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: PCER, Cohort, and Reference-Geometry Analyses

Ethan Yu[1,2][0000−0001−7665−3268] and Yunsik Son[2][0000−0002−2580−4393]⋆

[1] DEEP.FINE Co., Ltd., Seoul, Republic of Korea
[2] Department of Computer Science and Artificial Intelligence, Dongguk University, Seoul, Republic of Korea
sonbug@dongguk.edu

1 Scope and interpretation boundary

This supplement reports an exploratory geometric sensitivity analysis of the airway reference annotations. It is deliberately separate from the main results. The targets are peripheral reference branch endpoints, not lesions, and no named device is modelled. The analysis therefore does not estimate bronchoscope traversal, lesion access, procedural success, respiratory motion, CT-to-body divergence, or operator behaviour. No training or model inference was performed for this analysis.

2 PCER parameter sensitivity

The main analysis defines peripheral connected-endpoint recall (PCER) using the furthest 50% of reference branch endpoints, ranked by distance from the tracheal centroid in physical millimetres, and a 2 mm physical target tolerance. We varied the retained endpoint fraction (furthest 25%, 50%, or 75%) and tolerance (1 mm, 2 mm and 3 mm) while holding 26-connectivity and the tracheal seed rule fixed. Each setting was recomputed for all 103 cases and ten methods from the final predictions; no training or inference was repeated. The primary 50%/2 mm setting reproduced all 1,133 stored case-mask scores, including ground truth, with zero mismatches.

At the primary setting, mean PCER across methods is 78.6% for ATM'22, 53.8% for EXACT'09, and 85.8% for AeroPath. Across all nine settings, the corresponding ranges are 68.7–85.1%, 47.4–59.2%, and 80.6–89.4%. With the furthest 50% fixed, increasing the tolerance from 1 mm to 3 mm raises mean PCER by 9.0, 4.3, and 3.4 points, respectively. With 2 mm fixed, expanding from the furthest 25% to 75% raises the means by 7.2, 7.9, and 5.4 points, as more proximal endpoints enter the target set. The primary setting remains unchanged because it was predefined; sensitivity settings are not used to select a more favourable result.

⋆ Corresponding author.

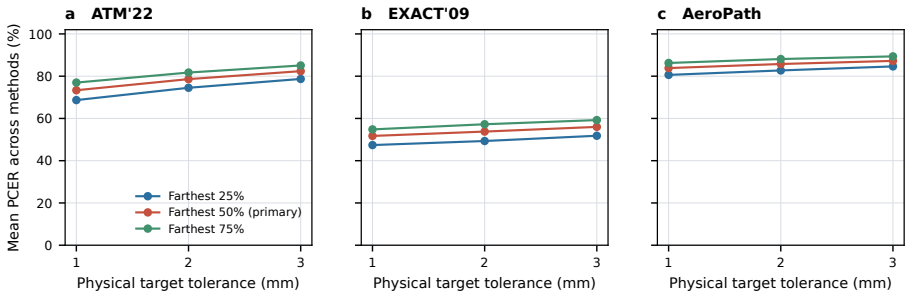

Fig. 1. PCER sensitivity to physical target tolerance and the proportion of furthest reference endpoints retained. Points are cohort means of case-level means across ten methods. These are digital evaluation choices, not device-feasibility constraints.

Table 1. Physical-distance erosion sensitivity. Association rows use the same post-hoc cohort-adjusted HC3 model; ρ and median absolute error compare automatic with annotation-derived variability at the same erosion radius.

Radius	β (PCER points/HU)	HC3 95% CI	p	Agreement ρ / median error
1 mm	−0.506	[−0.713, −0.300]	1.56×10^{-6}	0.838 / 3.99 HU
2 mm	−0.507	[−0.720, −0.294]	3.07×10^{-6}	0.935 / 2.04 HU
3 mm	−0.510	[−0.725, −0.296]	3.19×10^{-6}	0.942 / 1.84 HU

3 Physical-distance erosion sensitivity

The locked analysis uses two binary-erosion iterations. Because that operation is voxel-based, we repeated the exposure and automatic-measurement calculations with anisotropic Euclidean distance transforms, retaining lumen voxels more than 1 mm, 2 mm and 3 mm from background. All 103 cases retained at least 50 core voxels at every setting. No training or segmentation inference was repeated.

At the rounded 34 HU threshold, the 2 mm physical setting has sensitivity 79.2% (Wilson 95% CI 59.5–90.8%) and specificity 98.7% (93.2–99.8%). The threshold composition changes because the annotation-derived physical- erosion variability differs from the locked two-voxel value. The stable association slopes support erosion-scale robustness, not prospective validity or a specific acquisition mechanism.

4 Cohort-adjusted and method-wise analyses

The main post-hoc model regresses case-level mean PCER on continuous η and fixed cohort indicators, using HC3 inference. Here η is tracheal-lumen intensity variability; its slope is −0.508 PCER points/HU (95% CI −0.723 to −0.293; $p = 3.57 \times 10^{-6}$). At the pooled mean variability, the adjusted EXACT'09–ATM'22 difference is −13.70 points (95% CI −25.63 to −1.77; $p = 0.0245$), and the AeroPath–ATM'22 difference is +8.59 points (95% CI 3.50–13.68; $p =$

Table 2. Repeated-measures, covariate, and leave-one-cohort-out sensitivities. The GEE row reports robust sandwich inference; other rows use HC3 inference.

Analysis	β (PCER points/HU)	95% CI	p
Main post-hoc ten-method case mean	-0.508	$[-0.723, -0.293]$	3.57×10^{-6}
Method-adjusted GEE (1,030/103)	-0.508	$[-0.696, -0.319]$	1.28×10^{-7}
Endpoint/spacing-adjusted	-0.523	$[-0.738, -0.307]$	2.07×10^{-6}
Omit ATM'22 ($n = 47$)	-0.558	$[-0.800, -0.316]$	6.26×10^{-6}
Omit EXACT'09 ($n = 83$)	-0.272	$[-0.679, 0.135]$	0.190
Omit AeroPath ($n = 76$)	-0.501	$[-0.729, -0.274]$	1.60×10^{-5}

Table 3. Cohort composition, retained peripheral-target counts, and within-cohort variability–PCER associations. Upper means $\eta \geq 33.87$ HU; p_H is Holm-adjusted across the three cohort correlations.

Cohort	n	Below/upper	Targets, median [IQR]	Range	ρ	Raw p	p_H
ATM'22 internal val.	56	47/9	115 [87.8,140]	12–239	0.071	0.605	0.605
EXACT'09 external	20	8/12	80.5 [61.5,99.3]	36–168	-0.632	0.00282	0.00845
AeroPath external	27	22/5	59 [41.0,84.5]	26–151	-0.419	0.0294	0.0588

0.000944). The variability-by-cohort interaction test does not resolve heterogeneity ($\chi^2_2 = 2.98$, $p = 0.226$) and does not establish shared cohort slopes. Cohort-specific interaction-model slopes are -0.034, -0.555, and -0.591 points/HU for ATM'22, EXACT'09, and AeroPath, respectively. A within-cohort standardized-rank sensitivity gives $r = -0.194$ (20,000 stratified permutations, $p = 0.0495$). These analyses were added after inspecting the initial pooled and cohort results and are not presented as preregistered confirmation.

The main post-hoc estimand is the association with the unweighted average PCER of the fixed ten-method benchmark panel. A repeated-measures Gaussian GEE instead retains all 1,030 method–case rows, adds fixed method and cohort indicators, groups the ten predictions from each case under an exchangeable working correlation, and uses robust sandwich inference. Because each case has the same ten-method panel, its identical point estimate is expected; the GEE chiefly validates within-case dependence handling and robust uncertainty. A second case-level model adjusts for $\log_2$ target count and the acquisition covariates retained in the evaluation NIfTI files: slice spacing and mean in-plane spacing. Scanner manufacturer, scanner model, reconstruction kernel, dose, site, and pathology metadata were unavailable.

Median slice spacing [range] is 0.50 [0.50–5.00] mm for ATM'22, 1.00 [1.00–1.00] mm for EXACT'09, and 0.50 [0.499–1.25] mm for AeroPath. Median in-plane spacing [range] is 0.781 [0.607–0.896], 1.00 [1.00–1.00], and 0.703 [0.563–0.824] mm, respectively. Target-count distributions are reported in Table 3. The adjusted slope remains negative, but omission of EXACT'09 attenuates it and removes statistical significance; this supports the stated cohort-structured interpretation.

All ten method-specific pooled correlations have the same negative direction, but predictions from methods evaluated on the same cases are correlated and

Table 4. Method-wise secondary pooled intensity-variability–PCER associations and exploratory split means. Holm p adjusts the ten continuous correlations.

Method	ρ	Holm p	Below	Upper	Δ
nnU-Net v2	−0.363	0.00145	78.39	50.44	27.95
Skeleton Recall	−0.274	0.0254	91.03	71.94	19.08
vesselFM	−0.237	0.0322	72.56	55.42	17.13
DeNVeR adaptation	−0.331	0.00438	80.61	52.17	28.45
NETracer adaptation	−0.273	0.0254	80.25	55.92	24.33
GLCP paper objective	−0.372	0.00111	78.87	49.57	29.30
GLCP released code	−0.264	0.0254	73.02	50.43	22.59
vesselFM + Skel. Recall	−0.213	0.0322	89.39	75.48	13.92
DynUNet + Skel. Recall	−0.317	0.00657	81.39	54.65	26.73
Skel. Recall + shell	−0.354	0.00191	89.12	68.87	20.25

Table 5. Paired within-case PCER comparison with TD and BD. "Opposite" counts non-tied method pairs ordered in opposite directions. "Separated near ties" counts pairs within one TD/BD point but at least ten PCER points apart.

Comparator	Median τ_b [IQR]	Same top method	Opposite pairs	Separated near ties
TD	0.776 [0.690,0.855]	90/103 (87.4%)	426/4,138 (10.3%)	82/1,014 (8.1%)
BD	0.782 [0.689,0.852]	86/103 (83.5%)	374/3,970 (9.4%)	48/729 (6.6%)

are not independent replications. Table 4 reports this descriptive consistency in full.

5 PCER relative to TD and BD

PCER is strongly associated with TD across method-level cohort means, but it is not interchangeable with TD or BD within individual cases. We ranked the ten methods separately within every case. Table 5 reports Kendall τ_b, top-method agreement, and all non-tied paired method comparisons. The near-tie analysis is descriptive: it defines a conventional-metric near tie as at most one point and a material PCER separation as at least ten points; these thresholds were not used for hypothesis testing.

For a matched TD example, AeroPath case 12 gives vesselFM (TD 91.14, BD 79.49, Dice 88.36, precision 94.81, PCER 81.36) and GLCP (TD 90.15, BD 77.78, Dice 82.27, precision 88.97, PCER 27.12). For a matched BD example in the same case, baseline (TD 93.70, BD 88.03, Dice 86.25, precision 91.23, PCER 86.44) and DeNVeR (TD 93.04, BD 88.03, Dice 85.05, precision 91.88, PCER 37.29) have equal BD. These examples show PCER's incremental endpoint-level description; they do not establish clinical or method superiority.

6 Zero-score and preprocessing provenance

The predefined seed rule assigns PCER zero when a prediction does not intersect the reference trachea. Three of 1,030 case-method predictions meet this

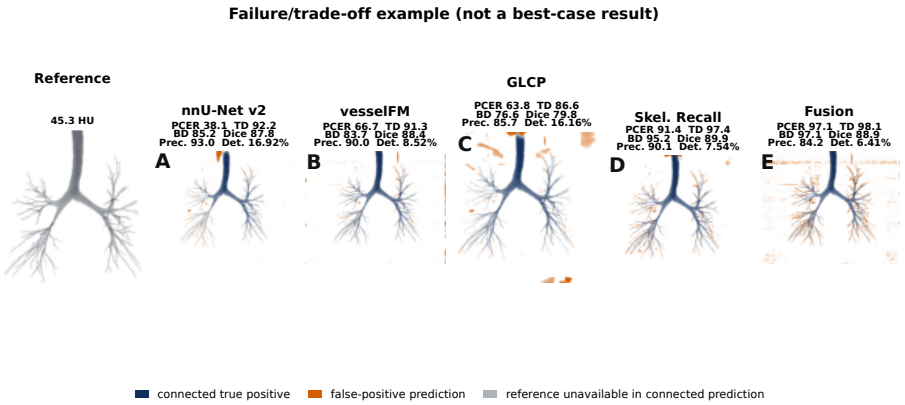

Fig. 2. Deliberate failure/trade-off example for EXACT'09 CASE09 (45.3 HU). The reference is unlettered; A–E are five segmentation methods, each labelled with PCER, TD, BD, Dice, precision, and detached foreground. Navy is trachea-connected true-positive prediction, orange is actual false-positive prediction, and light gray is reference airway unavailable in the connected prediction. "Det." reports detached foreground as a percentage of the raw prediction. The figure is qualitative and was retained to expose, not conceal, the connectivity–precision trade-off.

condition: EXACT'09 CASE03 for GLCP released code, and CASE14 for both the GLCP paper objective and released code. They remain in all aggregate and sensitivity analyses as zeros.

The EXACT'09 conversion manifest records z-axis reversal before evaluation for CASE03, CASE04, CASE06, CASE15, CASE19, and CASE20. This correction was based on image–label alignment and logged before method scoring; no case was removed because of its result.

7 Deliberate failure/trade-off example

Figure 2 retains EXACT'09 CASE09 as a deliberately labelled failure/trade-off example rather than evidence that one method is uniformly best. Fusion has PCER 97.1%, TD 98.1%, BD 97.1%, Dice 88.9%, and precision 84.2%, whereas Skeleton Recall has PCER 91.4%, TD 97.4%, BD 95.2%, Dice 89.9%, and precision 90.1%. Fusion's detached foreground is 6.41% of its prediction across 2,175 components; 98.7% of those detached voxels are false positives. Thus the case illustrates why PCER must be accompanied by overlap, precision, and false-positive reporting.

8 Reference geometry and marginal analysis

For each of 9853 peripheral reference endpoints from 103 cases (ATM'22 internal validation fold, $n = 56$; EXACT'09, $n = 20$; AeroPath, $n = 27$), we extracted

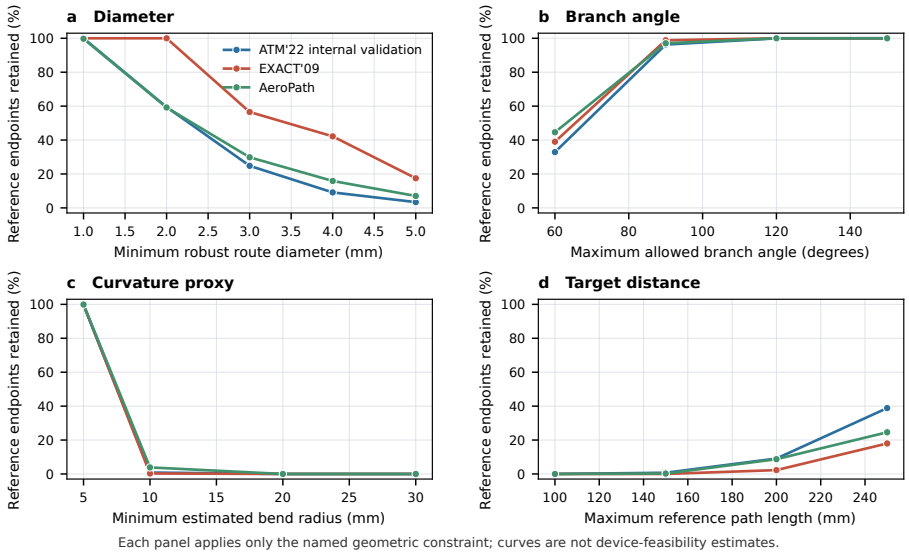

Fig. 3. Exploratory marginal reference-geometry sensitivity. Each point is the fraction of peripheral reference endpoints satisfying only the displayed diameter, branch-angle, curvature-proxy, or path-distance constraint. Curves are not device-feasibility estimates; the curvature proxy depends on centreline processing.

the trachea-to-endpoint reference-centreline path in physical millimetres. Four geometric quantities were computed: path length, a robust minimum diameter, maximum branch angle, and minimum estimated bend radius. The robust diameter is the fifth percentile of local reference-lumen diameter before the final 5 mm terminal approach. Bend radius is derived from the smoothed reference centreline and is sensitive to skeleton extraction, resampling, and smoothing; it is a curvature proxy rather than a device mechanics model.

Figure 3 reports four *marginal* sweeps. Each curve applies only the named constraint. We do not combine the constraints into a composite envelope because doing so would imply unvalidated device rules.

9 Exploratory findings

The curves show that any apparent feasibility fraction depends strongly on the selected constraint. At a 2 mm minimum robust diameter, endpoint retention is 59.4% for ATM'22, 100.0% for EXACT'09, and 59.2% for AeroPath. A 90° maximum branch-angle constraint retains 96.3–98.8% across cohorts. In contrast, raising the estimated minimum bend radius from 5 mm to 10 mm collapses retention from 99.5–100.0% to 0.3–3.9%. With a 250 mm maximum reference path length, retention ranges from 18.0% to 38.8%.

These findings support reporting sensitivity. They do not support a universal cutoff. In particular, the bend-radius result could primarily reflect centreline measurement choices rather than physical device behaviour. Peripheral connected-endpoint recall (PCER) remains a proxy for segmentation connectivity. It is not evidence of device or lesion reach.

10 Required next validation

The required next study is a lesion-level dataset containing real lesion coordinates, named-device specifications, expert-planned or clinically traversed routes, respiratory motion information, and CT-to-body registration or divergence measurements. Device diameter, bending, angle, curvature, and distance constraints must be predefined before method evaluation. The primary endpoint should be actual successful target access, with paired per-lesion comparisons between segmentation methods; PCER may be retained only as a secondary segmentation-connectivity measure.

Better airway connectivity may support planning, but clinical and device feasibility remains unvalidated.