LEEWAY: Same Anatomy, Same Score, Different Verdict Using Paired Falsification Tests in Dental CBCT

Behnood Mashayekhi[1][0009−0002−3062−4001], Arian Amiramjadi[1], and Parsa Hosseini[1]

Amesha, London, United Kingdom
behnood@amesha.co, arian@amesha.co, parsa@amesha.co
https://amesha.co

Abstract. Medical-image validation can be reproducible yet answer a different question from the intended geometric task. We test three interpretations in a dental-CBCT case study using patient-paired falsification. First, a corrected, patient-disjoint two-corpus laterality experiment with equal training counts and 20 refits shows early target-minus-source gaps of 1.39% and 1.98% shrink to 0.00% and 0.22% by 320 epochs; deterministic physical-rank rules achieve 1.000 accuracy without fitting. Second, an affine-compensated storage translation preserves every foreground world point but produces only 1.12% and 0.93% decision flips; the 95-case cohort has 38/95 complete-grid feasibility, all invariant controls have zero flips, and the prespecified promotion gate fails. Third, we prove that equal TP, FP, FN, predicted volume, and their named derived overlap scores do not identify a thresholded tooth–canal clearance proxy. A frozen actual-prediction construction yields 1,757 exact pairs in 64/92 patients; at 2 mm its patient-equal away-minus-toward action difference is 0.481 (95% CI $[0.402, 0.559]$), while 1,752/1,757 (99.7%) pairs require 3–6-mm shifts. These results reject the reconstructed transfer interpretation and establish grid-bounded metric non-identifiability for one dental cohort and one checkpoint. They do not estimate natural-error prevalence, clinical outcomes, safety, or architecture-general behavior.

Keywords: cone-beam CT · medical image validation · segmentation metrics · transfer evaluation · metamorphic testing

1 Introduction

In geometric uses of segmentation, a mask is often an intermediate object. The downstream question may instead concern a relation, for example, whether a predicted mandibular canal lies within a stated distance of a tooth. An overlap score compares sets, a transfer experiment compares fitting histories, and a clearance rule compares two anatomical objects in physical space. A pipeline may reproduce perfectly while evidence is attached to the wrong one of these objects.

Table 1. Original interpretations, corrected paired designs, and conclusions permitted by the frozen evidence.

Test	Original interpretation	Corrected intervention / control	Invariant; estimand or gate	Result → permitted conclusion
Role and convergence	Early target–source gap implies target difficulty or transfer cost.	Bilateral lower-canal task; equal 76-patient pools; patient-disjoint folds; same target patients; 20 refits; deterministic rank controls.	Target patients and task; patient-equal $\Delta_t(e)$; persistence and whether fitting is needed.	$1.39, 1.98\% \rightarrow 0.00, 0.22\%$; rank $1.000 \rightarrow$ reconstructed transfer interpretation rejected.
Same physical world	A storage-coordinate shortcut explains or rescues transfer.	Joint voxel shift and affine compensation; learned world/pair-centered models and rank controls.	Every world point, count, extent, pair, realized shift; ±20-mm flips; $\geq 5\%$, $\geq 80\%$ feasibility, stability, reversal/second-task gate.	1.12% at 38/95 and 0.93% at 191/191; controls zero → negative control, no promoted mechanism.
Exact overlap	Equal overlap/volume summaries imply the same clearance verdict.	Analytic omission plus actual-prediction half-mask shifts with opposite equal-norm vectors; retain every attempt.	TP, FP, FN, volume, named scores; structural checks; ≥ 20 discordant patients; patient-equal conditional difference.	1,757 rows/64 patients; $\widehat{\theta}_2 = .481 \rightarrow$ named scores do not identify this proxy within the constructed grid.

We encountered this problem in a two-corpus CBCT study whose original claim attributed a target-trained versus source-trained accuracy gap to target-corpus difficulty. Forensic reconstruction found mixed anatomy, resubstitution in one arm, truth-signed confidence, unequal patient roles, and a two-candidate answer encoded by physical left–right rank. Rather than cosmetically repair the manuscript, we ask a stricter question: which interpretation survives tests that hold the claimed scientific object fixed?

We contribute three conclusion-level tests: (1) a patient-disjoint transfer reconstruction with equal training counts, refits, convergence trajectories, and no-training sufficiency controls; (2) a storage-coordinate intervention that verifies equality of every foreground world point; and (3) a formal and empirical exact equivalence class of named confusion-count/volume metrics on which a tooth-conditioned action changes. This is neither a new metric nor a claim that Dice is generally inadequate. The evidence concerns the reconstructed two-corpus laterality task and one frozen dental-CBCT construction.

2 Closest Prior Work and Surviving Scope

Established guidance covers metric selection and failure modes [9,14]. Direction relative to critical anatomy can change downstream dosimetry [13]. Proximity and topology studies evaluate information absent from global overlap [10,15]; controlled-error propagation is also established [4]. Stothers constructs matched-Dice errors with high versus low hazard [16]. We claim neither matched-Dice nor hazard-perturbation priority; our narrower object certifies exact equality of

confusion counts and volume with patient-level accounting and a family seeded from a prediction.

Metamorphic testing [17], segmentation-specific variants [5], medical shortcut interventions [3], and padding/cropping position shortcuts [7] predate this study. Our same-world arm verifies physical-mask equality. Dental CBCT already motivates millimetric canal review and tooth–canal distance [11,8]; PMCanalSeg separates upper from lower canals [6], and we use only the lower labels.

3 Paired Falsification Protocol

Let z_i collect the role, masks, coordinate map, model state, and endpoint for patient i. A falsifier declares an equivalence relation $\sim_E$ over the object that should not alter a conclusion c and records

$$F_i(E) = \mathbb{I}\{c(z_i) \neq c(z_i')\}, \qquad z_i \sim_E z_i'. \tag{1}$$

Each test follows the same frozen sequence: **(1)** state the interpretation; **(2)** construct a paired intervention and control; **(3)** verify the object that must remain invariant; **(4)** compute a patient-level estimand with complete feasibility accounting; and **(5)** apply its prespecified validity or promotion gate. Undefined or clipped interventions are not zero-imputed. Repeated teeth, queries, and masks are averaged within patient.

A *no-training sufficiency control* is a deterministic rule with no fitted parameters; here it answers the two-candidate laterality query by physical-x or pair-centered rank. A *GT-intimate unit* is a patient–tooth unit with $d(G,T) \leq \tau$. The *patient-equal away-minus-toward difference* averages $a_\tau^{\text{away}} - a_\tau^{\text{toward}}$ over retained exact rows within each patient, then weights patients equally. The *33-cell simultaneous band* is one 95% maximum-deviation patient-bootstrap envelope over threshold-by-magnitude cells containing at least five patients; sparse cells receive no band.

3.1 Role correction and convergence

Each case provides two ground-truth mandibular-canal candidates and two laterality queries. A two-layer, 32-unit MLP scores a query concatenated with mask centroid, extent, elongation, and relative volume. This is a mask-derived laterality diagnostic, not an image encoder or clinical navigation model. Five outer folds follow a deterministic patient-level assignment rule. Within each fold both corpus training pools are restricted to 76 patients, and target- and source-trained models are evaluated on the same held-out target patients. We refit 20 seeds at 40, 80, 160, and 320 epochs. For target t the patient-level contrast is

$$\Delta_t(e) = n_t^{-1} \sum_{i \in t} \left[\overline{A}_{i,t\to t}(e) - \overline{A}_{i,s\to t}(e) \right], \tag{2}$$

where bars average refits. Patient-bootstrap intervals are conditional on the 20 fitted models; we also report the full refit range and crossed robustness band. The

no-training rank controls, plus constant and deterministic case-indexed random rules, test whether fitting is necessary.

3.2 Lossless storage-coordinate reparameterization

Let voxel index v map to world position $w = Lv + b$. Along the affine-derived left–right voxel axis e, an integer shift k is paired with

$$v' = v + ke, \qquad L' = L, \qquad b' = b - Lke, \tag{3}$$

so $L'v' + b' = Lv + b$. We test both signs at nominal $\{0, 2, 5, 10, 20\}$ mm. A case is eligible only if all candidates remain inside the array for the complete grid; there is no interpolation or clipping. Every occupied world point, candidate count, extent, pair assignment, axis, and realized shift is independently verified. The primary endpoint is the ±20-mm decision-flip fraction for the 40-epoch absolute-index MLP. Frozen promotion required at least 5% flips, 80% feasibility, stability across refits, and a transfer-sign reversal or nontrivial second task.

3.3 Exact-overlap action non-identifiability

Let Ω be a finite physical voxel lattice with Euclidean metric ρ, and let $S(T)$ be the 6-neighbour surface of tooth mask T. Define $\delta_T(x) = \min_{t \in S(T)} \rho(x, t)$ and, for a nonempty canal prediction P, $d(P, T) = \min_{p \in P} \delta_T(p)$. For a geometric proxy threshold τ, let $a_\tau(P; T) = \mathbb{I}\{d(P, T) > \tau\}$. This is not a validated treatment or injury threshold.

Proposition 1 (Critical-region omission). *Let G be a nonempty canal mask of N voxels and $C_\tau = \{x \in G : \delta_T(x) \le \tau\}$ contain m voxels, $1 \le m < N$. Set $P_{\text{crit}} = G \setminus C_\tau$. Delete instead the m voxels of G farthest from T (fixed lexicographic tie rule) to obtain P_{far}. Then $a_\tau(P_{\text{crit}}; T) = 1$ and $a_\tau(P_{\text{far}}; T) = 0$, while both predictions have $(\text{TP}, \text{FP}, \text{FN}) = (N - m, 0, m)$. Thus Dice, IoU, precision, recall, predicted volume, and symmetric volume similarity are identical; with $\kappa = m/N$, Dice is $2(1 - \kappa)/(2 - \kappa)$.*

All noncritical voxels precede critical voxels in the farthest-first order; an $m < N$ prefix cannot delete all of C_τ. This proves the action contrast, and equal deletion counts give the metric identities. The proposition says nothing about ASSD, HD95, surface Dice, or topology; the full proof is in the supplement.

Two frozen families instantiate the statement. Family A enumerates the analytic deletion on every eligible ground-truth patient–tooth pair and records connectivity and component diagnostics. Family B starts from a real thresholded canal prediction, derives a direction from its nearest surface pair with the ground-truth tooth, and shifts one spatial half by exact opposite integer vectors. Canal truth does not choose the direction; it is used for exact-match retention and truth-relative metric evaluation. Clipping and overlap with the fixed half fail closed. The primary population contains all exact rows, not only GT-intimate

units. For patient i, let $R_{i\tau}$ contain all retained exact rows and $\mathcal{I}_\tau$ the patients with at least one:

$$\widehat{\theta}_\tau = \frac{1}{|\mathcal{I}_\tau|} \sum_{i \in \mathcal{I}_\tau} \frac{1}{|R_{i\tau}|} \sum_{r \in R_{i\tau}} \left[a_\tau(P_{ir}^{\text{away}}; T_{ir}) - a_\tau(P_{ir}^{\text{toward}}; T_{ir}) \right]. \tag{4}$$

The grid $\tau \in \{1, 2, 3\}$ mm and displacements 0.3–6.0 mm in 0.3-mm steps was frozen after the earlier directional-replay route failed but before either exact-family enumeration; it is a new audit benchmark, not a retroactive preregistration of the original study.

4 Data, Anchor Model, and Inference

PMCanalSeg contributes 191 lower-canal cases at native spacing; all pass the frozen bilateral extraction and affine checks [6]. DentVoxel v1 contributes 100 cases at 0.3-mm spacing [18,19]; 95 contain the bilateral candidates for the transfer/same-world tests. The clearance construction uses predicted canal plus ground-truth tooth/canal in 92 of those cases after the three development cases are excluded, so it is not end-to-end. The prediction anchor is a public MONAI Attention U-Net checkpoint [12,1], whose model card describes ToothFairy2 training [2]. It was not fitted locally on the 92 evaluation cases, but image-level overlap with its original training set cannot be excluded. Frozen inference uses an intensity window of $[-200, 1800]$, reorients to RAS, applies 96^3 sliding windows with 0.25 overlap, thresholds at 0.5, and returns predictions losslessly to native space. The recipe was developed on three cases from the 95-case cohort and frozen before evaluation; those cases were excluded, leaving 92. Teeth and tooth identities are ground truth. We use de-identified research datasets under their source access terms; DentVoxel v1 is released under CC BY 4.0 [19]. We perform no new participant intervention and do not redistribute raw volumes, labels, or checkpoint weights.

The checkpoint anchors one constructed family in an actual model output. Its training images cannot be independently audited, and only one checkpoint, architecture, cohort, and ground-truth tooth pipeline are evaluated; witness coverage and conditional effects are therefore checkpoint- and cohort-specific. Three same-cohort cases informed the inference recipe and were excluded from the 92-case estimator, but this is not independently developed external validation.

Unless explicitly labeled as the crossed patient–refit robustness band, intervals resample patients 10,000 times from a fixed master seed with prespecified endpoint offsets, after averaging repeated rows within patient. The exact-family validation gate required complete attempt accounting, zero integer/metric/norm/vector discrepancies, independent all-row reconstruction, and at least 20 patients with an action-discordant exact row. A separate Family-B implementation recomputed all by-threshold intervals, the exact-pair Dice interval, and the simultaneous band defined above.

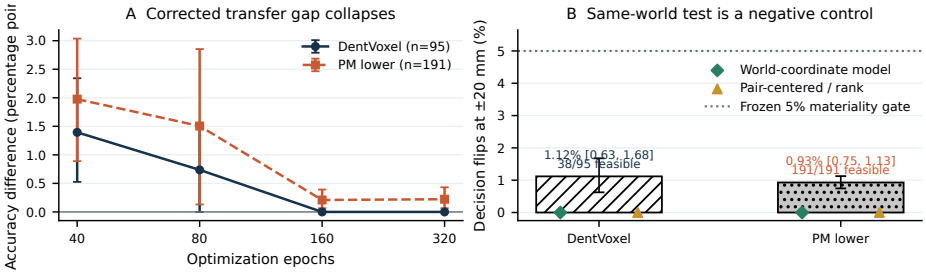

Fig. 1. Two negative audit results. **A**: patient-disjoint transfer contrasts across 20 refits; intervals resample patients after averaging refits. **B**: same-world flip rates and invariant controls. Complete-grid feasibility was 38/95 and 191/191; all verified world-point sets were identical (maximum drift 0 mm).

5 Results

5.1 The original transfer interpretation fails convergence and sufficiency checks

At 40 epochs, corrected gaps were 1.39% ([0.53, 2.34]%) for the 95-case target and 1.98% ([0.89, 3.04]%) for PM lower (Fig. 1A), while actual-refit ranges crossed zero: $[-3.16, 9.47]$ and $[-4.97, 9.95]$ percentage points. At 320 epochs the gaps were 0.00% and 0.22% ([0.07, 0.43]%). More decisively, physical-x and pair-centered rank achieved 1.000 accuracy in both corpora without fitting; constant and case-indexed random controls were approximately 0.5. The task is laterality indexing from ground-truth geometry, not learned CBCT navigation.

5.2 Same physical masks do not reverse the conclusion

All 2,400 prescribed fits completed, producing 1,204,800 patient/model/transform records. World-coordinate, pair-centered, and deterministic-rank controls had zero flips. The absolute-index model flipped 1.12% ([0.63, 1.68]%) and 0.93% ([0.75, 1.13]%) at ±20 mm (Fig. 1B). Only 40.0% of the 95-case corpus supported both signs across the full grid, below the frozen 80% feasibility gate; PM lower had 100% feasibility. No seed reached 5%, and no simultaneous band supported a transfer-sign reversal. This is a validated negative control, not evidence for a coordinate-shortcut rescue.

5.3 Exact overlap does not identify the clearance proxy

The real prediction anchor contains 92 patients and 809 posterior-tooth units, with mean canal Dice 0.576. At 2 mm, 49 units are GT-intimate and the prediction is clear for 16 in 12 patients. These are cohort descriptions, not a natural-error rate elsewhere.

Family A independently reconstructed all 92 cases and retained 169 exact pairs in 42 patients from 2,427 attempts. At 2 mm, patient-equal Dice was 0.9887 (95% CI $[0.9853, 0.9920]$); 48/49 critical deletions and all 49 far deletions preserved 26-connectivity. Thus the high-overlap theorem is data-instantiated, while its truth-derived construction remains explicit.

Family B recorded 16,180 attempts, of which 6,019 were geometrically feasible and 1,757 met exact overlap, equal norm, and opposite-vector constraints (Fig. 2). They span 64 patients; 61 had at least one action-discordant exact row anywhere on the frozen grid. The value 61/92 is constructed-grid witness coverage, not the prevalence of a natural model error. Patients without exact rows remain in feasibility denominators and are not assigned zero effect. At 2 mm, 827/1,757 rows were discordant across 57/64 patients; the patient-equal away-minus-toward action difference was 0.481 (95% CI $[0.402, 0.559]$). All named metric gaps, realized-norm gaps, and shift-vector sums were exactly zero under independent reconstruction.

Worked example. The displayed row was selected before visualization as the first under a prespecified deterministic ordering of all 2-mm-discordant exact rows, without filtering on Dice, magnitude, clearance, or appearance. Starting from one thresholded prediction, its tooth-side half was shifted by opposite equal-norm integer vectors; realized displacement was 5.739 mm, while the other half and anatomy remained fixed. Both arms intersected canal truth in 5,010 voxels and contained 22,439 predicted voxels. Dice (0.230525), IoU (0.130279), precision (0.223272), recall (0.238265), and volume similarity (0.967515) were identical. At $\tau = 2$ mm, the toward arm had $d = 0$ mm and action 0 (CLOSE), whereas the away arm had $d = 10.607$ mm and action 1 (CLEAR). This contributes +1 before within-patient averaging. It is a constructed actual-prediction witness, not an observed clinical error or prevalence estimate.

The rigid pairs are not equal-*high*-Dice examples: row-level median Dice was 0.286 (range $[0.017, 0.493]$). Their value is narrower: starting from actual predictions, an exact overlap-equivalence class can cross the proxy action boundary. The constructed tooth-relative direction and large-shift concentration prohibit inference about subtle errors, natural frequency, or clinical harm.

6 Discussion and Limitations

What is established. In the reconstructed two-corpus, mask-derived laterality diagnostic, convergence and deterministic sufficiency controls invalidate the original transfer interpretation. The same-world storage intervention fails its promotion gate. Within the frozen constructed dental-CBCT grid, the named count-derived overlap/volume summaries do not identify the specified tooth-conditioned clearance proxy. The falsification logic may transfer; these empirical findings remain dental-CBCT-specific.

What is not established. The experiments do not characterize domain transfer generally, architecture-independent robustness, natural witness prevalence, subtle

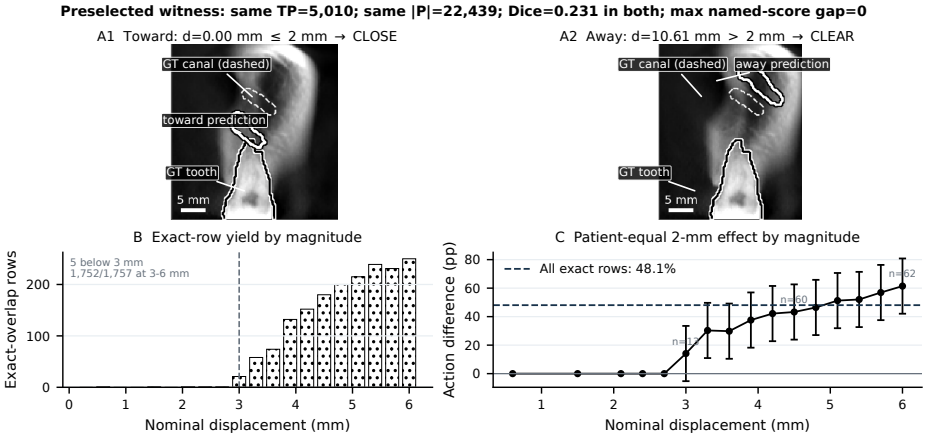

Fig. 2. Actual-prediction exact-overlap family. **A**: the illustrative witness was preselected by the same frozen deterministic ordering; identical counts and Dice yield opposite 2-mm verdicts. DentVoxel v1 is CC BY 4.0 [19]. **B**: just 5 exact rows occurred below 3 mm and none was action-discordant; yield is dominated by 3–6 mm edits. **C**: patient-equal 2-mm action difference by magnitude (simultaneous 95% bands for cells with at least five patients; sparse cells are shown without bands). Cohort conclusions use every frozen-grid row; the dashed line pools all exact rows and no magnitude is selected for inference.

or high-Dice errors, equality of surface or topology metrics, a validated clinical endpoint, treatment effect, injury risk, safety, or deployment.

The study has consequential limits. The transfer endpoint is a trivial two-candidate task; only two corpora are available; and the 20-mm complete-grid subset is selective in one corpus. The clearance thresholds are geometric proxies without outcome validation. The anchor uses one checkpoint and one evaluation cohort, with ground-truth teeth and unruled-out overlap with the checkpoint's training data. Family A uses truth to construct masks; Family B uses truth for exact-match retention and mostly requires large, low-Dice rigid edits. Connectivity is not guaranteed, annotation uncertainty is not propagated, and ASSD, HD95, surface Dice, and clDice are not held equal. The main paper contains every load-bearing fact; a separate supplement records proof details, complete denominators, and validation procedures.

Conclusion. Paired falsification rejected the reconstructed transfer interpretation; exact physical re-expression supplied no promoted replacement. A proof and frozen benchmark separately show that equal named overlap/volume summaries can coexist with different tooth-conditioned clearance verdicts.

Acknowledgements and Disclosure of Interests. No external funding was received. The authors declare no relevant competing interests.

References

1. abhijithkj3690: Alveolar canal segmentation: Public MONAI attention U-Net checkpoint. Hugging Face model repository (2026), https://huggingface.co/abhijithkj3690/alveolar-canal-segmentation, accessed 10 July 2026
2. Bolelli, F., et al.: Segmenting maxillofacial structures in CBCT volumes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5238–5248 (2025), https://openaccess.thecvf.com/content/CVPR2025/html/Bolelli_Segmenting_Maxillofacial_Structures_in_CBCT_Volumes_CVPR_2025_paper.html
3. Brown, A., et al.: Detecting shortcut learning for fair medical AI using shortcut testing. Nature Communications **14**, 4314 (2023). https://doi.org/10.1038/s41467-023-39902-7
4. Bruhns, M., et al.: Effects of segmentation errors on downstream analysis in highly multiplexed tissue imaging. PLOS Computational Biology **21**(9), e1013350 (2025). https://doi.org/10.1371/journal.pcbi.1013350
5. Hou, Z., et al.: SegTest: Metamorphic testing of image segmentation via guided instance-level test data augmentation. Software Testing, Verification and Reliability (2025). https://doi.org/10.1002/stvr.1910
6. Li, G., Lu, Y., Wu, G., Wang, L., Ma, R.: PMCanalSeg: A dataset for automatic segmentation of the pterygopalatine and mandibular canals from 3D CBCT images. Scientific Data **13**, 312 (2026). https://doi.org/10.1038/s41597-026-06620-w
7. Lin, M., et al.: Shortcut learning in medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. LNCS, vol. 15008, pp. 623–633 (2024). https://doi.org/10.1007/978-3-031-72111-3_59
8. Lyu, W., et al.: Assessing the spatial relationship between mandibular third molars and the inferior alveolar canal using a deep learning-based approach: A proof-of-concept study. BMC Oral Health (2025). https://doi.org/10.1186/s12903-025-06539-5
9. Maier-Hein, L., et al.: Metrics reloaded: Recommendations for image analysis validation. Nature Methods **21**(2), 195–212 (2024). https://doi.org/10.1038/s41592-023-02151-z
10. Ni, R., et al.: Geometrically focused training and evaluation of organs-at-risk segmentation via deep learning. Medical Physics **52**(7), e17840 (2025). https://doi.org/10.1002/mp.17840
11. Ntovas, P., et al.: Accuracy of artificial intelligence-based segmentation of the mandibular canal in CBCT. Clinical Oral Implants Research (2024). https://doi.org/10.1111/clr.14307
12. Oktay, O., et al.: Attention U-Net: Learning where to look for the pancreas. In: Proceedings of the 1st Conference on Medical Imaging with Deep Learning (2018), https://arxiv.org/abs/1804.03999
13. Poel, R., et al.: The predictive value of segmentation metrics on dosimetry in organs at risk of the brain. Medical Image Analysis p. 102161 (2021). https://doi.org/10.1016/j.media.2021.102161
14. Reinke, A., 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
15. Shit, S., et al.: clDice: A novel topology-preserving loss function for tubular structure segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16560–16569 (2021). https://doi.org/10.1109/CVPR46437.2021.01629

16. Stothers, D.: Beyond dice: Risk-normalized and hazard-aware evaluation of medical segmentation for image-guided robotics. In: NeurIPS 2025 Workshop on Evaluation and Standardization of AI for Robotics and Surgery (2025), https://openreview.net/forum?id=pY8YJHDCDp
17. Xie, X., et al.: Testing and validating machine learning classifiers by metamorphic testing. Journal of Systems and Software **84**(4), 544–558 (2011). https://doi.org/10.1016/j.jss.2010.11.920
18. Zhou, Y., Xu, Y., Khalil, B., Nalley, A., Tarce, M.: An open deep learning-based framework and model for tooth instance segmentation in dental CBCT. Clinical Oral Investigations **29**, 473 (2025). https://doi.org/10.1007/s00784-025-06578-w
19. Zhou, Y., Xu, Y., Tarce, M.A.: DentVoxel: A fully annotated dental CBCT dataset with 38 instance anatomical structures. figshare Dataset, version 1 (2026). https://doi.org/10.6084/m9.figshare.31239889.v1, CC BY 4.0