Evaluating Pulmonary Vessel Segmentation Beyond Dice: Connectivity, Caliber, and Clinical Validity

Literature review and conceptual analysis

Abstract

Pulmonary vessel segmentation benchmarks commonly rank algorithms by regional overlap, but a vascular tree can be volumetrically similar to a reference while containing clinically consequential breaks, false joins, missing peripheral branches, or biased diameters. This evidence synthesis develops an evaluation framework aligned with the structural and translational purposes of pulmonary vessel analysis. It distinguishes six dimensions: regional agreement, boundary accuracy, centerline connectivity, branch detection, morphometric fidelity, and clinical task validity. MorVess provides a useful focal example because it reports Dice, centerline Dice, 95th-percentile Hausdorff distance, branch-oriented measures, cross-dataset tests, and geometric comparisons of vessel volume and caliber. Its design also demonstrates that evaluation and learning are coupled: distance and thickness targets improve precisely the properties that voxel-wise supervision tends to neglect. The article argues that no scalar metric can represent a pulmonary vascular reconstruction adequately. Evaluation should be stratified by vessel size and anatomical level, estimate uncertainty at the patient level, test robustness across acquisition domains, and connect segmentation errors to downstream measurements or expert decisions. A tiered protocol is proposed for algorithm development, challenge benchmarking, and clinical validation. The goal is not to replace Dice but to place it inside a multidimensional evidence structure that distinguishes larger masks from better vascular models.

The Metric Defines the Object of Interest

An evaluation metric is not merely a neutral ruler. It encodes which disagreements matter and which can be averaged away. The Dice coefficient originated as a measure of association between sets (Dice, 1945) and became popular in medical segmentation because it summarizes foreground overlap in a compact, scale-normalized form. For organs and large lesions, that summary is often informative. For pulmonary vessels, it is incomplete because the foreground has a highly unequal distribution of calibers. Central vessels contain many voxels, whereas long peripheral branches contain few. A model can improve Dice by refining central boundaries while losing a distal branch that matters to vascular abundance or pruning.

This asymmetry changes the interpretation of rankings. A one-point Dice improvement may represent a useful reduction in widespread error, or it may reflect a small change around high-volume trunks. Conversely, a method that restores many thin branches may add true positives and some uncertain terminal predictions without producing a dramatic Dice gain. Evaluation must determine whether the intended object is a foreground region, a continuous tree, a set of named anatomical branches, or a measurement substrate. Different objects require different evidence.

Moccia et al. (2018) showed that vessel-segmentation studies vary widely in methods, datasets, and evaluation metrics. The field has since gained larger challenges and stronger learning systems, but the comparison problem remains. Self-configuring methods such as nnU-Net can optimize preprocessing, patch configuration, and training for a particular dataset (Isensee et al., 2021). Such systems are essential baselines, yet a high aggregate score does not reveal whether an anatomical topology is preserved. The evaluation framework must therefore separate pipeline competence from structural adequacy.

Dimension 1: Regional Agreement

Dice, sensitivity, precision, and related overlap measures remain the first dimension because they answer fundamental questions. Sensitivity estimates how much labeled vasculature is recovered. Precision estimates how much predicted foreground corresponds to the reference. Dice balances the two. Reporting all three helps distinguish conservative models that omit vessels from expansive models that introduce false positives. A single Dice score hides that tradeoff.

Regional metrics should be computed per patient and summarized with distributions, not only a pooled voxel count. Pooled analysis lets large scans or high-vessel-volume cases dominate. Mean, median, dispersion, confidence intervals, and paired differences against a baseline provide a more honest account. Taha and Hanbury (2015) review a broad range of three-dimensional segmentation measures and show why metric selection should be matched to object characteristics. Their analysis supports using complementary families rather than treating overlap as universal.

For pulmonary vessels, regional scores should also be stratified. Central and peripheral compartments can be defined by anatomical labels, distance from the hilum, branch generation, or reference caliber. The chosen rule must be declared and applied consistently. Caliber bins in physical millimeters are particularly useful because they expose the small-vessel bias of voxel-weighted objectives. If an algorithm gains overall Dice but loses sensitivity below a specified diameter, users can see the tradeoff instead of inferring it from a volume rendering.

Dimension 2: Boundary and Surface Accuracy

Boundary errors affect lumen diameter, cross-sectional area, and derived hemodynamic quantities. Surface-distance metrics therefore complement overlap. The Hausdorff distance identifies the largest separation between two point sets, as formalized for image comparison by Huttenlocher et al. (1993). In medical segmentation, the full maximum is highly sensitive to one outlier, so the 95th-percentile Hausdorff distance, or HD95, is commonly used. Average symmetric surface distance can describe typical boundary deviation, while HD95 emphasizes substantial localized failure.

These measures require physical spacing. A two-voxel error can mean less than a millimeter in-plane but several millimeters through thick slices. Distances computed after isotropic resampling may be easier to compare but can conceal interpolation artifacts. Reports should state whether metrics are calculated in native or resampled geometry and should convert coordinates to millimeters. Surface tolerances should be linked to the downstream use rather than borrowed uncritically from another organ.

Boundary metrics also need anatomical stratification. A one-millimeter error is proportionally small for a central vessel and catastrophic for a two-millimeter branch. Relative caliber error or normalized surface distance can add context, although normalization should not make noisy sub-voxel annotations appear precise. Manual vessel boundaries near the resolution limit are uncertain. Evaluation should estimate reader variability or define an acceptable tolerance band instead of treating every reference voxel as exact.

Dimension 3: Centerline Connectivity and Topology

A vascular segmentation is often converted to a skeleton or graph for branch analysis. Centerline discontinuity can invalidate that conversion even when the surrounding mask is mostly correct. Shit et al. (2021) introduced clDice to measure agreement between skeletons and opposite masks. It emphasizes whether predicted centerlines lie inside the reference and whether reference centerlines are covered by the prediction. Used with Dice, it distinguishes volumetric coverage from network continuity.

clDice is valuable but not sufficient. Skeletonization can create spurs in noisy masks, and a false shortcut may connect two correct branches while retaining favorable centerline overlap. Additional graph measures should count connected components, endpoints, bifurcations, cycles, and path-length discrepancies. Betti numbers can summarize components and loops, but their clinical meaning depends on the expected

anatomy. Pulmonary arterial and venous trees are largely branching structures; false cycles may indicate leakage between adjacent vessels, yet small peripheral loops or annotation conventions can complicate a rigid topological rule.

Path-based evaluation offers a more interpretable alternative. Select anatomically meaningful source and target points, then test whether a valid predicted path exists and how closely its centerline follows the reference. The fraction of reference branch length connected to the central trunk can quantify usable tree recovery. Such measures should be robust to small skeleton shifts through a declared spatial tolerance. The tolerance must not be so wide that parallel artery and vein segments are treated as the same path.

Morphology-aware learning and evaluation reinforce each other. Deep Distance Transform predicts a continuous distance field that represents tubular scale and supports geometry-aware refinement (Wang et al., 2020). Dynamic Snake Convolution adapts its sampling to curvilinear structures and adds a continuity-oriented constraint (Qi et al., 2023). These models make claims about properties beyond region overlap, so their tests should directly measure those properties. An architecture should not be called topology-preserving solely because its Dice improves.

Dimension 4: Branch and Caliber Recovery

Branch detection turns topology into a countable clinical and anatomical unit. A reference tree can be divided at bifurcations, and predicted segments matched by spatial proximity and path overlap. Branch precision and recall then reveal missed terminals and false extensions. Detection should be reported by caliber or branch order because major and minor branches differ greatly in visibility and consequence. The matching algorithm, pruning threshold, and treatment of short spurs must be published; otherwise branch scores are not reproducible.

The PARSE challenge was created to benchmark efficient segmentation of both main and branch pulmonary arteries (Luo et al., 2023). Its emphasis on multilevel anatomy and computational cost is important. A model that performs well on the trunk but poorly on branches is not interchangeable with one that recovers both. Challenge design can make this distinction visible by weighting compartments separately rather than allowing the large-vessel volume to determine the ranking.

Caliber fidelity deserves its own tests. Compare predicted and reference diameters along matched centerlines, reporting bias, absolute error, correlation, and distributional divergence. A high correlation does not exclude systematic overestimation, while a low mean error can conceal compensating biases at large and small scales. Bland-Altman analysis across vessel sizes can reveal proportional error. Measurements should use physically meaningful cross-sections perpendicular to the centerline when feasible, because axial widths vary with vessel orientation.

Small-vessel volume fraction is another useful measure, but it is sensitive to threshold definition and image resolution. A threshold stated as cross-sectional area or diameter in millimeters is more portable than a voxel count. Results should show how conclusions change across plausible thresholds. This sensitivity analysis is especially necessary when comparing normal-dose CT, low-dose CT, and scans reconstructed at different slice thicknesses.

MorVess as a Multidimensional Evaluation Case

MorVess is notable not only for its architecture but for the variety of evidence used to support its morphology-aware claims. Mao et al. (2026) jointly predict a binary vessel mask, a boundary-oriented distance map, and a centerline-derived thickness map. A lightweight 2.5D adapter incorporates information from five adjacent slices into a frozen Segment Anything Model encoder, while a global-local fusion block combines multiple semantic levels and geometric predictions. Training uses cross-entropy, Dice, clDice, distance, and thickness losses in two stages.

The primary evaluation spans Parse2022 and AIIB2023. Reported metrics include Dice, clDice, HD95, an average missing rate, and branch or detection-related measures. On Parse2022, MorVess reports mean Dice of 86.84, clDice of 83.22, and HD95 of 4.53 mm. On AIIB2023, it reports 94.31, 89.34, and 3.24 mm respectively. Ablation results associate distance-map supervision with a reduction in HD95 and thickness-map supervision with improved clDice. This alignment between intended mechanism and observed metric is stronger evidence than an isolated final score.

The study further examines total vessel volume, diameter-distribution agreement, tortuosity, and the fraction of small vessels. These outcomes move evaluation closer to vascular measurement. Cross-dataset experiments apply models without fine-tuning to HiPas pulmonary veins and ATM2022 airways, testing whether the learned priors transfer across appearance and anatomy. The authors report retained performance with expected decreases relative to in-domain testing. They also disclose a principal limitation: the 2.5D adapter sees only two neighboring slices on each side and assumes near-isotropic local geometry, limiting long-range depth reasoning and robustness to anisotropic CT.

Several cautions remain. Parse2022 and AIIB2023 are benchmarks, not a prospective multi-center clinical deployment. Some cross-domain targets are veins or airways rather than the same pulmonary artery task under a new hospital protocol. Reported geometric agreement is encouraging, but a biomarker can be precise against annotation and still fail to predict or monitor disease. The comparison includes bootstrap paired testing, yet confidence intervals and case-level failure analysis remain important for determining who benefits and where the model breaks.

Dimension 5: Robustness Across Domains and Diseases

Generalization should be evaluated along identifiable axes: scanner vendor, reconstruction kernel, radiation dose, contrast phase, slice spacing, field of view, disease, age, and institution. Combining all external cases into one score can conceal a failure concentrated in low-dose or fibrotic scans. The AIIB23 challenge specifically addresses imaging biomarkers in pulmonary fibrosis, where distorted parenchyma and abnormal tubular structures create a demanding setting (Nan et al., 2024). Performance on such data should be analyzed by disease severity and pattern, not only as a dataset average.

Cross-validation inside one dataset estimates sampling variation but not deployment shift. A rigorous study should reserve an untouched external site and report the full preprocessing path. Patient-level splitting is mandatory; slice-level splitting leaks nearly identical anatomy across training and test sets. Model selection, threshold tuning, and post-processing parameters must be fixed before external evaluation. If adaptation is allowed, the number and type of target-domain labels should be disclosed.

Robustness includes stability to technically plausible perturbations. Reconstructing the same scan with a different kernel, adding calibrated noise, altering resolution, or varying intensity windows can reveal whether morphology is supported by anatomy or by brittle texture. Test-retest scans can measure repeatability of branch count, volume, and caliber. A stable segmentation is not necessarily accurate, but instability sets an upper bound on the reliability of derived biomarkers.

Dimension 6: Clinical and Scientific Task Validity

The final dimension asks whether segmentation improves a real task. Chu et al. (2025) developed HiPaS for pulmonary artery-vein segmentation across contrast and non-contrast CT and evaluated it on multi-center data. They then used large-scale outputs to study vascular abundance in 11,784 participants. This design illustrates how segmentation evidence can progress from overlap to anatomical discovery. It also shows the importance of artery-vein distinction when biological interpretation differs between the two systems.

PaSAL adds another layer by combining binary vessel segmentation, graph construction, and labeling of 19 clinically defined vascular classes (Eppink et al., 2026). Expert review in that study showed weak correlation between standard quantitative metrics and perceived quality, a warning against assuming that benchmark

rank equals clinical plausibility. Anatomical labeling, surgical planning, embolism assessment, or pulmonary hypertension research may each value different branches and error types.

Clinical validation should predefine the intended use and compare decisions or measurements with and without algorithmic support. For morphometry, report agreement with expert measurements and associations with outcomes. For planning, measure whether relevant branches are identified and whether review time or correction burden decreases. Reader studies should randomize case order, blind method identity, and include representative difficult cases. Automation bias must be considered: a polished three-dimensional tree can make a false connection appear convincing.

Uncertainty, Annotation Quality, and Statistical Inference

Reference masks are observations, not perfect anatomy. Complete manual labeling of distal pulmonary vessels is slow, ambiguous, and dependent on windowing and reconstruction. Evaluation should describe the number and expertise of annotators, consensus process, software, stopping rules, and use of model-assisted correction. A subset with multiple independent readers allows estimation of interobserver variability. Claims smaller than that variability require caution.

Uncertainty should be reported spatially and structurally. Voxel probabilities can be calibrated, but branch-level uncertainty is often more useful. One can estimate the probability that a terminal remains connected to the trunk across an ensemble or perturbation set. Diameter intervals can be propagated into biomarker uncertainty. Cases with high uncertainty or disagreement among mask, distance, and thickness predictions may be routed for review.

Statistical comparisons must use the patient as the independent unit. Multiple metrics and datasets increase the risk of selective significance, so primary endpoints should be declared and secondary analyses labeled. Paired bootstrap or permutation procedures can estimate differences without assuming normality, but confidence intervals are more informative than thresholded p-values alone. Failure-case taxonomies should accompany averages: missing terminal branches, false artery-vein bridges, leakage into airway walls, truncation at patch boundaries, and errors around pathology.

A Tiered Evaluation Protocol

Tier 1 is development validation. It includes Dice, sensitivity, precision, surface distance, HD95, clDice, connected components, and inference cost on a fixed patient-level split. Results are stratified by caliber and central versus peripheral anatomy. Qualitative panels are selected by prespecified criteria rather than only by visual appeal.

Tier 2 is benchmark validation. It uses an untouched challenge test set, a strong self-configuring baseline, standardized post-processing, case-level uncertainty, and paired statistical comparisons. Branch recall, diameter error, and path connectivity are required alongside overlap. Compute time and memory are measured with declared hardware.

Tier 3 is external robustness. At least one independent institution and materially different acquisition protocol are evaluated without post hoc retuning. Results are stratified by scanner and disease, and test-retest or reconstruction perturbations assess measurement stability. Domain adaptation, if used, is treated as a separate experiment.

Tier 4 is task validation. Segmentation-derived biomarkers, anatomical labels, or clinical decisions are compared with reference measurements and expert performance. Reader correction burden, failure visibility, and uncertainty-based referral are assessed. This tier establishes whether a structurally better mask is also a more useful clinical instrument.

Conclusion

Dice remains a valuable measure of regional agreement, but it cannot determine whether a pulmonary vascular tree is connected, anatomically faithful, morphometrically unbiased, or clinically useful. Evaluation should combine regional overlap, physical surface error, centerline and graph integrity, branch and caliber recovery, cross-domain robustness, and downstream task validity. MorVess demonstrates the benefits of aligning morphology-aware supervision with morphology-aware evidence: distance and thickness fields are tested through boundary, topology, and geometric outcomes rather than Dice alone. Its results support this direction while also revealing the need for longer-range volumetric context and broader clinical validation. A tiered protocol makes claims proportional to evidence and turns benchmark comparison into a more meaningful question: not merely how many vessel voxels are recovered, but whether the reconstructed network can support dependable pulmonary vascular analysis.

References

Mao, F., Chen, Y., Wu, B., Lin, L., Dai, J., Li, Z., ... & Qin, F. (2026). MorVess: Morphology-Aware Pulmonary Vessel Segmentation Network. *arXiv preprint arXiv:2606.24214*.

Chu, Y., Luo, G., Zhou, L., Cao, S., Ma, G., Meng, X., ... & Gao, X. (2025). Deep learning-driven pulmonary artery and vein segmentation reveals demography-associated vasculature anatomical differences. *Nature Communications*, *16*, 2262.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297-302.

Eppink, J., Kervadec, H., van Capelleveen, J., Verhoeff, J., Senan, S., & Bohoudi, O. (2026). PaSAL: A deep learning pipeline for pulmonary artery-vein segmentation and anatomical labeling in thoracic CT. In *Proceedings of the 9th International Conference on Medical Imaging with Deep Learning* (pp. 3561-3592). PMLR.

Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. J. (1993). Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*(9), 850-863.

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

Luo, G., Wang, K., Liu, J., Li, S., Liang, X., Li, X., ... & Gao, X. (2023). Efficient automatic segmentation for multi-level pulmonary arteries: The PARSE challenge. *arXiv preprint arXiv:2304.03708*.

Moccia, S., De Momi, E., El Hadji, S., & Mattos, L. S. (2018). Blood vessel segmentation algorithms - review of methods, datasets and evaluation metrics. *Computer Methods and Programs in Biomedicine*, *158*, 71-91.

Nan, Y., Xing, X., Wang, S., Tang, Z., Felder, F. N., Zhang, S., ... & Gao, X. (2024). Hunting imaging biomarkers in pulmonary fibrosis: Benchmarks of the AIIB23 challenge. *Medical Image Analysis*, *97*, 103253.

Qi, Y., He, Y., Qi, X., Zhang, Y., & Yang, G. (2023). Dynamic Snake Convolution Based on Topological Geometric Constraints for Tubular Structure Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6070-6079).

Shit, S., Paetzold, J. C., Sekuboyina, A., Ezhov, I., Unger, A., Zhylka, A., ... & Menze, B. H. (2021). 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).

Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging*, *15*, 29.

Wang, Y., Wei, X., Liu, F., Chen, J., Zhou, Y., Shen, W., Fishman, E. K., & Yuille, A. L. (2020). Deep Distance Transform for tubular structure segmentation in CT scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3833-3842).