This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material

Complementary Reliability Axes for Aortic CTA Segmentation: An Empirical Audit

Thomas Olive

Reproducibility. All volumes are reoriented to RAS canonical and resampled by nnU-Net's default planner; external masks are converted to NIfTI at ingestion. Ground-truth Dice is the 23-class mean on AortaSeg24 and a binary aorta(+iliac) Dice on AVT/AMOS (AMOS aorta is label index 8). The 23-class prediction is collapsed to binary for AVT/AMOS evaluation and to trunk/left-iliac/right-iliac for TotalSegmentator. Pairwise Dice uses T=20 `Dropout3d`(p=0.1) passes; for the largest volumes the prediction accumulator is kept on CPU to stay within 24 GB VRAM. The largest-connected-component restriction is applied only for centerline extraction; Dice and MLD use the raw argmax.

Per-voxel variance and MC-sample count. Per-voxel softmax-variance aggregation of the MC-Dropout and TTA samples gives case-level Unc-AUROC ≈ 0.50 across datasets, which motivated the pairwise-argmax aggregator. The aggregator is robust to the sample budget (Table S1). T=10 recovers the T=20 Spearman ρ to within 3 % on every cohort, and failure-detection AUROC is flat from T=5. Only the in-distribution ρ gains materially from more passes (0.67 at T=5 to 0.74 at T=20), so T=20 is conservative. For completeness, kNN OOD-AUROC is 0.90 [0.83, 0.95] (AMOS) and 0.91 [0.84, 0.96] (TotalSegmentator), and within-AMOS Mahalanobis Unc-AUROC at τ=0.7 is 0.50 [0.43, 0.57].

Table S1. MC-sample-count ablation for the pairwise-Dice signal. Spearman ρ to ground-truth Dice and within-cohort Unc-AUROC (τ=0.5) at $T \in \{5, 10, 20\}$ MC passes, recomputed from the same saved per-pass argmaxes. The signal is largely insensitive to T; the in-distribution cohort has no catastrophic (Dice< 0.5) case, so its Unc-AUROC is undefined.

	Spearman ρ			Unc-AUROC (τ=0.5)		
Cohort (n)	T=5	T=10	T=20	T=5	T=10	T=20
AortaSeg24 ID (40)	0.67	0.72	0.74	—	—	—
AVT (56)	0.47	0.49	0.48	0.99	0.99	0.99
AMOS (300)	0.86	0.87	0.88	0.93	0.94	0.94

Table S2. Clinical measurement agreement. Top: per-case MLD MAE against per-dataset reference MLD (mm), MAE [95 % CI]; AMOS n=289 (11/300 cases had no extractable centerline). Middle: AVT aortic-dissection vs. non-dissection subgroup (Mann-Whitney; p-values uncorrected, and Bonferroni over the three tests keeps the MLD difference significant, p=0.011→0.033). Bottom: TotalSegmentator inter-tool volume agreement (Bland-Altman); the iliac positive bias is consistent with TotalSegmentator's documented iliac under-segmentation (Wasserthal et al., 2023). By comparison, sub-millimetre aortic-diameter agreement (0.3–0.8 mm) was reported only at Dice 0.95 (Sieren et al., 2021). Full citations are in the main paper.

MLD accuracy (mm)				
Dataset (n)	MAE [95 % CI]	Bias	SD	Degenerate %
AortaSeg24 test (40)	1.18 [1.00, 1.34]	+1.18	0.55	0.0
AVT (56)	3.16 [2.70, 3.72]	+2.85	2.42	7.1
AMOS (289)	3.83 [3.45, 4.23]	+0.11	5.15	24.0
AVT subgroup (Mann-Whitney, AD vs. non-AD)				
	Dice: p=0.22	Mahalanobis: p=0.56	MLD diff.: p=0.011	
TotalSegmentator inter-tool volume agreement (566 cases)				
Structure	Bias (mL)	SD (mL)	LoA (mL)	Mean (mL)
Aorta trunk	−14.99	63.25	[−139, +109]	194
R iliac tree	+4.27	12.06	[−19, +28]	17
L iliac tree	+2.86	11.66	[−20, +26]	16

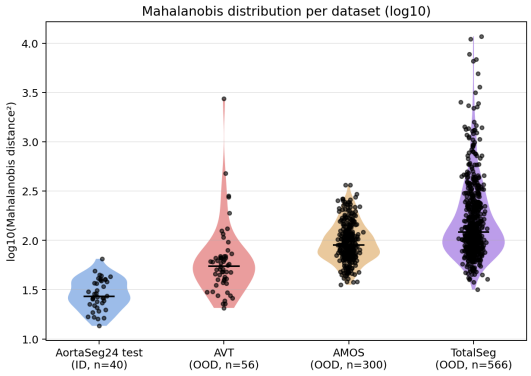

Fig. S1. Distribution of squared Mahalanobis distance per dataset (violin, $\log_{10}$ scale). Per-dataset medians: AortaSeg24 ID 27, AVT 55, AMOS 90, TotalSegmentator 121. The AMOS/TotalSegmentator separations are partly coarse modality and acquisition differences (see the main paper, Discussion).