When Voxels Are Not Equal: Spacing-Driven Metric Bias and Metric Uncertainty in Segmentation

Khadijah Mohammed and Irina Voiculescu

Oxford University Department of Computer Science
khadija.mohammed@cs.ox.ac.uk

Abstract. Medical images are acquired at different spatial resolutions, but segmentation models and evaluation metrics do not always account for these differences. We investigate how variation in CT slice thickness affects both the evaluation of segmentation boundaries and the uncertainty produced by a segmentation model. Using the KiTS23 dataset, we show that a boundary metric defined in voxel space can give systematically different results across different slice thicknesses, whereas defining the same neighbourhood in physical space removes this dependence more consistently. We also find that reported Dice scores depend on how results are aggregated across cases. Finally, model uncertainty at through-plane boundaries increases with slice thickness, although its relationship to annotation uncertainty remains to be established. These findings highlight several ways in which voxel size can influence both the measurement and interpretation of segmentation performance.

Keywords: Segmentation · evaluation · Slice thickness · boundary metrics · uncertainty · calibration.

1 Introduction

Deep learning models for 3D medical image segmentation are typically evaluated using overlap-based metrics such as the Dice similarity coefficient, computed per voxel and averaged across a test set. In computed tomography (CT), slice thickness varies substantially within and across clinical datasets, commonly ranging from under 1 mm to 5 mm or more [3]. This variation raises several questions about segmentation evaluation and uncertainty, which we investigate using KiTS23 (slice thickness 0.5-5.0 mm, 283 cases) with a trained nnU-Net model [6].

Boundary metric thickness dependence. The Symmetric Boundary Dice (SBD) [13] defines boundary agreement within a local neighbourhood whose size is specified in voxel coordinates. In the experiments considered here, we use a Moore neighbourhood of radius one voxel in each direction, giving a 3×3×3 neighbourhood; we refer to this implementation as Cubic SBD, or $\mathrm{SBD}_{\mathrm{cub}}$. For 5 mm thick slices this neighbourhood covers 15 mm in z, corresponding to a much

larger physical tolerance than for fine slices. We therefore evaluate a physical-space variant of SBD, in which the neighbourhood radius is specified in millimetres rather than voxels, giving a consistent physical boundary tolerance across different voxel sizes. We refer to this evaluation effect as *spacing-driven*: slice thickness is represented by the image z-spacing, which determines the physical extent of a fixed voxel-space neighbourhood.

Aggregation convention. We additionally observe that the choice of Dice aggregation convention shifts reported scores by up to 0.250 points and reverses sign between structures, a further evaluation confound not consistently reported alongside published results [10] (Section 4.3).

Entropy and slice thickness. A case acquired at 5 mm slice thickness can only be annotated to 5 mm precision in the through-plane direction. We ask whether z-boundary entropy is associated with slice thickness (Section 4.4).

Case-level association. We ask whether the per-case change in SBD introduced by the physical-space neighbourhood (ΔSBD) is associated with z-boundary entropy (Section 4.5).

We make four contributions: (1) we show that replacing a voxel-space SBD neighbourhood with a physical-space variant removes the observed thick–thin difference, robustly across several physical radii; (2) we show that aggregation convention materially changes reported Dice scores, with the magnitude and direction of the difference depending on structure; (3) we show that z-boundary entropy increases with slice thickness; (4) we find little evidence of a case-level association between ΔSBD and z-boundary entropy, despite both being associated with slice thickness.

2 Related Work

Common overlap metrics can be biased by foreground/background ratios [11], boundary metrics by large components [7], and Hausdorff Distance by asymmetric normalisation [2]. The INSTANCE challenge addressed anisotropic segmentation using Normalised Surface Dice [5], and KiTS19 prior work resampled to uniform 3 mm before training [3]. Wasserthal et al. showed Dice and NSD respond differently to spacing across two fixed-resolution models [12]; we test whether the same pattern emerges within a single model on a naturally heterogeneous dataset. Predictive entropy correlates inversely with Dice quality [9] and networks are overconfident at boundaries [8]; single-pass entropy correlates substantially less well with inter-rater disagreement than purpose-trained methods [1], motivating our regression test.

3 Methods

3.1 Dataset and model

We use the KiTS23 challenge dataset [4], with contrast-enhanced abdominal CT scans annotated for kidney, tumour, and cyst. In-plane spacing is approximately

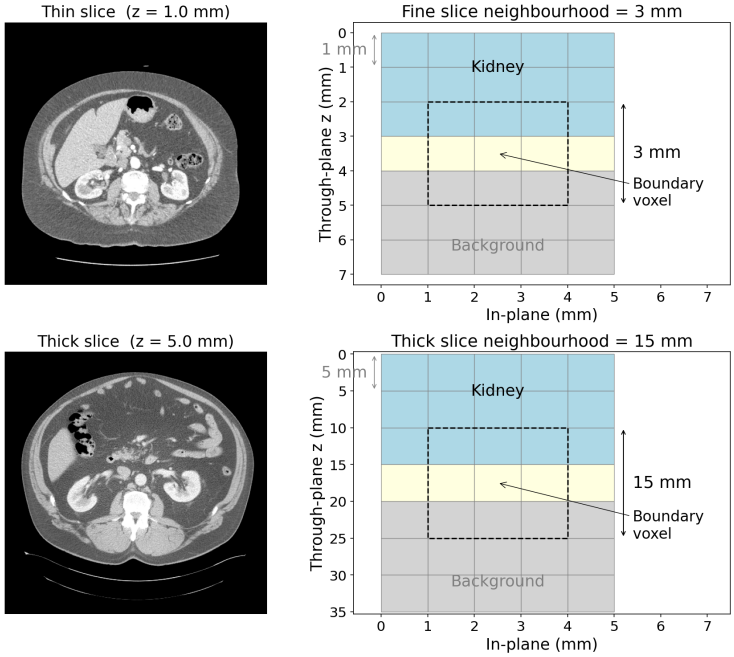

Fig. 1. Top row: thin-slice case (z=0.5 mm). Bottom row: thick-slice case (z=5.0 mm). Left: axial CT slices at 512×512 resolution with in-plane pixel size 0.703 mm. Right: schematic of 7 slices in the inferior-superior direction for fine (z=1 mm) and thick (z=5 mm). A 3×3×3 voxel neighbourhood covers 3 mm in z for fine slices but 15 mm for thick slices, forgiving boundary errors that would be large in physical space. $\mathrm{SBD}_{\mathrm{phys}}$ fixes the neighbourhood radius at 1.5 mm regardless of voxel size.

0.789 mm (dataset median). Slice thickness varies 0.5–5.0 mm. We report two evaluations: (i) 74 held-out cases (fast model, fold 0) for the three-scheme Dice analysis; (ii) 283 properly-split cross-validation cases across four folds for the SBD and entropy analyses, with softmax probabilities saved for entropy.

3.2 Stratification by slice thickness

Cases are grouped into three categories according to their native slice thickness: thin (t_z<2 mm), medium (2≤t_z<3 mm), and thick (t_z≥3 mm). In the 283-case multi-fold set: thin n=57, medium n=34, thick n=192. In the 74-case test set: thin n=24, medium n=11, thick n=39.

3.3 $\mathbf{SBD_{cub}}$ and $\mathbf{SBD_{phys}}$

For $\mathrm{SBD}_{\mathrm{cub}}$, we use a Moore neighbourhood with radius one voxel in each direction, giving a 3×3×3 neighbourhood in voxel coordinates.

For 5 mm thick slices this covers 15 mm in the through-plane direction (z).

For $\mathrm{SBD}_{\mathrm{phys}}$ the neighbourhood is specified by a physical radius r. For the principal experiments, we use r=1.5 mm. At the median in-plane spacing of 0.789 mm, this gives a radius of one voxel, matching $\mathrm{SBD}_{\mathrm{cub}}$ in the xy-plane. In the z direction, for a thickness t_z, we set $r_z = \lfloor r/t_z \rfloor$. Thus, given r=1.5 mm, we get r_z=1 when t_z=0.789 mm, and r_z=0 when t_z=5 mm.

Figure 1 illustrates the resulting difference between voxel- and physical-space neighbourhoods. We assess sensitivity to this choice, with results reported in Section 4.2.

More generally, for a given dataset, the physical radius r is chosen such that $\lfloor \frac{r}{s_{xy}^{\mathrm{med}}} \rfloor$=1, where s_{xy}^{med} represents the median spacing in the xy-plane. Specifically for KiTS, where s_{xy}^{med}=0.789 mm and we want $\lfloor \frac{r}{s_{xy}^{\mathrm{med}}} \rfloor$=1, radius r satisfies $0.789 \leq r < 1.578$ mm. In this range, the choice r=1.5 mm was selected for the principal experiments.

For fine slices (t_z=0.789 mm): $r_z = \lfloor \frac{1.5}{0.789} \rfloor$=1, identical to $\mathrm{SBD}_{\mathrm{cub}}$. For thick slices ($t_z$=5 mm): r_z=0, giving no z-extension beyond the centre voxel and a much tighter tolerance.This makes the boundary tolerance consistent in physical space regardless of scanner settings.

3.4 Three Dice aggregation schemes

Scheme A (pooled, unweighted): sum TP, FP, FN across all cases before computing Dice. Every voxel contributes equally regardless of physical size.
Scheme B (per-case averaged): compute Dice per case then average. Every case contributes equally regardless of anatomy volume. This is the KiTS23 official convention [4].
Scheme C (volume-weighted pooled): weight each voxel's contribution by its physical volume in mm^3 before pooling. Validated against nearest-neighbour resampling to within 5×10^{-4} Dice.

3.5 Boundary entropy, regression, and coupling

We compute per-voxel Shannon entropy $H(x) = -\sum_c p_c(x) \log_2 p_c(x)$ from the model's softmax output (available for 280 of 283 cases), averaged over ground-truth boundary voxels ($3 \times 3 \times 3$ erosion) and separately over the top and bottom axial slices only (*z-boundary*). We fit Spearman correlations and linear regressions of z-boundary entropy on slice thickness, comparing the observed slope against a partial-volume model baseline that predicts a negative relationship. We define the SBD difference as $\Delta\mathrm{SBD} = \mathrm{SBD}_{\mathrm{cub}} - \mathrm{SBD}_{\mathrm{phys}}$, so that positive values indicate higher scores under the cubic voxel-space neighbourhood. We additionally compute the Spearman correlation between per-case metric inflation (ΔSBD) and z-boundary entropy to test whether the two spacing-driven effects co-occur on the same cases.

Table 1. Overall segmentation performance of the trained nnU-Net model (Scheme B, per-case mean, n=74 held-out cases, fold 0). $\mathrm{SBD}_{\mathrm{phys}}$ computed at r=1.5 mm.

Class	**Dice**	**HD95 (mm)**	$\mathrm{SBD}_{\mathrm{cub}}$	$\mathrm{SBD}_{\mathrm{phys}}$
Kidney	0.905	12.1	0.830	0.798
Tumour	0.666	39.0	0.567	0.525
Cyst	0.236	56.9	0.218	0.191

Table 2. Kidney SBD by slice thickness group: cubic vs. physical neighbourhood at r=1.5 mm (283 cases, 4 folds). $\Delta\mathrm{SBD} = \mathrm{SBD}_{\mathrm{cub}} - \mathrm{SBD}_{\mathrm{phys}}$.

Group	n	$\mathrm{SBD}_{\mathrm{cub}}$	$\mathrm{SBD}_{\mathrm{phys}}$	ΔSBD
Thin (t_z<2 mm)	57	0.930	0.943	−0.013
Medium (2≤t_z<3 mm)	34	0.949	0.924	+0.025
Thick (t_z≥3 mm)	192	0.961	0.932	+0.029

4 Results

This section reports both descriptive observations obtained directly from the computed metrics and results supported by statistical analysis. We first describe the observed effects of slice thickness and aggregation convention (Sections 4.1–4.3), before testing the entropy–thickness relationship and case-level associations statistically (Sections 4.4–4.5).

4.1 KiTS segmentation analysis

Unless stated otherwise, all $\mathrm{SBD}_{\mathrm{phys}}$ values in this paper are computed for physical radius r=1.5 mm. Section 4.2 reports that the sign reversal persists across $r \in \{1.0, 1.5, 2.0, 2.5\}$ mm, confirming the result is not sensitive to this choice.

Table 1 reports overall performance (Scheme B, per-case mean) across all three structures for the 74-case test set.

4.2 $\mathrm{SBD}_{\mathrm{phys}}$ corrects spacing-driven boundary metric inflation

Table 2 shows $\mathrm{SBD}_{\mathrm{cub}}$ and $\mathrm{SBD}_{\mathrm{phys}}$ by thickness group. $\mathrm{SBD}_{\mathrm{cub}}$ is 0.031 higher for thick than for thin slices; $\mathrm{SBD}_{\mathrm{phys}}$ reverses this to −0.011 (thick $\mathrm{SBD}_{\mathrm{phys}}$ 0.932 minus thin $\mathrm{SBD}_{\mathrm{phys}}$ 0.943). This sign reversal is a key result.

In order to confirm robustness, $\mathrm{SBD}_{\mathrm{phys}}$ was also evaluated for other radii $r \in \{1.0, 1.5, 2.0, 2.5\}$ mm. The sign reversal persisted for all values: the difference in $\mathrm{SBD}_{\mathrm{phys}}$ between the thick and thin groups was −0.011, −0.012, −0.015, −0.014 respectively. The sign reversal observed at the principal radius r=1.5 mm is robust to the choice of radius over the range tested.

Table 3. Three Dice aggregation schemes by structure (74-case test set).

Structure	Scheme A (pooled)	Scheme B (per-case)	Scheme C (vol-weighted)	A−B
Kidney	0.922	0.905 ± 0.182	0.920	+0.017
Tumour	0.548	0.666 ± 0.284	0.637	−0.117
Cyst	0.485	0.236 ± 0.315	0.415	+0.249

Table 4. Boundary predictive entropy by slice thickness group, kidney class (280 cases).

Group	n	Z-boundary entropy	Full boundary entropy
Thin	55	0.0027	0.0235
Medium	34	0.0057	0.0210
Thick	191	0.0164	0.0257

4.3 Dice aggregation scheme materially changes the reported score

Table 3 reports all three schemes. For kidney, Schemes A and B differ by 0.017, whereas the differences are substantially larger for tumour and cyst. For cyst, Scheme A gives 0.485 compared with 0.236 under Scheme B, a difference of 0.249 Dice points. For tumour, the direction of the difference is reversed: Scheme A gives 0.548 compared with 0.666 under Scheme B. Thus, both the magnitude and direction of the difference between aggregation schemes depend on the structure being evaluated.

4.4 Z-boundary entropy increases with slice thickness beyond what information loss predicts

Table 4 reports z-boundary entropy by slice thickness group. We examine the relationship between slice thickness and boundary entropy through four types of analysis.

These four analyses address the relationship between entropy and slice thickness relationship at increasing levels of specificity. Analysis 1 tests whether thickness groups differ at all. Analysis 2 tests whether the specific thin-thick comparison is significant. Analysis 3 tests whether the relationship is monotonic across all 280 cases continuously. Analysis 4 quantifies the magnitude of the relationship.

Analysis 1. A Kruskal–Wallis test assesses whether z-boundary entropy differs across the thin, medium, and thick slice thickness groups. It ranks all 280 entropy values and checks whether the three groups cluster differently within that ranking. This yields H=14.0, p<0.001, providing evidence of differences between the groups.

Analysis 2. A pre-specified Mann–Whitney U test compares thin versus thick groups directly, testing whether cases with larger slice thickness tend to rank higher entropy. This yields p<0.001, with rank-biserial correlation r_{rb}=0.30.

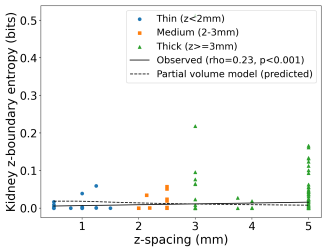

Fig. 2. Z-boundary entropy as a function of native slice thickness (kidney class, 280 cases). Solid line: observed regression (Spearman rho=0.233, p<0.001). Dashed line: partial volume model prediction, which expects entropy to decrease with thickness. The observed positive slope is in the opposite direction, suggesting the model tracks annotation imprecision beyond information loss alone.

Analysis 3. A Spearman rank correlation of z-boundary entropy on slice thickness tests whether there is a monotonic relationship between slice thickness and entropy across all 280 cases continuously, rather than just comparing groups. This yields ρ=0.233, p<0.001.

Analysis 4. A linear regression of entropy on slice thickness gives a slope of 2.3×10^{-3} bits per mm. To quantify uncertainty in the estimated slope, bootstrapping yields a 95% confidence interval of $[5.8\times10^{-4}, 4.0\times10^{-3}]$. As the interval excludes zero, the estimated association is positive.

All analyses used `scipy.stats`: `kruskal` for Analysis 1, `mannwhitneyu` (alternative='greater') for Analysis 2, `spearmanr` for Analysis 3, and `linregress` with 1000 bootstrap resamples for Analysis 4.

This association is also observed for tumour (ρ=0.152, p=0.011) and cyst (ρ=0.182, p=0.031), suggesting the relationship is not specific to the kidney.

Our simple partial-volume model predicts the opposite relationship. As slice thickness increases, fewer voxels at the boundary contain mixed tissue, and boundary entropy should therefore decrease. We find that z-boundary entropy increases significantly with slice thickness (Spearman ρ=0.233, p<0.001). This pattern is consistent with increased annotation imprecision as through-plane resolution becomes coarser, although other explanations cannot be excluded.

4.5 Metric inflation and entropy show little case-level association

Despite both being associated with slice thickness, we find little evidence of a case-level association between metric inflation and the uncertainty signal. Table 5 tests whether cases with larger metric inflation also tend to have higher z-boundary entropy. Row 1 tests whether ΔSBD is associated with z-boundary entropy across all 280 cases. Row 2 repeats this within the 191 thick cases only, where both effects are most pronounced and any association would be most detectable.

Rows 3–4 test a different but related question: whether the raw $\mathrm{SBD}_{\mathrm{cub}}$ and $\mathrm{SBD}_{\mathrm{phys}}$ scores are individually associated with entropy within thick cases. If

Table 5. Metric–entropy coupling. Spearman correlations are reported with bootstrap 95% confidence intervals.

Analysis	n	ρ	p	**95% CI**
ΔSBD vs. z-boundary entropy (all 280 cases)	280	+0.083	0.166	[−0.117, +0.117]
ΔSBD vs. z-boundary entropy (thick only)	191	−0.097	0.184	[−0.144, +0.141]
$\mathrm{SBD}_{\mathrm{cub}}$ vs. z-boundary entropy (thick only)	191	+0.043	0.554	[−0.143, +0.143]
$\mathrm{SBD}_{\mathrm{phys}}$ vs. z-boundary entropy (thick only)	191	+0.068	0.352	[−0.136, +0.144]

either raw score showed a strong association with entropy, this would suggest the result in Row 1 was specific to the choice of ΔSBD as the metric variable.

All correlations are small and no significant association was detected ($p>0.16$ for all analyses). The bootstrap 95% confidence intervals all include zero.

5 Discussion

5.1 $\mathrm{SBD}_{\mathrm{phys}}$ corrects a systematic, deterministic metric bias

The $\mathrm{SBD}_{\mathrm{cub}}$ inflation arises from a well-defined geometric mechanism: for 5 mm slices, the 3×3×3 voxel neighbourhood covers 15 mm in z, forgiving boundary errors that would be large in physical space. $\mathrm{SBD}_{\mathrm{phys}}$ removes this by fixing the tolerance in physical space. The sign reversal ($\mathrm{SBD}_{\mathrm{cub}}= +0.031$; $\mathrm{SBD}_{\mathrm{phys}}= -0.011$) confirms that $\mathrm{SBD}_{\mathrm{phys}}$ is not simply a scaled version of $\mathrm{SBD}_{\mathrm{cub}}$ but corrects a systematic bias in the wrong direction. Figure 3 shows the case-level relationship between metric inflation and z-boundary entropy. The practical implication is that boundary-metric results on datasets with variable thickness depend on how the matching neighbourhood is defined. In our experiments, changing from a voxel-space to a physical-space neighbourhood reverses the direction of the thick–thin difference, from +0.031 to −0.011.

5.2 Predictive entropy is consistent with annotation imprecision beyond information loss

The z-boundary entropy finding is the most theoretically interesting result. A simple partial volume model predicts that entropy should decrease with slice thickness. Thicker voxels have fewer mixed-tissue boundary voxels per unit volume, reducing ambiguity. The observed opposite direction (Spearman ρ=0.233, $p<0.001$; see Section 4.4) is consistent with the hypothesis that the model has learned, during training on heterogeneous data, to produce higher uncertainty at through-plane boundaries where annotation precision is limited by slice thickness. However, alternative explanations cannot be excluded from this analysis alone.

If this interpretation holds, it would be an encouraging calibration property: the model would be more uncertain where ground truth is less reliable.

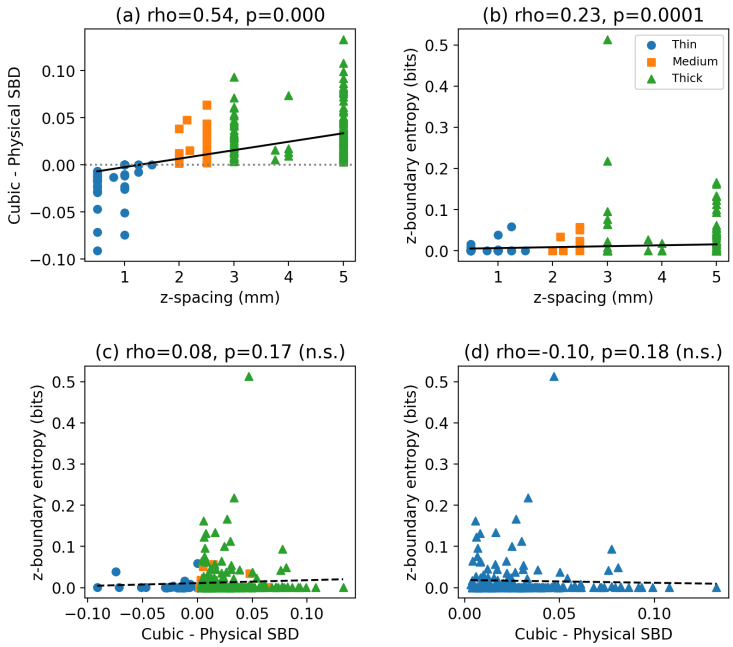

Fig. 3. Metric inflation (ΔSBD) and z-boundary entropy are both associated with slice thickness (panels a, b). Panels (c) and (d) show no significant case-level association between ΔSBD and entropy, across all 280 cases (c)and within the 191 thick cases only (d).

5.3 Practical implications

Both metric inflation and z-boundary entropy are associated with thicker slices, but we find little evidence of a case-level association between them. SBD_{phys} addresses the metric effect by defining boundary tolerance in physical space. The entropy result may provide complementary information about uncertainty at through-plane boundaries, although its relationship to annotation reliability remains to be established. Purpose-trained methods such as EDUE [1] provide a means of investigating this relationship directly against inter-rater disagreement.

5.4 Aggregation as an underexamined source of variance

The 0.249-point gap between Schemes A and B for cyst, and the sign reversal for tumour, show that aggregation convention is not a neutral choice. The KiTS23 challenge documents its use of Scheme B [4]. Differences in aggregation convention can therefore complicate cross-paper comparisons, unless the convention is reported explicitly.

6 Limitations and Future Work

This study uses a single architecture, CT dataset, annotation source, and set of model predictions. The generalisability of the entropy–slice thickness relationship should therefore be evaluated across additional datasets and architectures, including MRI, where differences in acquisition and image formation may introduce distinct effects. Its reproducibility across independent training runs and across different slice thicknesses also remains to be established. Future work should additionally compare purpose-trained uncertainty methods [1] with single-pass entropy.

7 Conclusion

Variable slice thickness can affect segmentation evaluation because a fixed voxel-space boundary neighbourhood represents different physical tolerances at different voxel sizes. Replacing the voxel-space $\mathrm{SBD}_{\mathrm{cub}}$ neighbourhood with $\mathrm{SBD}_{\mathrm{phys}}$ reverses the observed thick–thin difference, an effect that is robust across the physical radii tested. We also find that Dice aggregation convention can substantially alter reported scores, with the magnitude and direction of the differences varying between structures.

Contrary to the prediction of our simple partial-volume model, z-boundary entropy increases significantly with slice thickness. Increased annotation imprecision at coarser through-plane resolution is one possible explanation, although the present analysis does not establish this mechanism. Despite both metric inflation and boundary entropy being associated with slice thickness, we find little evidence of an association between them at the case level. Together, these findings demonstrate the importance of accounting explicitly for voxel size and evaluation convention when assessing segmentation performance and uncertainty on heterogeneous CT data.

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

References

1. Abutalip, K., et al.: LEXU: Learning from expert disagreement for single-pass uncertainty estimation in medical image segmentation. In: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 112–122. Springer, Cham (2025). https://doi.org/10.1007/978-3-032-06593-3_11
2. Aydin, O., Taha, A., Hilbert, A., et al.: On the usage of average Hausdorff distance for segmentation performance assessment: hidden error when used for ranking. European Radiology Experimental **5**, 4 (2021). https://doi.org/10.1186/s41747-020-00200-2
3. Heller, N., Isensee, F., Maier-Hein, K., et al.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge. Medical Image Analysis **67**, 101821 (2021). https://doi.org/10.1016/j.media.2020.101821

4. Heller, N., et al.: The 2023 kidney tumor segmentation challenge (KiTS23). https://kits-challenge.org/kits23/ (2023)
5. INSTANCE challenge: 3D anisotropic intracranial hemorrhage segmentation on non-contrast head CT. Grand Challenge. https://instance.grand-challenge.org/
6. Isensee, F., Jaeger, P., Kohl, S., Petersen, J., Maier-Hein, K.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**, 203–211 (2021)
7. Jaus, A., et al.: Every component counts: rethinking the measure of success for medical semantic segmentation in multi-instance segmentation tasks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3904–3912 (2025). https://doi.org/10.1609/aaai.v39i4.32408
8. Liang, W., Zhang, W., Yue, L., Xu, M., Maennel, O., Chen, W.: Calibrating on medical segmentation model through signed distance. In: Proceedings of the 34th ACM International Conference on Information and Knowledge Management. pp. 1778–1787 (2025). https://doi.org/10.1145/3746252.3761101
9. Mehrtash, A., et al.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Transactions on Medical Imaging **39**(12), 3868–3878 (2020). https://doi.org/10.1109/TMI.2020.3006437
10. Reinke, A., Tizabi, M., Baumgartner, M., et al.: Understanding metric-related pitfalls in image analysis validation. Nature Methods **21**(2), 182–194 (2024). https://doi.org/10.1038/s41592-023-02150-0
11. Taha, A., Hanbury, A.: Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Medical Imaging **15**, 29 (2015). https://doi.org/10.1186/s12880-015-0068-x
12. Wasserthal, J., et al.: TotalSegmentator: robust segmentation of 104 anatomical structures in CT images. Radiology: Artificial Intelligence p. e230034 (2023). https://doi.org/10.1148/ryai.230024
13. Yeghiazaryan, V., Voiculescu, I.: A family of boundary overlap measures for the evaluation of medical image segmentation. Journal of Medical Imaging **5**(1), 015006 (2018)