigns one class per voxel, we threshold each organ channel separately. Lower thresholds miss fewer organ voxels but add predicted-positive voxels, so we report both effects.

3.2 Calibrating the recall floor

For each organ, we seek the highest threshold that still certifies a population-mean missed-voxel rate no greater than α ; a higher threshold usually gives fewer false positives. Under RCPS [5], we select the largest λ_k whose $(1 - \delta)$ upper confidence bound (UCB) on risk is at most α : $\lambda_k = \sup\{\tau : \text{UCB}_{1-\delta}(R_k(\tau)) \leq \alpha\}$. Because risk is monotone in τ , acceptable thresholds form a prefix of the grid and need no grid-wide multiplicity correction [5]. The selected threshold satisfies

$$\mathbb{P}_{\text{cal}}(R_k(\lambda_k) \leq \alpha) \geq 1 - \delta, \quad (2)$$

where λ_k depends on the random calibration sample and $\mathbb{P}_{\text{cal}}$ is taken over that sample.

What the guarantee means. Equation (2) says that at least a $1-\delta$ fraction of calibration samples produce a threshold whose population-mean future loss is at most α . It does not promise $L_k \leq \alpha$ for every patient. CRC instead gives $\mathbb{E}[L_k(\lambda_k)] \leq \alpha$ [6], where the expectation includes both the calibration sample and future case and has no δ . CRC may need fewer cases because it controls only this overall expectation. Its guarantee is not interchangeable with high-probability RCPS control.

The three selection rules. We compare two RCPS bounds, HB and WSR, with CRC. Each rule selects the largest grid point λ_k satisfying its condition. Here $\hat{R}$ is mean FNR across calibration cases, L_i is case i 's loss, and n is the number of calibration cases:

$$\begin{aligned} \text{HB [5]} & \quad \min\{e^{-nh_1(\hat{R} \wedge \alpha, \alpha)}, e \mathbb{P}[\text{Bin}(n, \alpha) \leq \lceil n\hat{R} \rceil]\} \leq \delta \\ \text{WSR [8]} & \quad \max_{t \leq n} \prod_{i \leq t} (1 + \eta_i(\alpha - L_i)) \geq 1/\delta \\ \text{CRC [6]} & \quad (n\hat{R} + B)/(n+1) \leq \alpha \end{aligned}$$

HB is the standard RCPS bound: it is tight when observed risk is near zero but does not use observed variance. WSR is variance-adaptive and can be tighter when losses have low variance, as they often do for a recall floor. Here $h_1(a, b) = a \log(a/b) + (1-a) \log((1-a)/(1-b))$ and $B=1$ bounds the loss. The $\wedge$ clamp prevents HB from certifying a threshold once $\hat{R} \geq \alpha$. WSR has one tuning choice, the predictable betting fraction $\eta_i = \sqrt{2 \log(1/\delta) / (n\hat{\sigma}_{i-1}^2)}$. The lagged variance $\hat{\sigma}_{i-1}^2$ makes η_i depend only on $L_1, \dots, L_{i-1}$; we cap it at $1/(1-\alpha)$ so each factor remains non-negative. We tune no other hyperparameter.

Marginal, not simultaneous, control. We calibrate each organ independently. The guarantee is therefore marginal for one organ; it does not control the chance that at least one of 12 organs exceeds α . For HB and WSR, replacing δ with $\delta/12$ gives simultaneous high-probability control by a union bound; we report its extra cost in Section 5.3. CRC has no δ to divide, so simultaneous expectation control needs a different method that we do not study. Finite samples also limit which risks can be certified. With $\delta=0.10$, HB needs about 45 cases for $\alpha=0.05$ and 22 for $\alpha=0.10$, independent of the model; WSR needs 46/23, and CRC needs $\lceil 1/\alpha - 1 \rceil = 19/9$.

3.3 Transfer and bound-aware recalibration

We first freeze the AMOS-calibrated λ_k and measure FNR on RAOS (the transfer audit). We then re-certify each organ on n local RAOS cases, repeating the analysis over random splits and values of n . Because all recalibration data come from one RAOS site, these are optimistic single-site estimates.

A recall floor does not limit over-segmentation. We therefore report whole-volume precision = TP/PP and false-positive ratio $\text{FP}_k = \max(\text{PP} - \text{TP}, 0)/|G_k|$ for present organs, where PP is the predicted-positive count and $\text{TP} = \text{recall} \cdot |G_k|$. Once the sample is large enough, $\tau \rightarrow 0$ can control FNR by predicting almost everything. We instead report the *usable* budget: the smallest n that controls held-out FNR—in at least a $1-\delta$ fraction of draws for HB and WSR,

or in mean for CRC—and gives mean held-out precision ≥ 0.5 . This precision cutoff is illustrative, not clinically validated. In deployment, the appropriate cutoff may differ by organ; we did not evaluate organ-specific cutoffs and leave them to future clinical validation.

How we define the organ groups. At the reference setting (WSR, $\alpha=0.10$, $n=25$), six organs meet both conditions: liver, spleen, both kidneys, stomach, and bladder. We call them **Tier-1**; the other six form **Tier-2**. These labels summarize this experiment and are not clinical categories.

All three methods use the same threshold search, with local RAOS cases replacing AMOS. HB and WSR differ only in the finite-sample bound applied to the same losses; this can change the selected threshold, false-positive cost, and required sample size. CRC also provides a different, weaker guarantee. We therefore report CRC separately and do not treat its case counts as equivalent to HB or WSR counts.

4 Experimental Setup

Model and data. We use a 3D full-resolution nnU-Net [1] trained on AMOS CT [2]. AMOS and RAOS use the same nnU-Net plan and preprocessing, but their predictors differ. AMOS calibration uses out-of-fold validation softmax: each case is predicted by the single fold model that held it out (236 of 243 cases are usable; seven corrupted exports are skipped). This cross-fitted set is not the one fixed predictor assumed by the formal guarantee. The main RAOS analysis instead averages all five fold models. A threshold calibrated on one predictor has no formal guarantee for another, so we also run a matched fold 0 analysis (Section 5.2).

The shifted cohort contains all 163 RAOS [3] cases: Set 1 has 67 oncology cases, Set 2 has 22 partial-excision cases, and Set 3 has 74 full-excision cases. All calibration and risk-control analyses use cached softmax outputs after training; we do not retrain the model.

Shared label space. We study only the 12 organs defined consistently in AMOS and RAOS. We exclude the merged `prostate_uterus` class, which would create spurious failures, and the aorta and IVC, which occur only in AMOS. Sex and contrast are audit strata only. FNR is defined only when an organ is present, so each organ has fewer than 163 eligible cases and resection can reduce that number. From Set 1 to Set 3, for example, right-kidney presence falls from 67 to 63 cases and gallbladder presence from 67 to 34. The recall loss therefore excludes errors caused by organ absence or resection.

Protocol. We use $\alpha \in \{0.05, 0.10\}$ and $\delta = 0.10$; an AMOS calibration/test split of 150/86 cases repeated over 20 seeds; 1001 thresholds in steps of 0.001; 2000 case-level bootstrap samples (seed 2026) for transfer-audit CIs; and 200 random re-certification draws per budget. The illustrative precision criterion is 0.5.

5 Results

5.1 In-distribution positive control

The empirical AMOS positive control passes. Over 20 calibration/test splits of the 236 cases, every organ has realized test $\text{FNR} \leq \alpha$ in at least 95% of splits; the target is $1 - \delta = 0.90$. Mean held-out FNR, pooled across organs and both α values, is 0.028. This split-refit result differs from the in-sample $\hat{R}(\lambda_k) \approx 0.06$ in Table 1. CRC makes no $1 - \delta$ claim, so a lower split-wise rate, such as 0.55 for the stomach, is not a violation. Tier-2 organs (Section 5.3) already select $\tau \approx 0$ on AMOS, showing that their low attainable precision is not caused only by transfer.

Simpler thresholds provide no per-organ control even in distribution. A fixed $\tau=0.5$ violates the floor for 8/12 organs at $\alpha=0.05$ and 6/12 at $\alpha=0.10$. One global threshold violates it for 5–6/12 organs: it is too conservative for easy organs and insufficient for hard ones.

5.2 The floor does not survive transfer

When the frozen AMOS thresholds are applied to RAOS, cohort-mean FNR exceeds α for 6/12 organs at $\alpha=0.05$ and 7/12 at $\alpha=0.10$ (Figure 2 and Table 1). Of the other five organs at $\alpha=0.10$, both adrenals have the vacuous predict-everything threshold $\lambda_k=0.000$. Only bladder, liver, and stomach remain below α with a non-vacuous threshold. This non-exceedance is an observed result, not proof that control transfers.

The change differs by organ. Duodenum FNR rises from 0.063 to 0.297, and right-kidney FNR from 0.064 to 0.143; both pairs have non-overlapping CIs. AMOS FNR is well below α , so the higher RAOS values are not caused by an already loose in-distribution fit.

The predictor mismatch has limited effect here. Using fold 0 alone changes per-organ FNR by 0.002 on average (≤ 0.006) and flips no exceedance verdicts (7/12 and 6/12 in both arms). Fully matched model 0 calibration and deployment also agrees, giving 2/12 against the ensemble’s 2/12 at the same 59-case budget. Matching the predictor does not restore exchangeability.

Calibration size changes the exceedance count. With 59 AMOS calibration cases over 500 draws, typical exceedances fall from 7/12 to 2/12 at $\alpha=0.10$ and from 6/12 to 0/12 at $\alpha=0.05$. This apparent improvement comes from wider margins lowering λ_k and increasing predictions: the liver falls from 0.892 to 0.614 and from 0.525 to 0.013, respectively. At $\alpha=0.05$, 59 cases barely clear the 45-case floor and vacuous thresholds rise from 3/12 to 8/12. At $\alpha=0.10$, the right kidney, duodenum, pancreas, and left kidney still exceed in 72.6%, 71.6%, 23.2%, and 17.6% of draws; each other organ stays below 16%.

5.3 Within our grid, WSR needs fewer cases than HB

At $\alpha=0.10$, HB and WSR need about 22–25 local RAOS cases before any threshold can be certified (Figure 3 and Table 2); this reflects finite-sample limits.

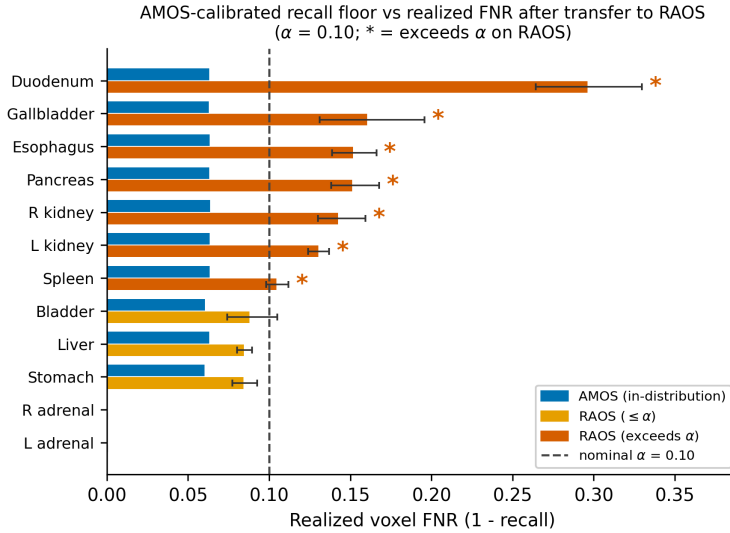


Fig. 2. Transfer audit ($\alpha=0.10$). Per-organ FNR for AMOS in-distribution vs. frozen-threshold transfer to RAOS, with 95% CIs; highlighted organs exceed α on RAOS.

With the illustrative precision criterion, WSR re-certifies all Tier-1 organs with 25 cases and HB with 30–40—a $1.2\text{--}1.6\times$ range within high-probability control. CRC meets its separate expectation-level criterion with 10–15 cases; because this guarantee is weaker, its counts are not equivalent.

At a fixed 25-case budget, WSR gives a usable floor and HB does not. WSR attains precision $0.62\text{--}0.90$ (liver 0.90 , spleen 0.87), whereas HB selects a near-pass-through threshold (liver 0.39 , ≤ 0.01 for the rest). The resulting false-positive cost differs by $4\text{--}1800\times$. That band is organ-dependent and its top end is one organ: right kidney $3.6\times$, liver $133\times$ ($21 \rightarrow 0.16$ FP voxels per true voxel), spleen $1789\times$. The re-certified WSR floor still controls recall: held-out RAOS FNR stays an order of magnitude below α (Table 2) and was $\leq \alpha$ in all 200 draws for all six organs.

Expectation control can hide poor individual cases. Pooling held-out case losses over 200 draws at $n=25$ (Table 2, right), CRC keeps mean risk below α , but up to 17.4% of individual cases exceed α . The 99th percentile is far above the mean, and the worst gallbladder and right-kidney cases approach a total miss. WSR at the same budget exceeds α on at most 1.0% of cases; HB yields a non-degenerate threshold for only two organs. CRC’s smaller case count therefore accompanies a much heavier individual-case tail. Its expectation guarantee is not equivalent to high-probability RCPS control.

Simultaneous control over all 12 organs. Replacing the marginal δ with a Bonferroni $\delta/12$ roughly doubles the usable budget (Table 2), and costs one organ: counting all 12, the number reaching the usability criterion anywhere

Table 1. Frozen AMOS thresholds λ_k and realized voxel FNR at $\alpha=0.10$ on AMOS and the clinically shifted RAOS cohort, with 95% CI. Boldface marks RAOS FNR exceeding α ; both adrenals’ $\lambda_k=0.000$ is vacuous; $n_{\text{AMOS}}/n_{\text{RAOS}}$ are present-organ case counts.

Organ	λ_k	AMOS FNR	RAOS FNR (95% CI)	n_{AMOS}	n_{RAOS}
Bladder	0.277	0.061	0.088 [0.074, 0.105]	233	163
Duodenum	0.024	0.063	0.297 [0.264, 0.330]	236	163
Esophagus	0.050	0.064	0.152 [0.139, 0.166]	235	161
Gallbladder	0.051	0.063	0.161 [0.131, 0.196]	222	123
L adrenal	0.000	0.000	0.000 [0.000, 0.000]	236	163
L kidney	0.690	0.064	0.130 [0.124, 0.137]	235	160
Liver	0.892	0.063	0.085 [0.080, 0.089]	236	163
Pancreas	0.107	0.063	0.152 [0.138, 0.168]	236	163
R adrenal	0.000	0.000	0.000 [0.000, 0.000]	236	156
R kidney	0.741	0.064	0.143 [0.130, 0.159]	234	152
Spleen	0.861	0.064	0.105 [0.098, 0.112]	235	163
Stomach	0.282	0.061	0.085 [0.077, 0.093]	233	163

within the ≤ 100 -case grid falls from 8 to 7 for WSR and from 7 to 6 for HB. Those totals exceed the six Tier-1 organs because pancreas and esophagus also qualify, but only at large budgets. At $\alpha=0.05$ a simultaneous statement is out of reach within the grid: HB certifies none of the 12 and WSR only the liver. Covering every organ at once is therefore materially more expensive than the per-organ budgets suggest.

Tier-2 organs stay out of reach. The six Tier-2 organs (pancreas, duodenum, gallbladder, esophagus, both adrenals) do not reach a usable floor at the 25-case reference budget. Pancreas and esophagus need larger calibration sets (WSR $\sim 40/\sim 80$ cases; CRC $\sim 20/\sim 40$), while gallbladder, duodenum and both adrenals never cross the precision criterion within our grid.

The groups are stable across several precision cutoffs. At the reference point, Tier-1 precision is at least 0.615 and Tier-2 at most 0.0019, so any cutoff in $(0.002, 0.615]$, applied globally or per organ, preserves the groups. Budgets are stable from 0.25 to 0.5; at 0.75, bladder, stomach, and right kidney move to Tier-2.

Which shift axis is associated with the exceedances? Macro-averaged FNR is nearly flat across Sets 1–3 (0.111/0.111/0.122; Figure 4), so exceedances are not concentrated along the surgical gradient. Contrast and sex are likewise weak axes. Missing scanner and protocol descriptors preclude attribution to a specific mechanism.

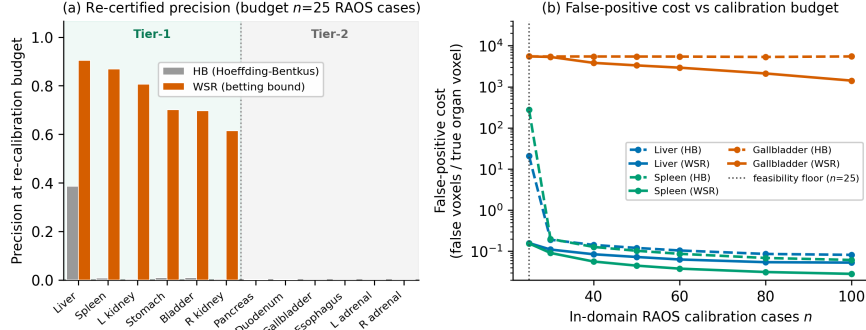
Re-certifying the per-organ recall floor on RAOS: usable budget by concentration bound ($\alpha = 0.10$)


Fig. 3. Re-certification analysis ($\alpha=0.10$). (a) Precision at a fixed 25-case budget, HB vs. WSR. (b) False-positive cost vs. in-domain calibration cases.

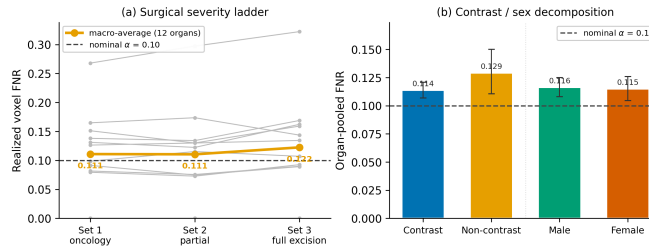
 Realized FNR on RAOS by surgical-shift set, contrast and sex ($\alpha = 0.10$)


Fig. 4. (a) Surgical-shift ladder is nearly flat. (b) Contrast/sex decomposition with α and 95% CIs.

6 Discussion

What this adds beyond a shift warning. An audit must report calibration size, vacuous thresholds, the deployed predictor, guarantee type, and false-positive cost. Here the RCPS bound changed the observed local sample size needed for a useful threshold. Our illustrative results motivate prospective testing of a Tier-1 workflow with ~ 25 local cases and manual review for Tier-2 organs; they do not validate a clinical referral rule.

Limitations. We study one model and one source-target pair; generalization requires other architectures, modalities, and cohorts. We do not compare re-certification with retraining, domain adaptation, or other uncertainty methods. The main calibration and deployment predictors differ, and the matched check covers one fold of one model. Recalibration uses one site, requires labels of unmeasured cost, and yields budget estimates without confidence intervals.

Table 2. Within high-probability control, WSR needs fewer cases than HB.

Left: usable local RAOS budgets for the six Tier-1 organs; ranges span organs. A usable threshold controls held-out FNR and has precision ≥ 0.5 ; “simult.” uses Bonferroni $\delta/12$ over all 12 at $\alpha=0.10$. HB and WSR provide high-probability control; CRC provides expectation control, so their counts are not equivalent. *Right:* held-out case loss at $n=25$, $\alpha=0.10$. Each column reports the worst statistic among organs with a non-vacuous threshold (CRC 9, WSR 6, HB 2); columns can refer to different organs. At $\alpha=0.05$, HB never certifies the right kidney within our grid.

Rule	Usable budget n			Held-out loss, column-wise worst				
	$\alpha=.10$	$\alpha=.05$	simult.	mean	p_{95}	p_{99}	max	$\%>\alpha$
<i>High-probability control: $\mathbb{P}(R_k \leq \alpha) \geq 1 - \delta$</i>								
WSR	25	50–80	50–60	.012	.043	.092	.64	1.0
HB	30–40	60–100	60–100	.002	.010	.020	.29	0.0
<i>Expectation control: $\mathbb{E}[L_k] \leq \alpha$</i>								
CRC	10–15	25	—	.066	.175	.922	1.00	17.4

The recall floor is not a full safety case: control is marginal per organ (and expectation-level for CRC), covers only known-present organs, and excludes hallucination and resection error. Independently thresholded masks may overlap, so whole-volume precision does not establish label-map usability. A per-organ ROI could reduce false-positive counts; its effect on the Tier-1/Tier-2 grouping remains untested.

7 Conclusion

AMOS-calibrated per-organ recall floors did not reliably transfer to RAOS. WSR met our high-probability criterion with fewer local cases than HB; CRC needed fewer under its weaker expectation guarantee but had a heavier individual-case tail. Tier-2 organs required larger sets or never met the illustrative precision criterion. The practical insight is that, once shift broke the guarantee, the choice of concentration bound—not the frozen model—set the local annotation budget needed to restore it in our audit, making bound selection a reportable design decision in deployment-time re-certification.

Acknowledgments. This work was funded by the Deutsche Forschungsgemeinschaft (DFG), project number MA 6791/1-1.

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

References

1. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**(2), 203–211 (2021)

2. Ji, Y., Bai, H., Ge, C., Yang, J., et al.: AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In: *NeurIPS Datasets and Benchmarks* (2022)
3. Luo, X., Li, Z., Zhang, S., Liao, W., Wang, G.: Rethinking abdominal organ segmentation (RAOS) in the clinical scenario: a robustness evaluation benchmark with challenging cases. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. LNCS, vol. 15009, pp. 531–541. Springer Nature Switzerland (2024)
4. Luo, X., Liao, W., Xiao, J., Chen, J., Song, T., Zhang, X., Li, K., Metaxas, D.N., Wang, G., Zhang, S.: WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from CT image. *Med. Image Anal.* **82**, 102642 (2022)
5. Bates, S., Angelopoulos, A., Lei, L., Malik, J., Jordan, M.I.: Distribution-free, risk-controlling prediction sets. *J. ACM* **68**(6), 1–34 (2021)
6. Angelopoulos, A.N., Bates, S., Fisch, A., Lei, L., Schuster, T.: Conformal risk control. In: *International Conference on Learning Representations (ICLR)* (2024)
7. Angelopoulos, A.N., Bates, S., Candès, E.J., Jordan, M.I., Lei, L.: Learn then test: calibrating predictive algorithms to achieve risk control. *arXiv:2110.01052* (2021)
8. Waudby-Smith, I., Ramdas, A.: Estimating means of bounded random variables by betting. *J. R. Stat. Soc. B* **86**(1), 1–27 (2024)
9. Tibshirani, R.J., Foygel Barber, R., Candès, E.J., Ramdas, A.: Conformal prediction under covariate shift. In: *NeurIPS*, pp. 2530–2540 (2019)
10. Barber, R.F., Candès, E.J., Ramdas, A., Tibshirani, R.J.: Conformal prediction beyond exchangeability. *Ann. Stat.* **51**(2), 816–845 (2023)
11. Vovk, V., Gammerman, A., Shafer, G.: *Algorithmic Learning in a Random World*. Springer (2005)
12. Lei, J., G’Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L.: Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.* **113**(523), 1094–1111 (2018)
13. Angelopoulos, A.N., Bates, S.: Conformal prediction: a gentle introduction. *Found. Trends Mach. Learn.* **16**(4), 494–591 (2023)
14. Teneggi, J., Stayman, J.W., Sulam, J.: Conformal risk control for semantic uncertainty quantification in computed tomography. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. LNCS, vol. 15973, pp. 45–55. Springer (2025)
15. Cheung, M.Y., Veeraraghavan, A., Balakrishnan, G.: COMPASS: Robust feature conformal prediction for medical segmentation metrics. In: *International Conference on Learning Representations (ICLR)* (2026)
16. Mossina, L., Friedrich, C.: Conformal prediction for image segmentation using morphological prediction sets. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. LNCS, vol. 15963, pp. 78–88. Springer (2025)
17. Wundram, A.M., Fischer, P., Muehlebach, M., Koch, L.M., Baumgartner, C.F.: Conformal performance range prediction for segmentation output quality control. In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (UNSURE)*, MICCAI Workshop. Springer (2024)
18. Dai, M., Luo, W., Li, T.: Statistical management of the false discovery rate in medical instance segmentation based on conformal risk control. *arXiv:2504.04482* (2025)
19. Mehrtens, H.A., Bucher, T., Brinker, T.J.: Pitfalls of conformal predictions for medical image classification. In: *UNSURE 2023, MICCAI Workshop*. LNCS, vol. 14291, pp. 198–207. Springer (2023)

20. Rickmann, A.-M., Xu, M., Wolf, T.N., Kovalenko, O., Wachinger, C.: HALOS: hallucination-free organ segmentation after organ resection surgery. In: Information Processing in Medical Imaging (IPMI). LNCS, vol. 13939, pp. 667–678. Springer (2023)
21. Gonzalez, C., Gotkowski, K., Bucher, A., Fischbach, R., Kaltenborn, I., Mukhopadhyay, A.: Detecting when pre-trained nnU-Net models fail silently for Covid-19 lung lesion segmentation. In: MICCAI. LNCS, vol. 12907, pp. 304–314. Springer (2021)
22. Wasserthal, J., Breit, H.-C., Meyer, M.T., et al.: TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiol. Artif. Intell.* **5**(5), e230024 (2023)
23. Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Trans. Med. Imaging* **39**(12), 3868–3878 (2020)
24. Maier-Hein, L., Reinke, A., Godau, P., et al.: Metrics reloaded: recommendations for image analysis validation. *Nat. Methods* **21**(2), 195–212 (2024)