Supplementary Information

Supplementary Table 1. Comparison of Delineation Performance Metrics Across the Five Centers.

Metrics	Delineation	Mean value (95% CI)					Welch's ANOVA	
		Center 1	Center 2	Center 3	Center 4	Center 5	F (DFn, DFd)	P-value
vDSC	M	0.866 (0.862, 0.869)	0.837 (0.833, 0.841)	0.856 (0.852, 0.861)	0.879 (0.875, 0.883)	0.867 (0.863, 0.872)	54.50 (4.00, 4470)	6.35E-45
	AI	0.896 (0.893, 0.899)	0.901 (0.899, 0.904)	0.897 (0.894, 0.900)	0.889 (0.885, 0.892)	0.891 (0.888, 0.895)	10.1 (4.00, 4402)	4.17E-08
	Aa	0.902 (0.900, 0.905)	0.905 (0.902, 0.907)	0.904 (0.901, 0.907)	0.897 (0.894, 0.901)	0.895 (0.891, 0.898)	7.53 (4.00, 4376)	4.82E-06
HD95 (mm)	M	6.51 (6.17, 6.85)	10.70 (9.95, 11.50)	7.15 (6.60, 7.70)	7.01 (6.24, 7.78)	6.51 (6.07, 6.95)	25.83 (4.00, 4522)	3.34E-21
	AI	5.64 (5.43, 5.85)	5.54 (5.35, 5.72)	5.85 (5.45, 6.24)	6.14 (5.67, 6.61)	5.46 (5.17, 5.76)	2.00 (4.00, 4188)	9.25E-02
	Aa	4.91 (4.73, 5.10)	5.50 (5.08, 5.92)	5.04 (4.53, 5.56)	5.24 (4.72, 5.76)	5.43 (4.75, 6.10)	2.07 (4.00, 3963)	8.14E-02
sDSC	M	0.819 (0.815, 0.823)	0.780 (0.776, 0.784)	0.810 (0.805, 0.816)	0.861 (0.856, 0.865)	0.813 (0.807, 0.819)	153.20 (4.00, 4485)	5.19E-123
	AI	0.869 (0.866, 0.872)	0.877 (0.874, 0.880)	0.874 (0.870, 0.878)	0.870 (0.866, 0.874)	0.861 (0.856, 0.867)	8.27 (4.00, 4366)	1.22E-06
	Aa	0.880 (0.876, 0.883)	0.882 (0.880, 0.885)	0.885 (0.881, 0.889)	0.883 (0.880, 0.887)	0.866 (0.861, 0.870)	11.30 (4.00, 4377)	3.74E-09
VRI	M	0.258 (0.252, 0.264)	0.313 (0.306, 0.321)	0.265 (0.258, 0.273)	0.228 (0.220, 0.235)	0.269 (0.259, 0.280)	65.46 (4.000, 4446)	7.02E-54
	AI	0.217 (0.212, 0.223)	0.203 (0.198, 0.208)	0.215 (0.208, 0.222)	0.235 (0.227, 0.242)	0.231 (0.223, 0.239)	15.53 (4.00, 4347)	1.28E-12
	Aa	0.207 (0.201, 0.213)	0.197 (0.192, 0.202)	0.199 (0.193, 0.206)	0.215 (0.208, 0.222)	0.223 (0.215, 0.231)	9.87 (4.00, 4354)	7.67E-08
Time (min)	M	61.3 (56.7, 65.8)	64.4 (59.6, 69.1)	59.5 (53.3, 65.8)	87.2 (79.8, 94.7)	67.9 (60.7, 75.2)	10.16 (4.00, 397)	7.65E-08
	Aa	12.8 (11.7, 13.9)	9.4 (8.9, 9.9)	13.8 (12.7, 14.9)	15.4 (14.0, 16.8)	11.5 (9.6, 13.4)	28.66 (4.00, 359.4)	2.91E-17

All values are presented as mean value with 95% confidence interval. Welch's ANOVA followed by Games-Howell post hoc corrections was used to test for significant differences across the five centers. All statistical tests were two-sided. Abbreviations: M, Manual delineation; AI, AI delineation; Aa, AI-assisted delineation; vDSC, volumetric Dice Similarity Coefficient; HD95, 95th percentile Hausdorff Distance; sDSC, surface Dice Similarity Coefficient; VRI, Volumetric Revision Index.

Supplementary Table 2. Comparison of vDSC in initial and later delineations by less-experienced physicians

Name	Metrics*	Initial	Later	Name	Metrics*	Initial	Later
	OAR number	55 (0)	53 (2)		OAR number	154 (0)	175 (1)
CKJ	Median (Q1, Q3)	0.904 (0.859,	0.922 (0.825,	**LZ**	Median (Q1, Q3)	0.895 (0.811,	0.892 (0.791,
	P-value	0.47			P-value	0.46	
	OAR number	32 (1)	41 (3)		OAR number	98 (1)	99 (0)
CSY	Median (Q1, Q3)	0.902 (0.819,	0.901 (0.816,	**MJ**	Median (Q1, Q3)	0.916 (0.853,	0.921 (0.824,
	P-value	0.75			P-value	0.27	
	OAR number	110 (0)	110 (0)		OAR number	152 (2)	161 (4)
CY	Median (Q1, Q3)	0.886 (0.797,	0.868 (0.778,	**QLF**	Median (Q1, Q3)	0.902 (0.818,	0.889 (0.808,
	P-value	0.042			P-value	0.067	
	OAR number	142(1)	154 (0)		OAR number	120 (1)	152 (2)
HR	Median (Q1, Q3)	0.902 (0.819,	0.901 (0.816,	**SLL**	Median (Q1, Q3)	0.783 (0.646,	0.880 (0.714,
	P-value	0.75			P-value	0.00014	
	OAR number	54 (1)	54 (1)		OAR number	187 (0)	176 (0)
HJ	Median (Q1, Q3)	0.897 (0.759,	0.876 (0.762,	**SYH**	Median (Q1, Q3)	0.895 (0.814,	0.894 (0.825,
	P-value	0.38			P-value	0.78	
	OAR number	165 (0)	141 (2)		OAR number	153 (1)	175 (1)
HXH	Median (Q1, Q3)	0.816 (0.648,	0.828 (0.686,	**WJL**	Median (Q1, Q3)	0.877 (0.784,	0.846 (0.745,
	P-value	0.019			P-value	0.029	
	OAR number	77 (0)	74 (3)		OAR number	152 (2)	165 (0)
JXL	Median (Q1, Q3)	0.932 (0.871,	0.93 (0.874,	**WT**	Median (Q1, Q3)	0.895 (0.835,	0.895 (0.845,
	P-value	0.28			P-value	0.73	
	OAR number	150 (4)	173 (3)		OAR number	87 (1)	110 (0)
LBH	Median (Q1, Q3)	0.879 (0.796,	0.883 (0.796,	**XZY**	Median (Q1, Q3)	0.848 (0.769,	0.810 (0.737,
	P-value	0.86			P-value	0.052	
	OAR number	363 (0)	383 (2)		OAR number	164 (1)	185 (2)
LC	Median (Q1, Q3)	0.897 (0.832,	0.884 (0.828,	**YCS**	Median (Q1, Q3)	0.878 (0.788,	0.887 (0.799,
	P-value	0.27			P-value	0.61	
	OAR number	106 (4)	88 (2)		OAR number	142 (1)	151 (3)
LJ	Median (Q1, Q3)	0.900 (0.823,	0.896 (0.813,	**ZL**	Median (Q1, Q3)	0.876 (0.781,	0.822 (0.825,
	P-value	0.44			P-value	0.013	
LM	OAR number	187 (0)	187 (0)		OAR number	153 (1)	150 (4)
	Median (Q1, Q3)	0.890 (0.807,	0.878 (0.798,	**ZMY**	Median (Q1, Q3)	0.893 (0.827,	0.892 (0.825,
	P-value	0.38			P-value	0.99	
	OAR number	183 (4)	163 (2)		OAR number	164 (1)	186 (1)
LT	Median (Q1, Q3)	0.877 (0.802,	0.906 (0.802,	**ZZQ**	Median (Q1, Q3)	0.910 (0.851,	0.900 (0.863,
	P-value	0.0074			P-value	0.91	
	OAR number	98 (1)	87 (1)				
LYF	Median (Q1, Q3)	0.852 (0.737,	0.852 (0.723,				
	P-value	0.94					

"Initial" and "later" refer to the two chronological segments of manual contouring, divided based on the median contouring date for each physician. "Dropout/missing" refers to OARs that were not generated by the AI, not delineated by the physicians, or data that was lost. Comparisons were performed using the two-sided Mann-Whitney U test**.** Abbreviation: vDSC, volumetric Dice similarity coefficients; OAR, organs at risk; Q1 and Q3, the first and third quartile.

Supplementary Table 3. Characteristics of physicians involved in manual and AI-assisted delineation

Level of expertise	Center	Physicians	Cases delineated in this study#			Experience in radiotherapy	
			CT scans	Manual delineation	AI-assisted delineation	Years*	Annual patient volume (Thoracic/breast tumor)
Less-experienced	01	CY	20	20	20	3	>100
		LC	68	68	68	2	50-100
		LM	34	34	34	1	<50
	02	HR	34	34	34	5	>100
		HXH	28	28	28	8	<50
		LT	33	33	33	9	<50
		LZ	30	30	30	13	<50
		SLL	28	28	28	5	<50
		SYH	34	34	34	4	<50
		WJL	31	31	31	2	<50
		YCS	33	33	33	5	<50
		ZL	34	34	34	7	<50
	03	HJ	11	11	11	9	<50
		JXL	14	14	14	2	>100
		LJ	18	18	18	11	<50
		LYF	17	17	17	13	<50
		MJ	18	18	18	9	<50
		XZY	18	18	18	5	<50
	04	CKJ	10	10	10	4	50-100
		CSY	7	7	7	15	<50
		ZZQ	34	34	34	4	>100
	05	LBH	30	30	30	13	<50
		QLF	29	29	29	8	<50
		WT	29	29	29	4	<50
		ZMY	29	29	29	2	<50
Experienced	01	CYJ	34	34	34	8	50-100
		WHH	35	35	35	5	>100
		WJ	36	36	36	14	>100
		WX	34	34	34	12	>100
	02	TF	32	32	32	10	>100
	03	LY	14	14	14	9	>100
		WJ	19	19	19	13	>100
		YWX	14	14	14	9	50-100
		ZML	17	17	17	5	>100
	04	HBJ	29	29	29	9	50-100
		MYY	31	31	31	8	50-100
		ZB	27	27	27	10	50-100

Experienced physicians are defined as those with more than 3 years of experience in radiotherapy for over 50 thoracic and breast cancer cases annually, while others are considered less experienced. None of the physicians listed participated in the ground truth generation process. # Delineators generated manual and AI-assisted contours for the same patient's images before and after a washout period. *Years of clinical experience in delineation of target volume and OARs prior to participation in this study.

Supplementary Table 4. Information of software and tools used for delineation in the study

Delineation	Software/tools
Manual and AI-assisted* delineation	PVmed Contouring version 1.0#
AI delineation (auto-segmentation)	Res-SE net (iCurveE)#
Expert A & B	
Center 01	MIM software version 7.3.3; Pinnacle version 9.8
Center 02	Pinnacle version 9.2; Varian Eclipse version 15.6
Center 03	Pinnacle version 16.2
Center 04	Monaco version 5.4
Center 05	Monaco version 5.4
Expert C	PVmed Contouring version 1.0#

*Refers to the manual modification of auto-segmentation results. #PVmed Contouring software (Registration No. 20203210802) and iCurveE (Registration Number: 20253210066) are registered with the National Medical Products Administration (NMPA) in China.

Supplementary Table 5. Patient and imaging data distribution across training, validation, and test cohorts in model development

Patient & Images Information	Training Set	Validation Set	Testing Set	Total
Total images, numbers	497	165	165	827
Sex (%)				
Male	263 (52.9%)	99 (60.0%)	84 (50.9%)	446 (53.9%)
Female	234 (47.1%)	66 (40.0%)	81 (49.1%)	381 (46.1%)
Tube Voltage (kV) (%)				
100	2 (0.4%)	0 (0%)	0 (0%)	2 (0.2%)
120	349 (70.2%)	95 (57.6%)	74 (44.8%)	518 (62.7%)
140	146 (29.4%)	70 (42.4%)	91 (55.2%)	307 (37.1%)
Tube Current (mA)				
Mean (SD)	226.9 (84.08)	223.1 (58.37)	218.9 (47.67)	224.6 (73.37)
Median (Q1, Q3)	217.0 (118.0, 229.0)	217.0 (188.0, 229.0)	217.0 (189.0, 229.0)	217.0 (188.0, 229.0)
Min, Max	54.0, 595.0	83.0, 595.0	83.0, 595.0	54.0, 595.0
Contrast administration				
Non-contrast	153 (30.8%)	35 (21.2%)	38 (23.0%)	226 (27.3%)
Contrast-enhanced	344 (69.2%)	130 (78.8%)	127 (77.0%)	601 (72.7%)
Matrix size				
512x512	497 (100%)	165 (100%)	165 (100%)	827 (100%)
Pixel spacing (mm)				
Mean (SD)	1.13 (0.161)	1.13 (0.168)	1.20 (0.144)	1.14 (0.161)
Median (Q1, Q3)	1.13 (0.99, 1.26)	1.15 (0.99, 1.29)	1.21 (1.11, 1.34)	1.15 (1.02, 1.28)
Min, Max	0.73, 1.37	0.80, 1.37	0.83, 1.37	0.73, 1.37
CT slices				
Median (Q1, Q3)	115 (101, 145)	109 (98, 134)	105 (95, 126)	112 (98, 142)
Min, Max	67, 367	73, 271	84, 230	67, 367
Slice Thickness				
2 mm	1 (0.2%)	0 (0%)	1 (0.6%)	2 (0.2%)
2.5 mm	27 (5.4%)	4 (2.4%)	3 (1.8%)	34 (4.1%)
3 mm	323 (65.0%)	97 (58.8%)	81 (49.1%)	501 (60.6%)
3.75 mm	3 (0.6%)	0 (0%)	0 (0%)	3 (0.4%)
5 mm	143 (28.8%)	64 (38.8%)	80 (48.5%)	287 (34.7%)

Abbreviations: OARs, organs at risk; SD, standard deviation; Q1, first quartile; Q3, third quartile.

Supplementary Table 6. Technical specifications for iCurveE preprocessing, architecture, and training.

Category	Parameter / Component	Description / Specification
Data Preprocessing	Standardization Strategy	Resampling to 1×1 mm pixel spacing; Nearest-neighbor interpolation
	ROI Extraction	1. Fixed-threshold segmentation (≥ -300 HU) 2. Connected-component analysis (thoracic region extraction) 3. Cropping with external margin expansion
	Intensity Normalization	1. Clipping to soft-tissue window (Width: 400, Level: 50) 2. Min-max normalization to range [0, 1]
	Final Input Resolution	384 × 384 pixels (Resampled via nearest-neighbor interpolation)
Data Augmentation	Techniques & Parameters	• Gaussian noise: Standard deviation (SD) = 0.01 • Gamma correction: gamma ∈ [0.6, 1.5] • Random rotation: ±30° • Random shift: ±10% • Random scale: ±10%
	Implementation Policy	Independent application with probability P = 0.5
Model Architecture	Network Backbone	U-Net style encoder-decoder integrating Res-SE blocks
	Fundamental Unit Specification (Res-SE Block)	1. Residual Bottleneck: Three conv layers (1×1, 3×3, 1×1) with residual connection; each followed by BN and ReLU. 2. SE Block: Global pooling → FC layer → ReLU & Sigmoid activation for channel-wise recalibration.
	Initial Block	3×3 Conv – Batch Normalization (BN) – ReLU (First layer of Encoder)
	Encoder Path	5 Res-SE blocks, each followed by a 2×2 max-pooling layer for downsampling (Feature channels: 64 → 1024)
	Decoder Path	4 Res-SE blocks combined with transposed convolutional layers for upsampling (Feature channels: 1024 → 64)
	Feature Fusion	Skip connections merging multi-scale features from Encoder to Decoder
	Output Processing	1×1 Convolution →Softmax activation → Argmax
	Output Dimensions	12-channel probability vector (11 OARs + 1 background)
	Model Scale	14.13 million trainable parameters
Post-processing	Hole Filling	Morphological dilation and erosion (Threshold: < 150 pixels)
	Component Selection	Retention of the largest connected component for each organ
	Spatial Restoration	Inverse resampling to original image dimensions using recorded coordinates
Training & Implementation	Hardware & Framework	NVIDIA GTX 1080-Ti (11GB); TensorFlow 1.13.1
	Optimizer Configuration	Adam: alpha=0.001, beta1=0.9, beta2=0.999, epsilon=1e-7
	Loss Function	Dice loss (Loss components for missing OAR labels excluded)
	Hyperparameters	Batch size: 64; Initial learning rate: 0.01; Max epochs: 200
	Early Stopping Strategy	Stop if validation loss shows no improvement ($\Delta < 0.001$) over 15 epochs
	Model Selection Policy	Checkpoint with the highest average vDSC on validation set

Abbreviations: FC, Fully Connected; HU, Hounsfield Units; OAR, Organ at Risk; ReLU, Rectified Linear Unit; ROI, Region of Interest; SD, Standard Deviation; SE, Squeeze-and-Excitation; vDSC, Volumetric Dice Similarity Coefficient.

Supplementary Table 7. Performance metrics of the iCurveE model on the development datasets (training, validation, and test)

Organs at Risk (OARs)	Training Set		Validation Set		Test Set	
	vDSC	HD95	vDSC	HD95	vDSC	HD95
Spinal Cord	497 (0)		165 (0)		165 (0)	
Median (Q1, Q3)	0.856 (0.836, 0.880)	1.135 (0.996, 1.270)	0.849 (0.831, 0.878)	1.160 (0.993, 1.334)	0.833 (0.812, 0.855)	1.277 (1.140, 1.367)
Left Humeral head	497 (0)		163 (2)		165 (0)	
Median (Q1, Q3)	0.934 (0.926, 0.942)	1.137 (0.996, 1.275)	0.939 (0.930, 0.950)	1.135 (0.977, 1.281)	0.928 (0.914, 0.943)	1.270 (1.122, 1.367)
Right Humeral head	497 (0)		163 (2)		165 (0)	
Median (Q1, Q3)	0.935 (0.926, 0.942)	1.145 (0.996, 1.297)	0.939 (0.929, 0.949)	1.139 (0.977, 1.291)	0.931 (0.916, 0.944)	1.211 (1.074, 1.365)
Aorta	497 (0)		165 (0)		165 (0)	
Median (Q1, Q3)	0.943 (0.933, 0.950)	1.133 (0.994, 1.261)	0.940 (0.927, 0.951)	1.145 (0.986, 1.286)	0.928 (0.919, 0.939)	1.207 (1.111, 1.350)
Trachea	496 (1)		165 (0)		165 (0)	
Median (Q1, Q3)	0.905 (0.892, 0.918)	1.137 (1.000, 1.270)	0.886 (0.874, 0.898)	1.162 (1.002, 1.314)	0.881 (0.870, 0.891)	1.242 (1.135, 1.363)
Bronchial tree	497 (0)		165 (0)		165 (0)	
Median (Q1, Q3)	0.905 (0.891, 0.918)	1.139 (1.007, 1.280)	0.898 (0.873, 0.916)	1.174 (1.002, 1.365)	0.890 (0.867, 0.901)	1.268 (1.126, 1.367)
Esophagus	497 (0)		165 (0)		165 (0)	
Median (Q1, Q3)	0.866 (0.840, 0.887)	1.133 (0.991, 1.260)	0.858 (0.824, 0.888)	1.145 (0.986, 1.286)	0.838 (0.808, 0.864)	1.207 (1.111, 1.348)
Left Lung	497 (0)		165 (0)		165 (0)	
Median (Q1, Q3)	0.974 (0.971, 0.977)	1.133 (0.991, 1.261)	0.972 (0.966, 0.977)	1.174 (1.042, 1.331)	0.964 (0.960, 0.969)	1.207 (1.111, 1.343)
Right Lung	497 (0)		165 (0)		165 (0)	
Median (Q1, Q3)	0.981 (0.978, 0.983)	1.129 (0.987, 1.258)	0.976 (0.970, 0.980)	1.162 (1.002, 1.304)	0.969 (0.966, 0.973)	1.207 (1.111, 1.343)
Heart	497 (0)		165 (0)		165 (0)	
Median (Q1, Q3)	0.982 (0.980, 0.984)	0.988 (0.000, 1.179)	0.977 (0.973, 0.980)	1.160 (0.990, 1.300)	0.973 (0.969, 0.975)	1.211 (1.114, 1.350)
Liver	484 (13)		160 (5)		162 (3)	
Median (Q1, Q3)	0.985 (0.982, 0.987)	0.000 (0.000, 0.000)	0.979 (0.974, 0.983)	1.155 (0.985, 1.301)	0.974 (0.969, 0.978)	1.207 (1.112, 1.347)

The establishment of the ground truth is based on the consensus of three radiation oncologists, each with over 15 years of experience in radiotherapy, following the delineation guidelines from RTOG 1106. For each OAR within each dataset, the number of available contours is listed, with the number of missing contours shown in parentheses. **Abbreviations:** OAR, organ at risk; vDSC, volumetric Dice similarity coefficient; HD95, 95th percentile Hausdorff distance in millimeters (mm); Q1, first quartile; Q3, third quartile.

Supplementary Table 8. Metrics definitions and calculation formulas

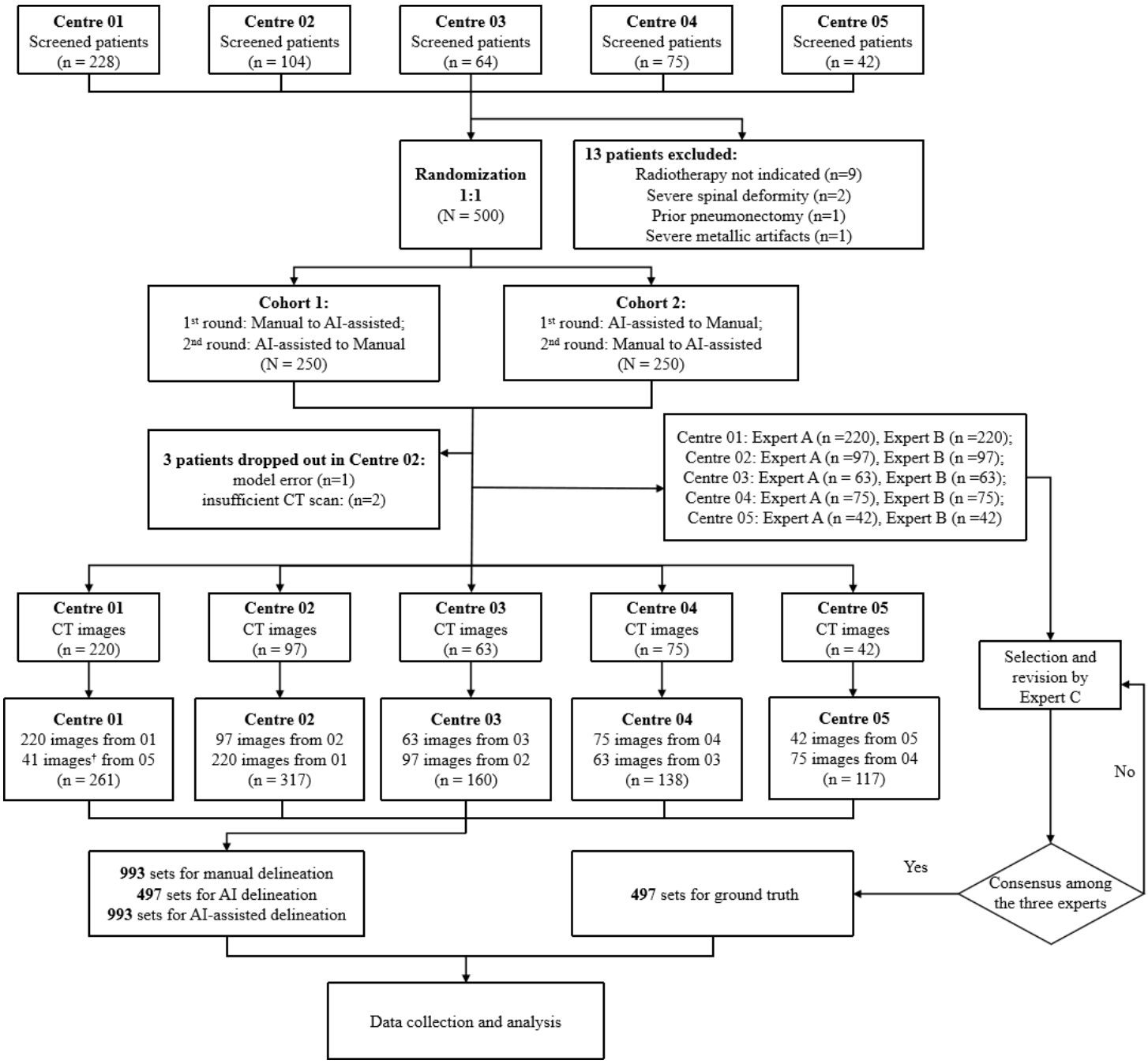
Metric	Definition
Volume Dice similarity coefficient, vDSC	$vDSC = 2 \times V_{A \cap B} / (V_A + V_B)$
95th percentile Hausdorff Distance, HD95	HD95(A, B) = max (h95(A, B), h95(B, A)), where h95(A, B) is the 95th percentile of the shortest distances from all points on surface A to surface B, and vice-versa for h95(B, A).
Surface Dice similarity coefficient, sDSC	$sDSC = (\|S(A) \cap S(B)\tau\| + \|S(B) \cap S(A)\tau\|) / (\|S(A)\| + \|S(B)\|)$, where S(A) and S(B) are the sets of points on the surfaces of A and B, $S(B)\tau$ represents the points on surface B that are within the tolerance τ of surface A, and $S(A)\tau$ represents the points on surface A that are within the tolerance τ of surface B.
Contouring time (min)	Manual delineation time is recorded from the moment the CT image is displayed on the contouring platform. The AI-assisted delineation time includes the runtime of the auto-segmentation model, the transfer of auto-segmentation results to the contouring platform, and the subsequent manual modification period.
Time efficiency improvement rate (%)	Time efficiency improvement rate = (manual delineation time - AI-assisted delineation time)/ manual delineation time*100%
Volumetric revision index, VRI*	$[(V_A - V_{A \cap B}) + (V_B - V_{A \cap B})] / V_A$
Recall, Rec	$Rec = V_{A \cap B} / V_A$
Precision, Pre	$Pre = V_{A \cap B} / V_B$
Relative volume difference, RVD	$RVD = \left\| V_A - V_B \right\| / V_A$
Subjective assessment of AI delineation using a Likert scale	Evaluated on a 1-5 Likert scale: 1 - Strongly dissatisfied, 2 - Dissatisfied, 3 - Neutral, 4 - Satisfied, 5 - Strongly satisfied.

In the formulas, A represents the ground truth and B represents the manual, AI or AI-assisted delineation, while V_A and V_B represent their respective volumes. A tolerance of τ=2 mm was used in this study. *When the delineated volume by the physician closely matches the ground truth in size but differs significantly in shape, RVD may not accurately reflect the expert's modification effort. To address this, we introduced the VRI parameter, which considers both the added and removed volumes during modifications.

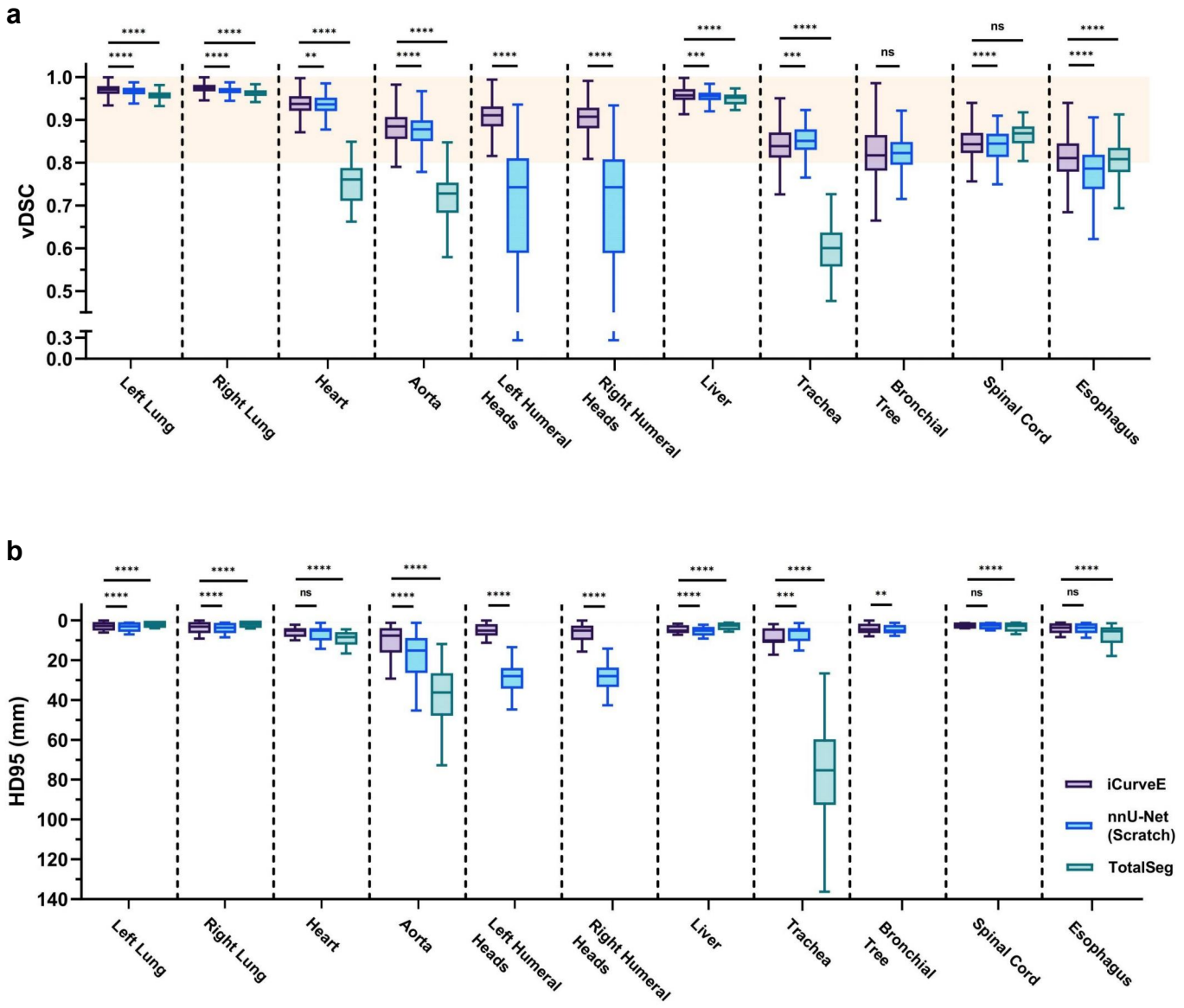
Supplementary table 9. Summary of adverse events (AEs) and serious adverse events (SAEs)

Number (percent)	Center 01	Center 02	Center 03	Center 04	Center 05
AEs	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
SAEs	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

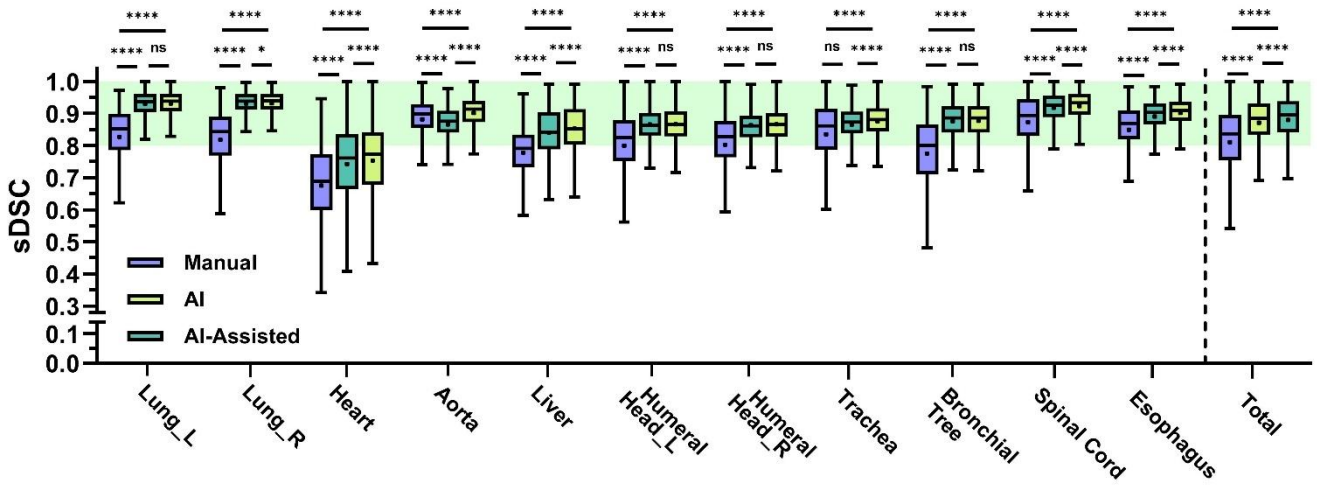
AEs were defined as unfavorable medical occurrences during CT imaging, such as a significant allergic reaction or severe anxiety, that led to scan interruption. SAEs were defined as events resulting in significant health deterioration or life-threatening injury. No such events were recorded during the study.



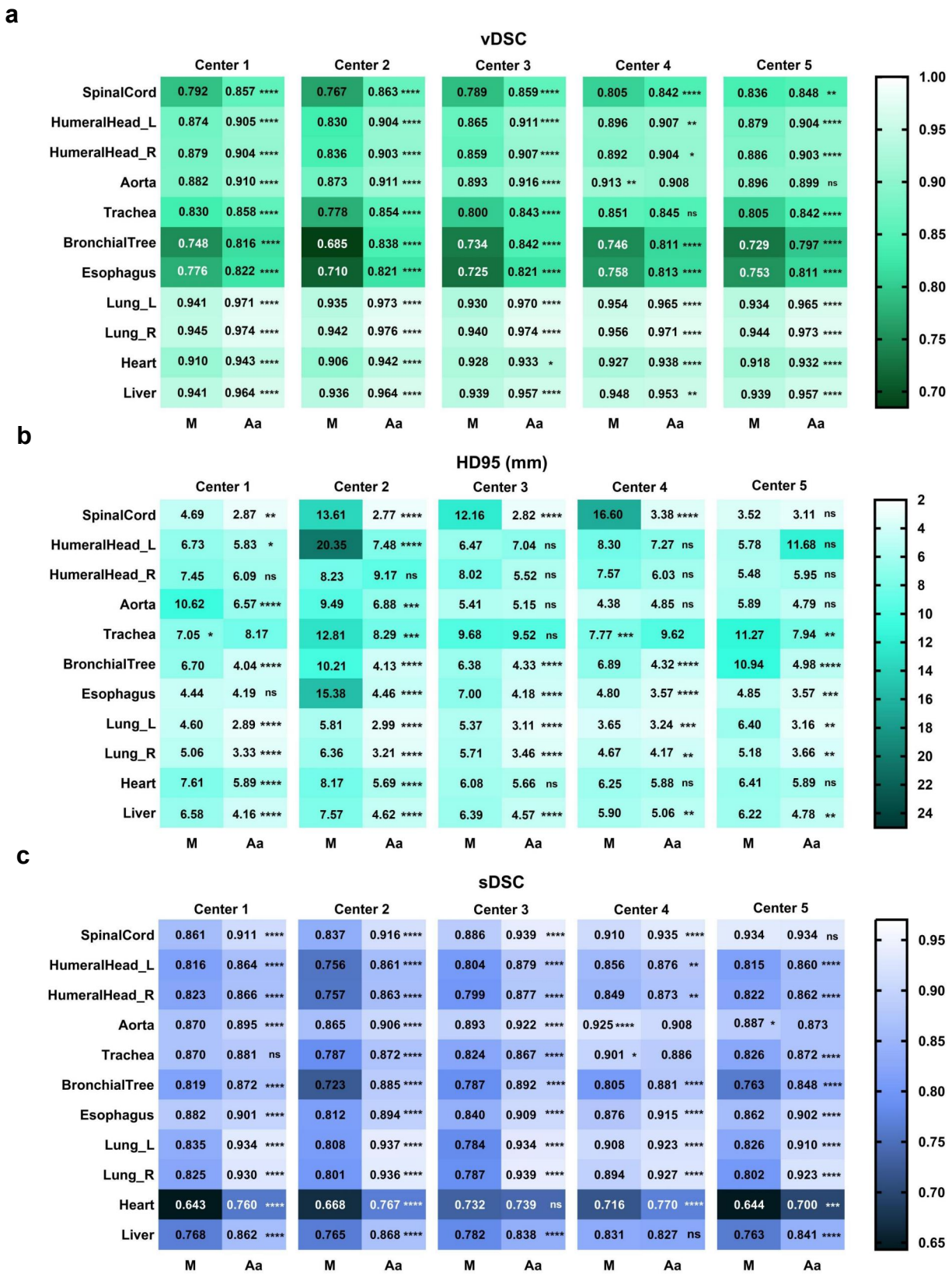
Supplementary Figure 1 | CONSORT flow diagram of CT image inclusion, delineation, and data analysis.
Experts A and B were independent across centers. Expert A, B, and C were not involved in generating manual and AI-assisted delineations. †One image from Center 05 failed to transfer to Center 01, resulting in its manual, AI, and AI-assisted delineations being performed only at Center 05.



Supplementary Figure 2 | Performance benchmark of iCurveE against the nnU-Net and TotalSegmentator models.
(a) Comparison of the volumetric Dice Similarity Coefficient (vDSC) between the iCurveE (n = 497 delineation sets), nnU-Net (n = 497 delineation sets), and TotalSegmentator (n = 497 delineation sets) models across 11 organs at risk (OARs). The orange shaded area highlights the high-accuracy range (vDSC ≥ 0.80). **(b)** Comparison of the 95% Hausdorff Distance (HD95) between the three models. The nnU-Net model was trained from scratch on the same internal development cohort, whereas the pre-trained open-source TotalSegmentator (v2.4.0) model was utilized.[1] The segmentation scope of TotalSegmentator does not include the left/right humeral heads and the bronchial tree. Box plot definitions: center line, median; point, mean; box limits, interquartile ranges (IQR); whiskers, 1.5×IQR, with precise sample sizes (n, number of OARs) for each box provided in the Source Data file. Statistical comparisons were performed using two-sided Wilcoxon signed-rank tests. Statistical significance was determined using two-sided Wilcoxon signed-rank tests. $^{**}p < 0.01$; $^{***}p < 0.001$; $^{****}p < 0.0001$; ns, not significant. Source data and exact p-values are provided as a Source Data file.

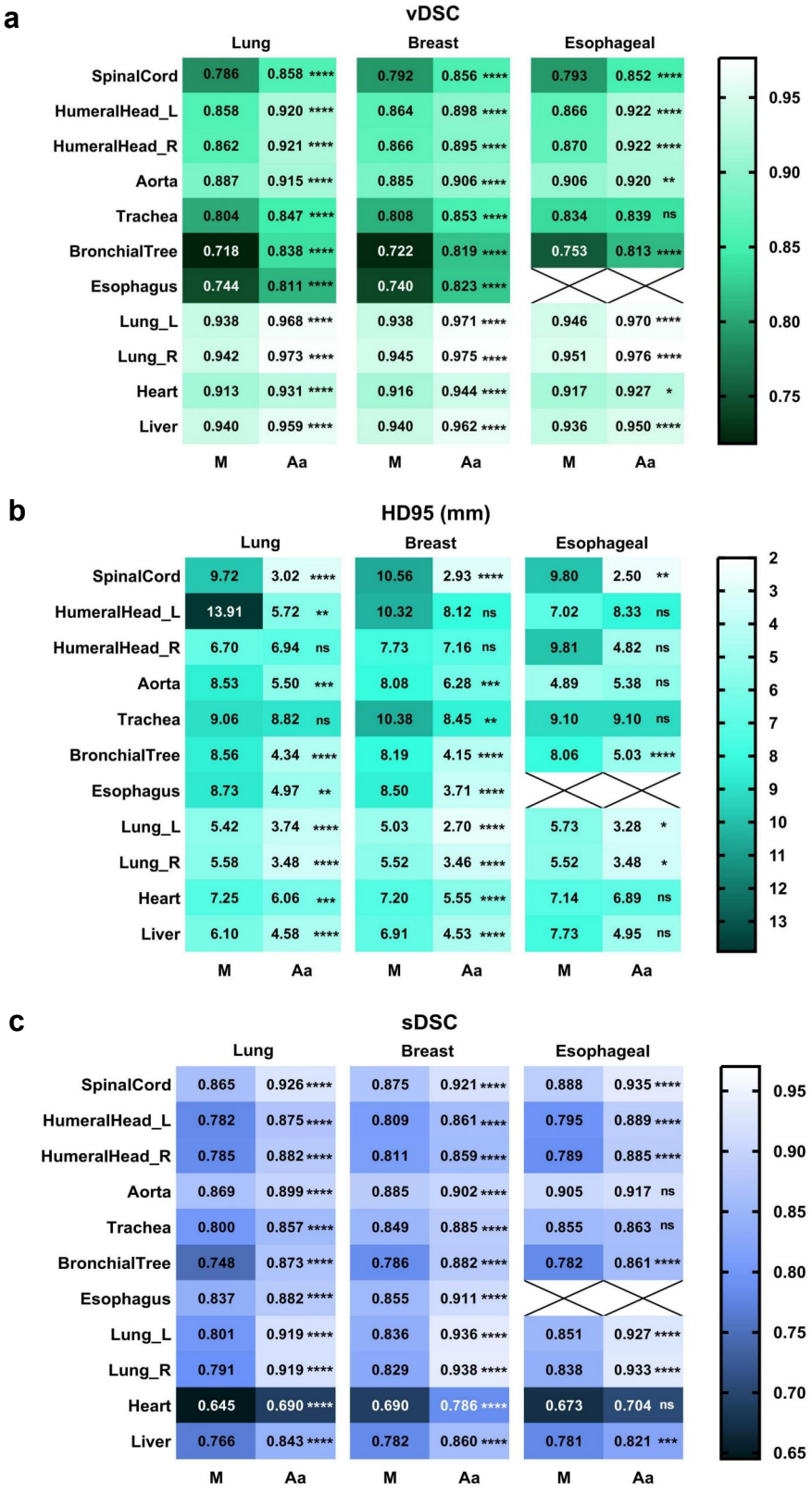


Supplementary Figure 3 | Comparison of delineation accuracy by surface Dice Similarity Coefficient (sDSC). Paired comparison of delineation accuracy for manual (n = 993 delineation sets), AI (n = 497 delineation sets), and AI-assisted (n = 993 delineation sets) methods measured by the sDSC. Statistical significance was assessed using two-sided Friedman tests followed by Dunn's multiple comparisons test. Box plot definitions: center line, median; point, mean; box limits, interquartile ranges (IQR); whiskers, 1.5×IQR, with precise sample sizes (n, number of OARs) for each box provided in the Source Data file. The green shaded area highlights the high-accuracy range (sDSC ≥ 0.80). Statistical significance is denoted by asterisks: *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001; ns, not significant (p > 0.05). Source data and exact p-values are provided as a Source Data file.

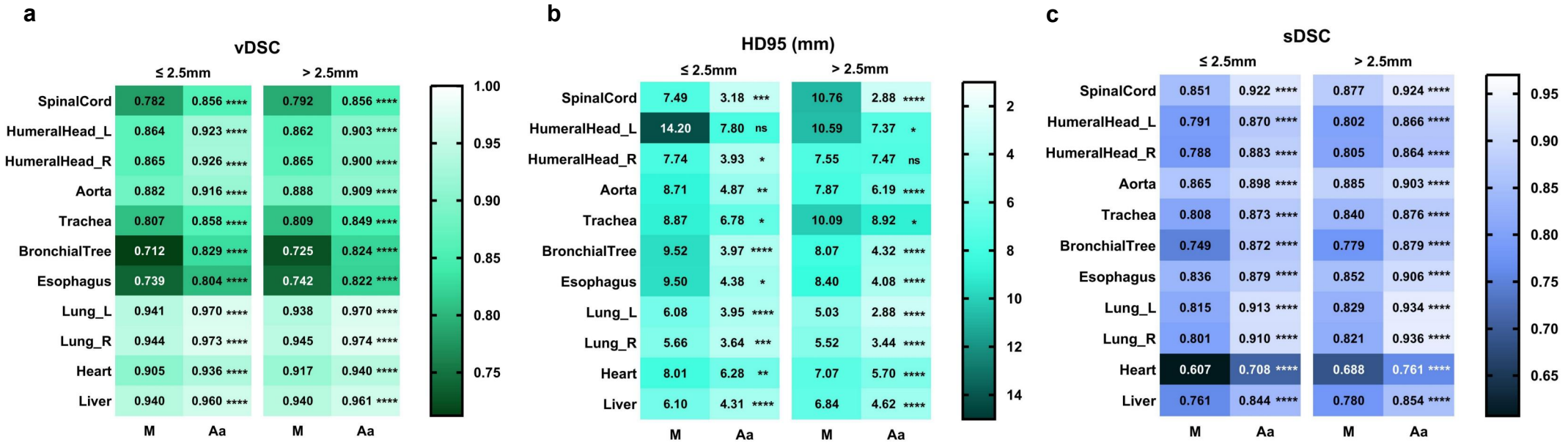


Supplementary Figure 4 | Paired comparison of manual and AI-assisted delineation performance in different centers.

Heatmaps comparing the performance of manual (M) and AI-assisted (Aa) delineation for 11 OARs across the five participating centers (n = 261, 317, 160, 138, and 117 delineation sets for Centers 1–5, respectively). Performance was evaluated using **(a)** volumetric Dice similarity coefficient (vDSC), **(b)** 95th percentile Hausdorff Distance (HD95), and **(c)** surface DSC (sDSC). Cell colors indicate the mean value for each group. Asterisks denote statistically significant superior performance in the paired comparison within each center (i.e., a higher vDSC/sDSC or lower HD95). Statistical comparisons were performed using two-sided Wilcoxon signed-rank tests. Statistical significance is denoted by asterisks: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$; ns, not significant ($p > 0.05$). Source data and exact p-values are provided as a Source Data file.

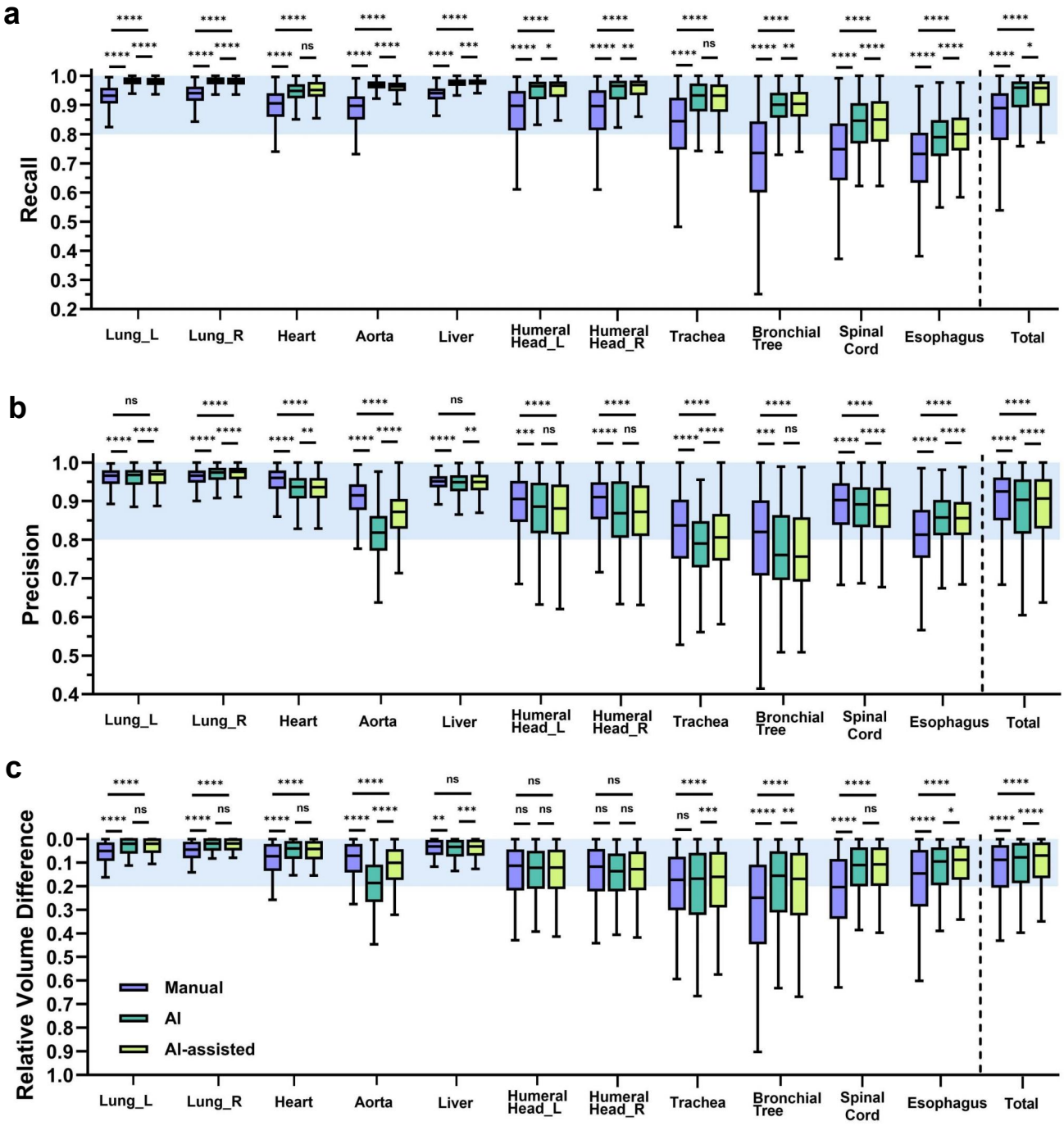


Supplementary Figure 5 | Paired comparison of manual and AI-assisted delineation performance stratified by cancer type. Heatmaps comparing the performance of manual (M) and AI-assisted (Aa) delineation for organs at risk (OARs), stratified by lung (n = 287 delineation sets)), breast (n = 638 delineation sets), and esophageal cancer (n = 68 delineation sets) cohorts. Performance was evaluated using the **(a)** volumetric Dice similarity coefficient (vDSC), **(b)** 95th percentile Hausdorff Distance (HD95), and **(c)** surface DSC (sDSC). Cell colors indicate the mean value for each group. Asterisks denote a statistically significant improvement of the AI-assisted method over the manual method within each cancer type. The crossed-out cells indicate that the esophagus was not evaluated as an OAR in the esophageal cancer cohort. Statistical comparisons were performed using two-sided Wilcoxon signed-rank tests. Statistical significance is denoted by asterisks: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$; ns, not significant ($p > 0.05$). Source data and exact p-values are provided as a Source Data file.

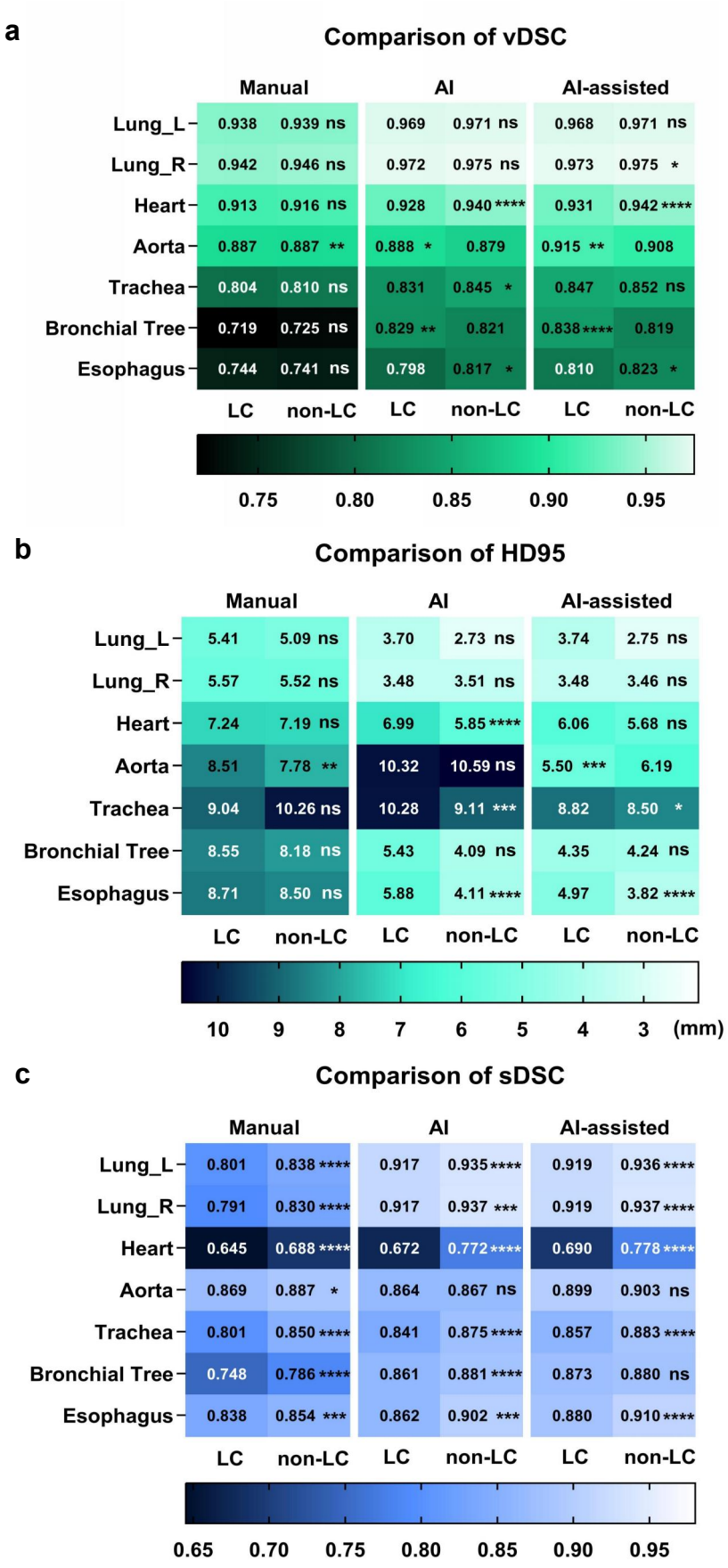


Supplementary Figure 6 | Paired comparison of manual and AI-assisted delineation performance stratified by CT slice thickness.

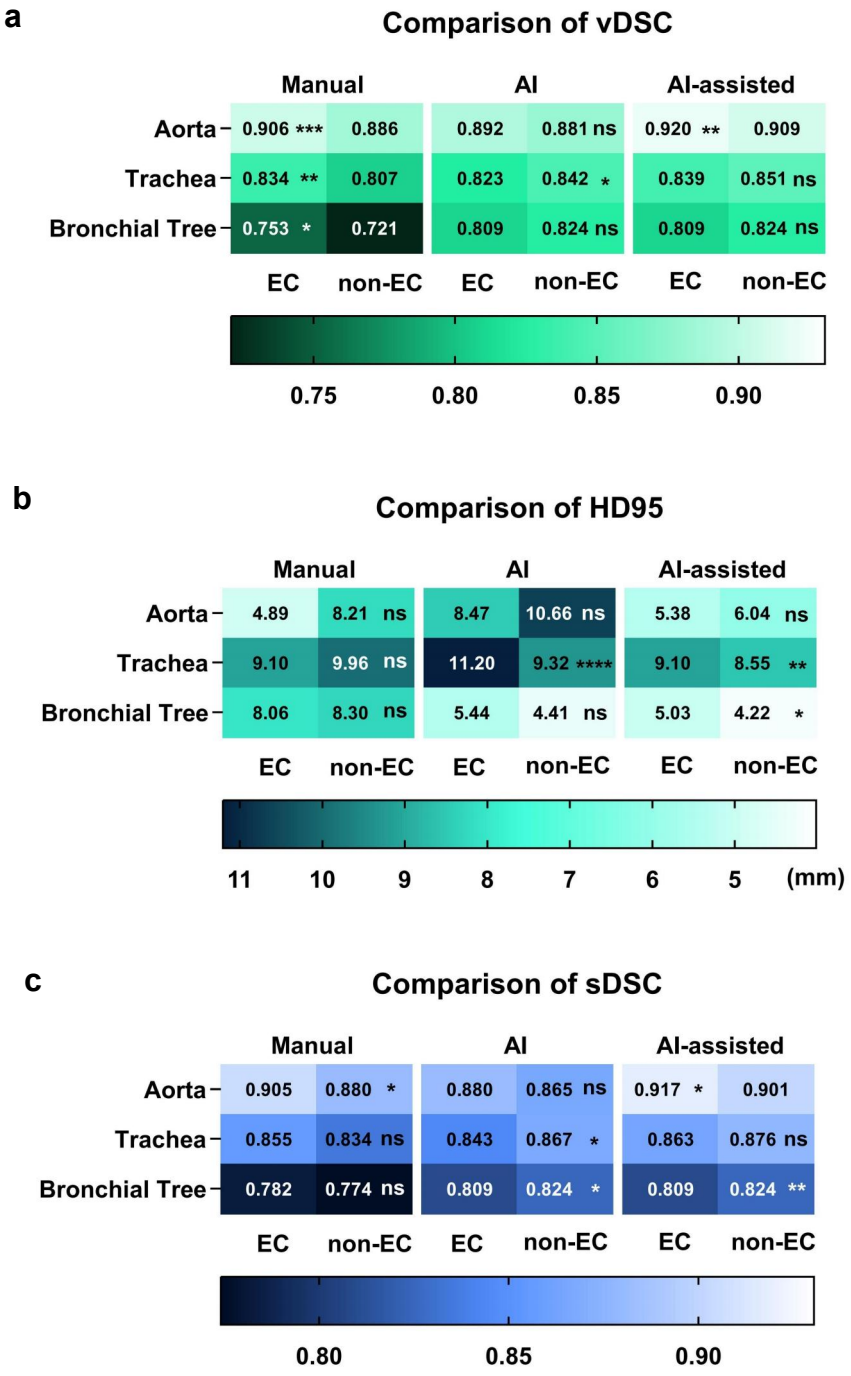
Heatmaps comparing the performance of manual (M) and AI-assisted (Aa) delineation for 11 OARs, stratified by CT slice thickness [≤ 2.5 mm (n = 150 delineation sets) and > 2.5 mm (n = 843 delineation sets)]. Performance was evaluated using **(a)** volumetric Dice similarity coefficient (vDSC), **(b)** 95th percentile Hausdorff Distance (HD95), and **(c)** surface DSC (sDSC). Cell colors indicate the mean value for each group. Asterisks denote statistically significant superior performance in the paired comparison within each slice thickness subgroup (i.e., a higher vDSC/sDSC or lower HD95). Statistical comparisons were performed using two-sided Wilcoxon signed-rank tests. Statistical significance is denoted by asterisks: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$; ns, not significant ($p > 0.05$). Source data and exact p-values are provided as a Source Data file.



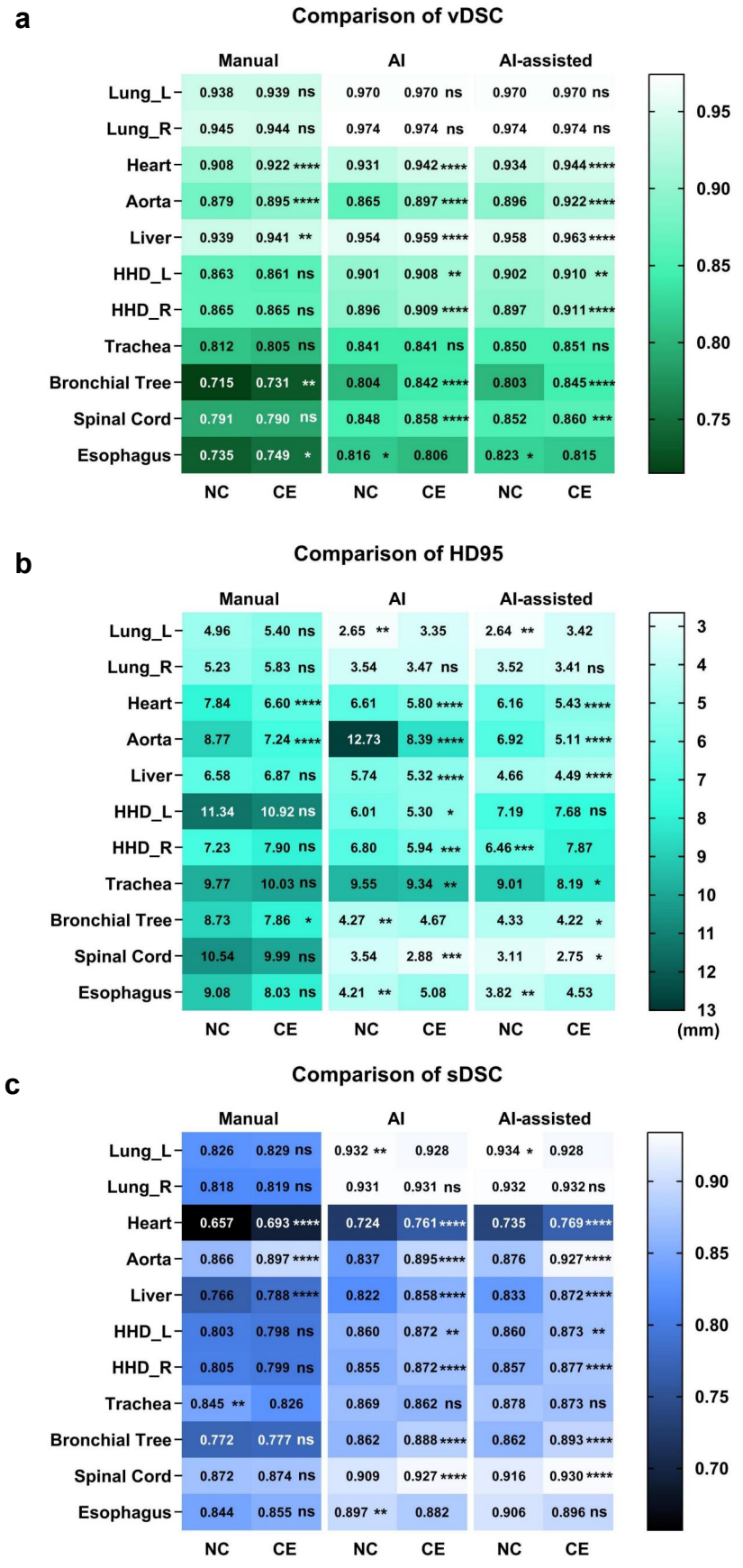
Supplementary Figure 7 | Comparison of delineation accuracy by recall, precision, and relative volume difference (RVD). Paired comparison of delineation accuracy for manual (n = 993 delineation sets), AI (n = 497 delineation sets), and AI-assisted (n = 993 delineation sets) methods measured by the **(a)** Recall, **(b)** Precision, and **(c)** RVD. Statistical significance was assessed using two-sided Friedman tests followed by Dunn's multiple comparisons test. In the box plots, the horizontal line represents the median values. Box plot definitions: center line, median; point, mean; box limits, interquartile ranges (IQR); whiskers, 1.5×IQR, with precise sample sizes (n, number of OARs) for each box provided in the Source Data file. The shaded area highlights the high-performance range (RR ≥ 0.8, PR ≥ 0.8, and RVD ≤ 0.2). Statistical significance is denoted by asterisks: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$; ns, not significant ($p > 0.05$). Source data and exact p-values are provided as a Source Data file.



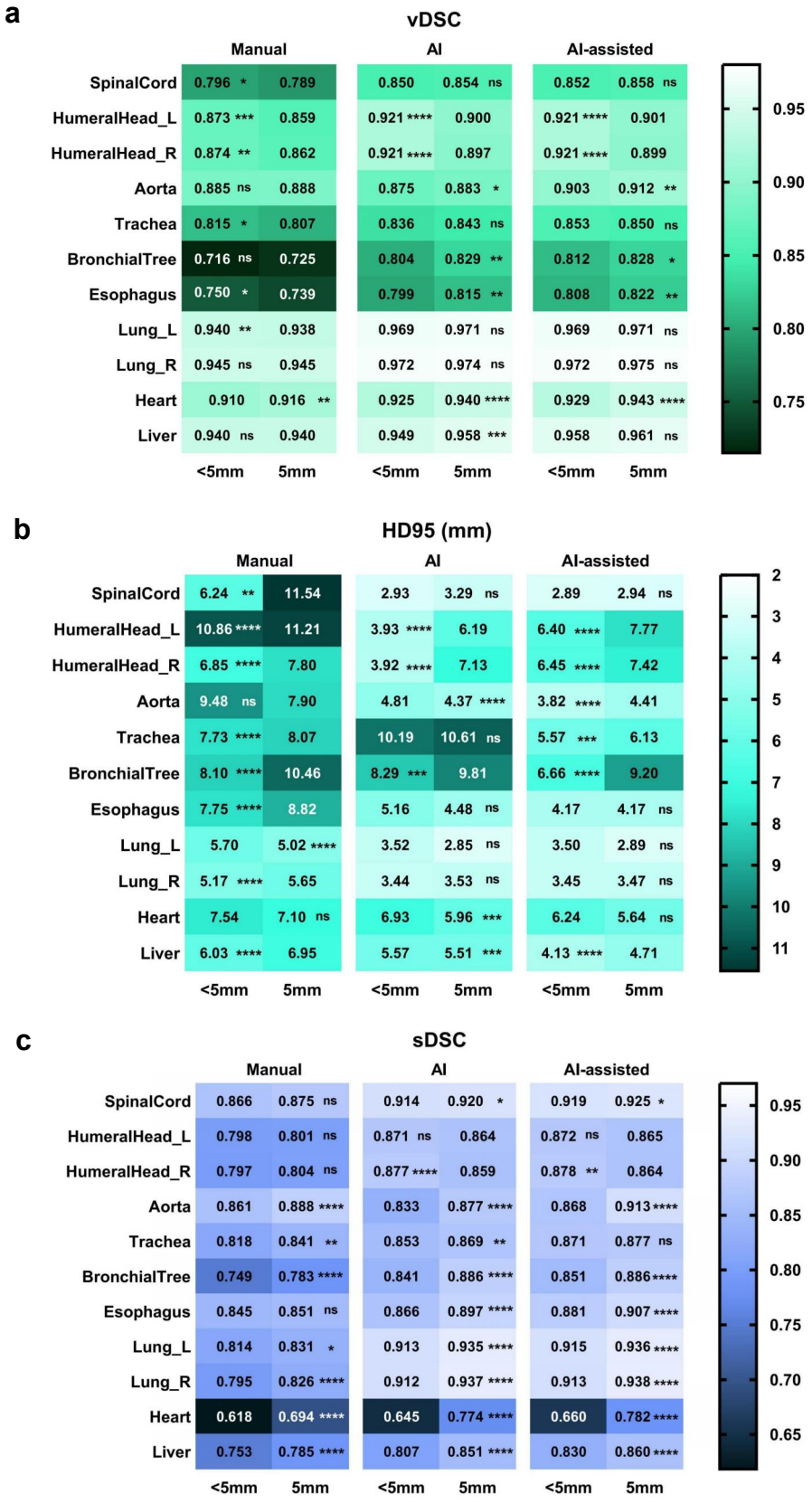
Supplementary Figure 8 | Delineation performance in patients with lung cancer versus non-lung cancer.
Heatmaps comparing delineation performance between patients with lung cancer (LC) (n = 285 delineation sets) and those without (non-LC) (n = 608 delineation sets). The analysis focuses on organs at risk (OARs) particularly susceptible to mass effect from tumor lesions. Performance was evaluated using **(a)** volumetric Dice similarity coefficient (vDSC), **(b)** 95th percentile Hausdorff Distance (HD95), and **(c)** surface DSC (sDSC). Cell colors indicate the mean value for each group. An asterisk in a cell indicates that the corresponding group (LC or non-LC) demonstrated a statistically superior performance over the other group (i.e., a higher vDSC/sDSC or lower HD95). Statistical comparisons were performed using two-sided Mann–Whitney *U* tests. Statistical significance is denoted by asterisks: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$; ns, not significant ($p > 0.05$). Source data and exact p-values are provided as a Source Data file.



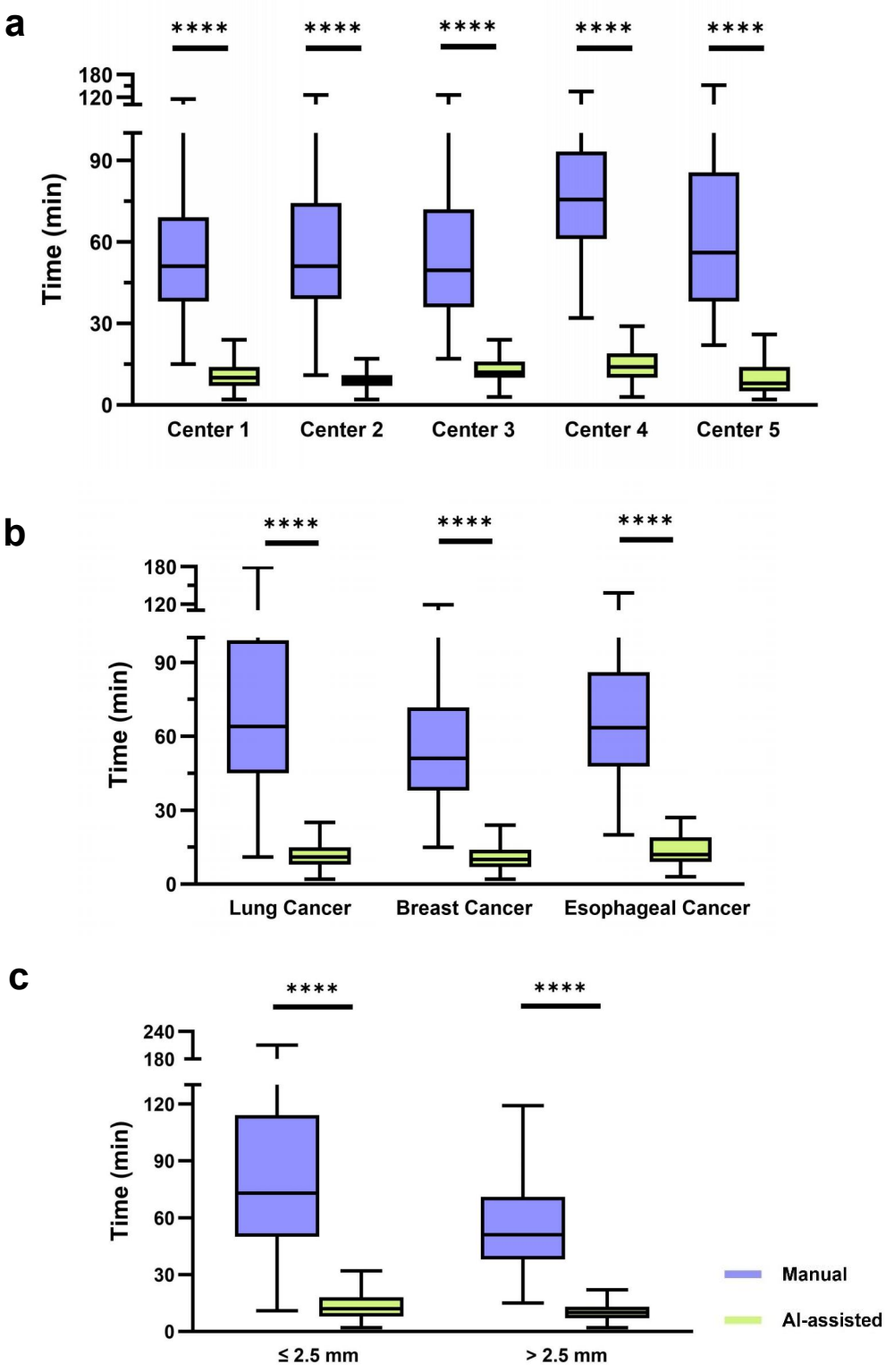
Supplementary Figure 9 | Delineation performance in patients with esophageal cancer versus non-esophageal cancer. Heatmaps comparing delineation performance between patients with esophageal cancer (EC) (n = 67 delineation sets) and those without (non-EC) (n = 826 delineation sets). The analysis focuses on organs at risk (OARs) particularly susceptible to mass effect from tumor lesions. Performance was evaluated using **(a)** volumetric Dice similarity coefficient (vDSC), **(b)** 95th percentile Hausdorff Distance (HD95), and **(c)** surface DSC (sDSC). Cell colors indicate the mean value for each group. An asterisk in a cell indicates that the corresponding group (EC or non-EC) demonstrated a statistically superior performance over the other group (i.e., a higher vDSC/sDSC or lower HD95). Statistical comparisons were performed using two-sided Mann–Whitney *U* tests. Statistical significance is denoted by asterisks: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$; ns, not significant ($p > 0.05$). Source data and exact p-values are provided as a Source Data file.



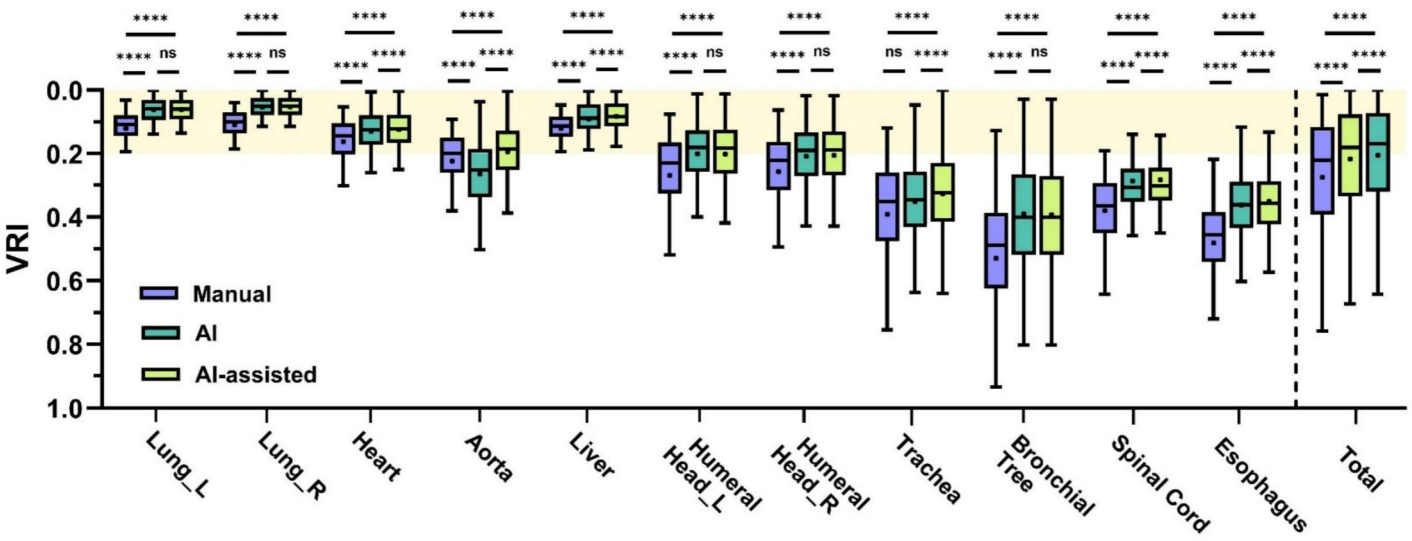
Supplementary Figure 10 | Delineation performance on non-contrast versus contrast-enhanced CT scans.
Heatmaps comparing delineation performance between non-contrast (NC) (n = 492 delineation sets) and contrast-enhanced (CE) (n = 501 delineation sets) CT scans. Performance was evaluated using **(a)** volumetric Dice similarity coefficient (vDSC), **(b)** 95th percentile Hausdorff Distance (HD95), and **(c)** surface DSC (sDSC). Cell colors represent the mean value for each group, where lighter colors indicate superior performance. An asterisk in a cell indicates that the corresponding scan type (NC or CE) yielded a statistically superior performance over the other. Statistical comparisons were performed using two-sided Mann–Whitney *U* tests. Statistical significance is denoted by asterisks: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$; ns, not significant ($p > 0.05$). Abbreviations: HHD_L/R, left/right humeral head. Source data and exact p-values are provided as a Source Data file.



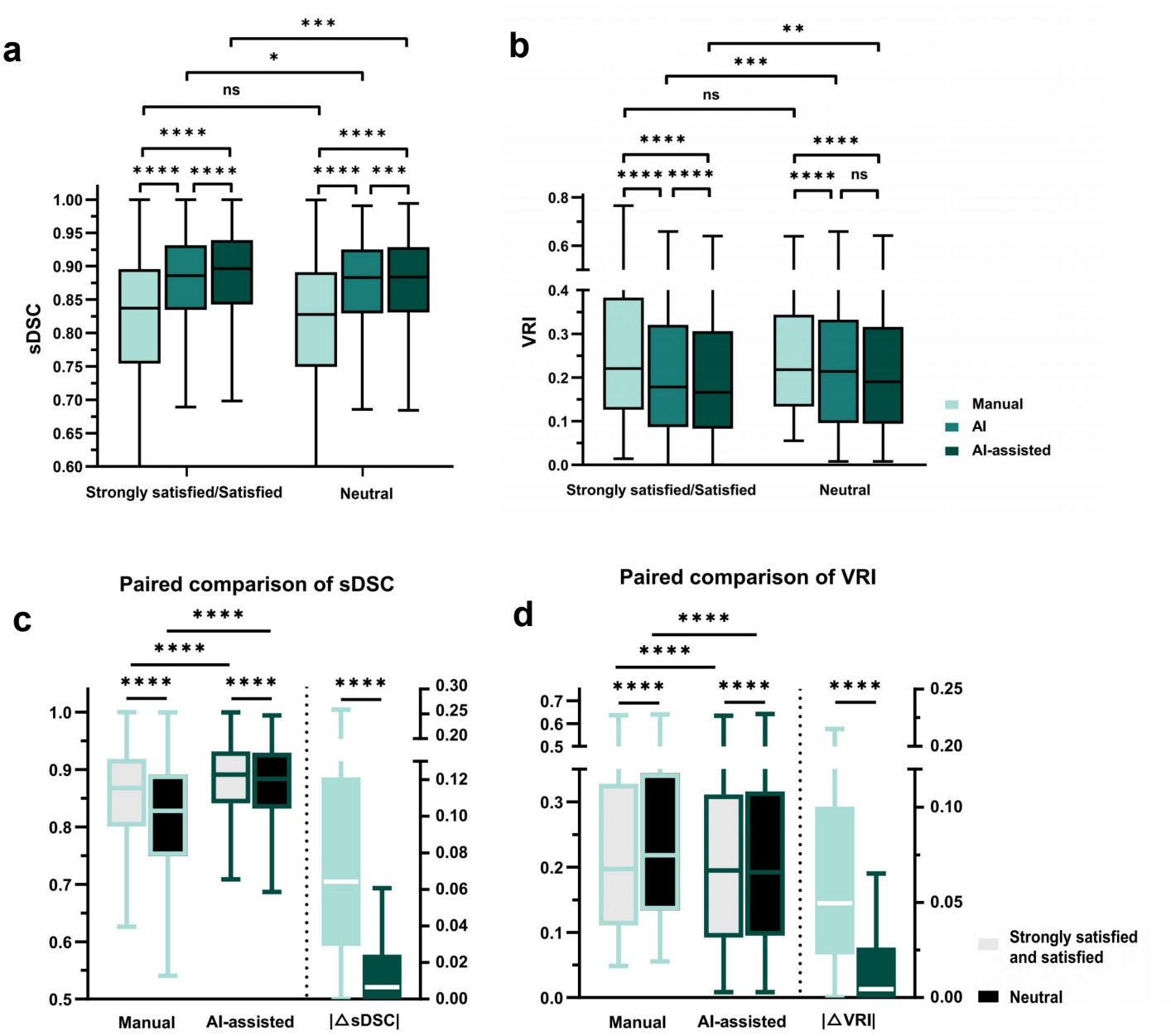
Supplementary Figure 11 | Delineation performance stratified by CT slice thickness.
Heatmaps comparing delineation performance between patient cohorts with thin-slice (<5 mm) (n = 241 delineation sets) and thick-slice (5 mm) (n = 752 delineation sets) CT scans. Performance was evaluated using **(a)** volumetric Dice similarity coefficient (vDSC), **(b)** 95th percentile Hausdorff Distance (HD95), and **(c)** surface DSC (sDSC). Cell colors represent the mean value for each group. Cell colors represent the mean value for each group, where lighter colors indicate superior performance for all three metrics. An asterisk in a cell indicates that the corresponding slice thickness group (<5 mm or 5 mm) demonstrated a statistically superior performance over the other. Statistical comparisons were performed using two-sided Mann–Whitney *U* tests, with significance denoted by asterisks: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$; ns, not significant ($p > 0.05$). Source data and exact p-values are provided as a Source Data file.



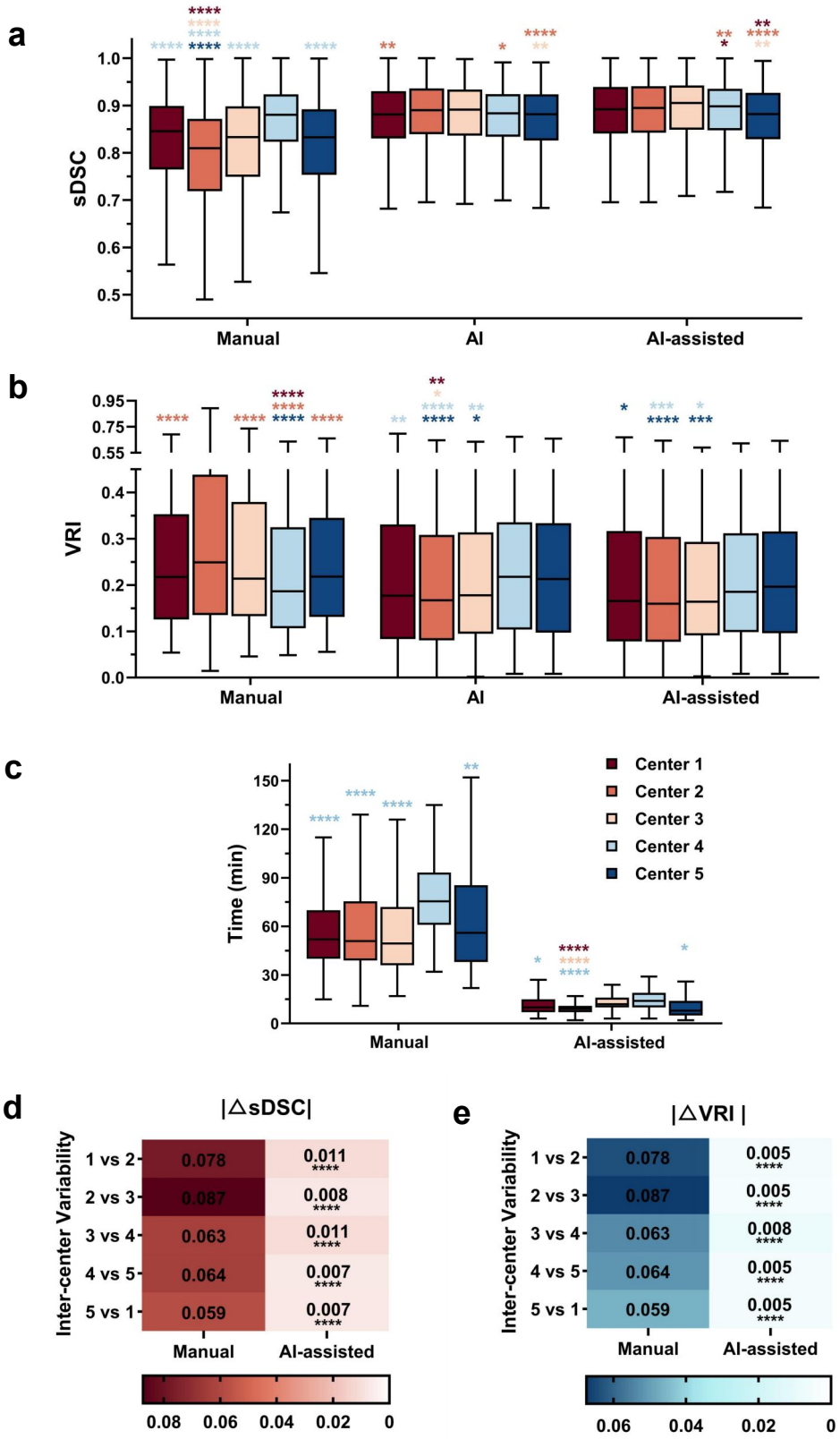
Supplementary Figure 12 | Paired comparison of manual and AI-assisted contouring time in different subgroups. Box plots show the total contouring time for the manual (purple) and AI-assisted (green) methods, stratified by **(a)** participating centers (n = 261, 317, 160, 138, and 117 delineation sets for Centers 1–5, respectively), **(b)** cancer types (n = 287, 638 and 68 delineation sets for lung, breast, and esophageal cancer, respectively), and **(c)** CT slice thickness (n = 150 and 843 delineation sets for ≤ 2.5 mm and > 2.5 mm, respectively). Box plot definitions: center line, median; point, mean; box limits, interquartile ranges (IQR); whiskers, 1.5×IQR, with precise sample sizes (n, number of delineation sets) for each box provided in the Source Data file. Statistical comparisons were performed using two-sided Wilcoxon signed-rank tests, with significance denoted by asterisks: ns, not significant; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$. Source data and exact p-values are provided as a Source Data file.



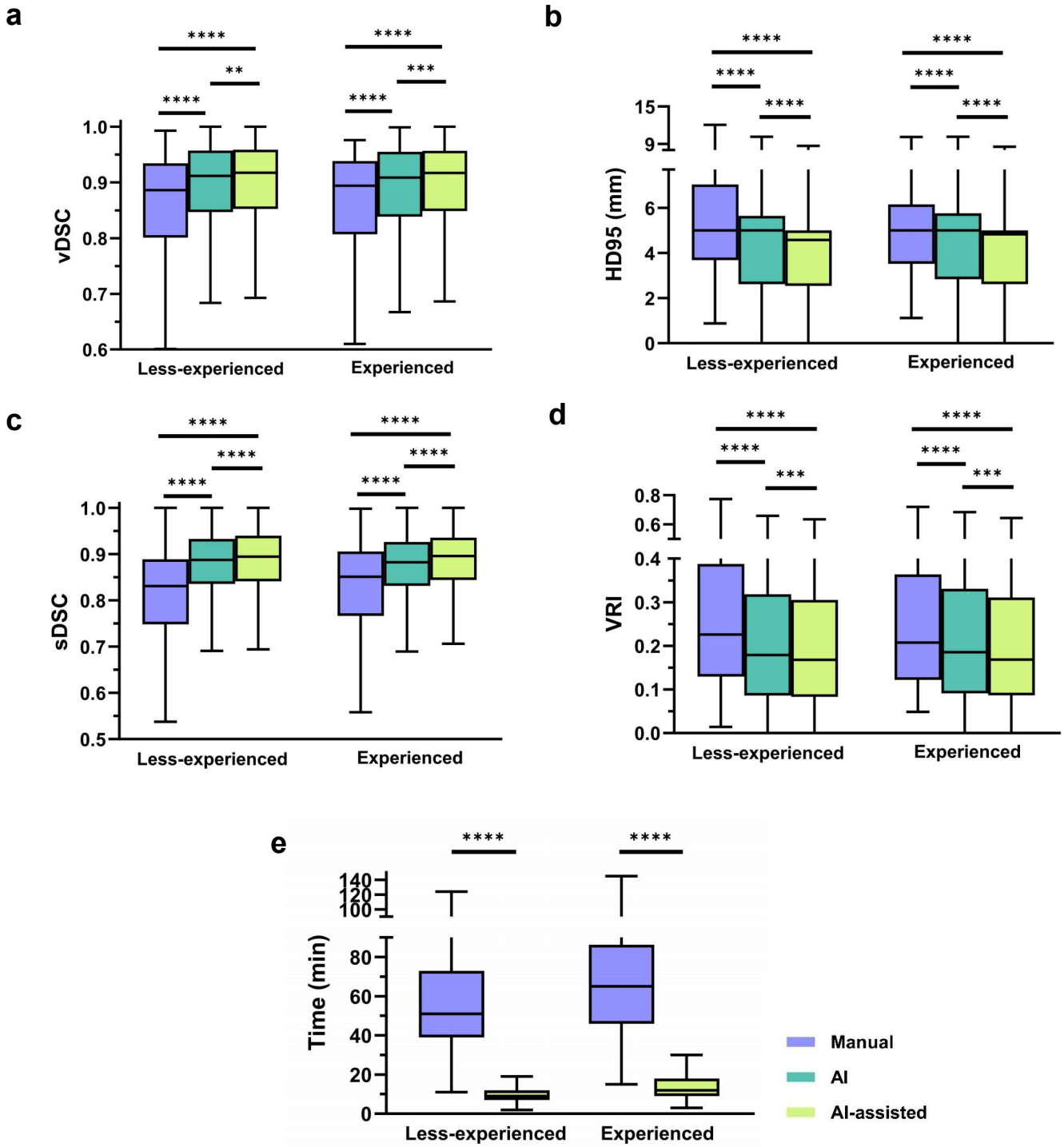
Supplementary Figure 13 | Expert revision workload measured by the Volumetric Revision Index (VRI). Paired comparison of the VRI for manual (n = 993 delineation sets), AI (n = 497 delineation sets), and AI-assisted (n = 993 delineation sets). In the box plots, the horizontal line and the point represent the median and mean values, respectively. Statistical significance was assessed using two-sided Friedman tests followed by Dunn's multiple comparisons test. The box indicates the interquartile range (IQR), and the whiskers extend to the Tukey range (1.5 times the IQR). The yellow shaded area represents the range where the VRI is less than 0.2. Statistical significance is denoted by asterisks: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$; ns, not significant ($p > 0.05$). Source data and exact p-values are provided as a Source Data file.



Supplementary Figure 14 | Impact of physician satisfaction on delineation performance metrics.
(a, b) Comparison of **(a)** surface Dice Similarity Coefficient (sDSC) and **(b)** Volumetric Revision Index (VRI) in positive attitude (Strongly satisfied/Satisfied) (n = 897 delineation sets) and neutral attitude (n = 95 delineation sets) groups across three contouring methods. Statistical comparisons among methods were performed using two-sided Friedman tests followed by Dunn's multiple comparisons test, while comparisons between attitude groups were analyzed using two-sided Mann–Whitney *U* tests.**(c, d)** Paired comparison of **(c)** sDSC and **(d)** VRI between the positive and neutral groups in manual (n = 95 delineation sets) and AI-assisted (n = 95 delineation sets) delineations. The |Δ| values represent the absolute difference of performance metrics between the positive and neutral groups in manual and AI-assisted delineations. Statistical significance was assessed using two-sided Wilcoxon signed-rank tests. Box plot definitions: center line, median; box limits, interquartile ranges (IQR); whiskers, 1.5×IQR, with precise sample sizes (n, number of OARs) for each box provided in the Source Data file. Statistical significance is denoted by asterisks: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$; ns, not significant ($p > 0.05$). Source data and exact p-values are provided as a Source Data file.

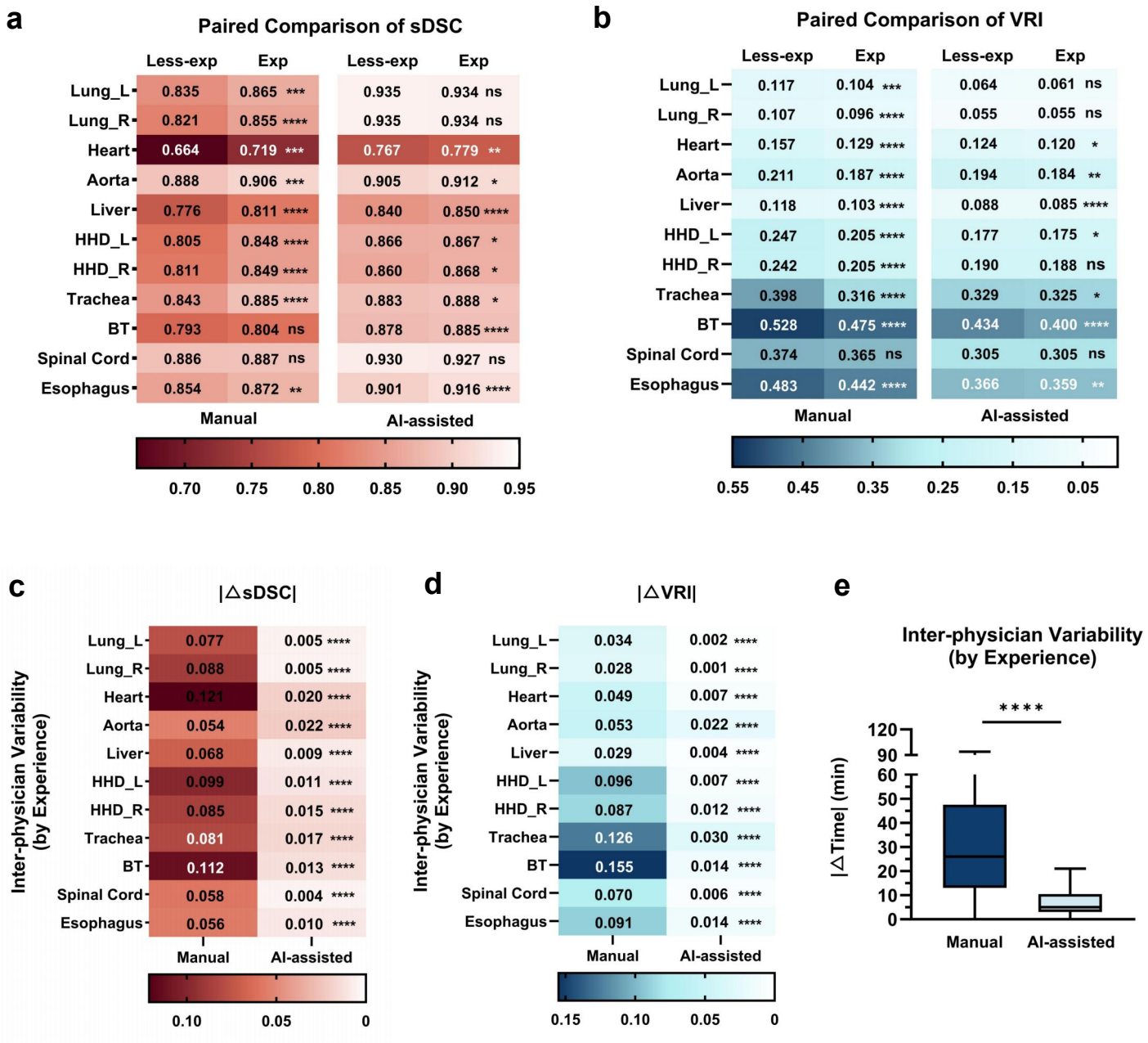


Supplementary Figure 15 | Impact of AI-assisted delineation on inter-center variability.
(a-c) Box plots comparing **(a)** surface Dice similarity coefficient (sDSC), **(b)** volumetric revision index (VRI), and **(c)** contouring time across the five participating centers (n = 261, 317, 160, 138, and 117 delineation sets for Centers 1–5, respectively) for the different methods. Statistical significance was assessed using two-sided Welch's ANOVA followed by Games-Howell multiple comparisons test. For pairwise comparisons, the symbol's position marks the inferior performer, while its color identifies the superior one. Box plot definitions: center line, median; interquartile ranges (IQR); whiskers, 1.5×IQR, with precise sample sizes (n, number of OARs or delineation sets) for each box provided in the Source Data file. Heatmaps showing the median inter-center variability, quantified as the absolute difference (|Δ|) between paired centers (n = 220, 97, 63, 75, and 41 delineation sets for the 5 pairs, respectively), in **(d)** sDSC and **(e)** VRI for manual versus AI-assisted delineations. For these heatmaps, an asterisk indicates the method with significantly lower variability. Statistical significance is denoted by asterisks: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$; ns, not significant ($p > 0.05$). Source data and exact p-values are provided as a Source Data file.

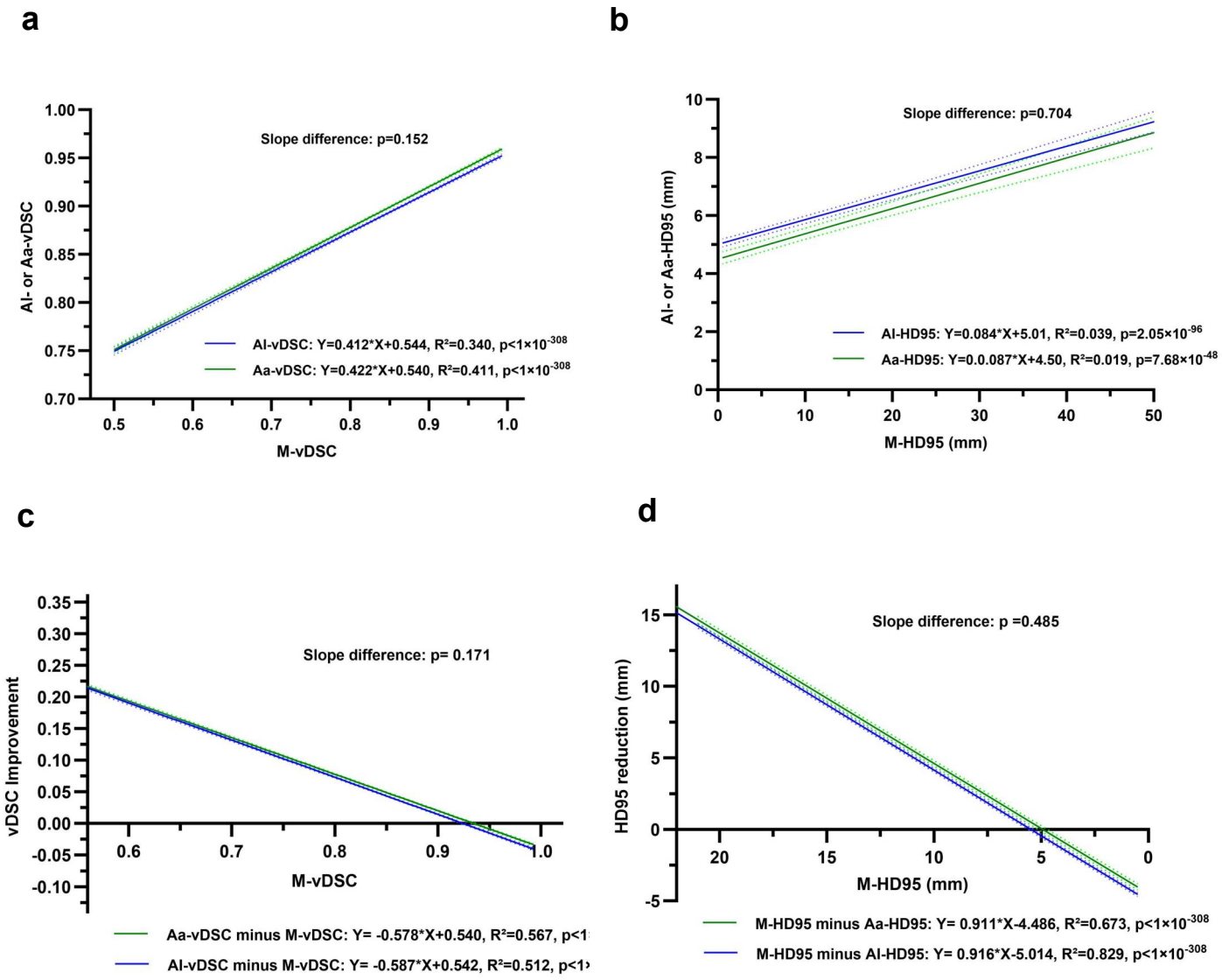


Supplementary Figure 16 | Performance comparison of manual, AI, and AI-assisted delineation methods, stratified by physician experience.

Box plots comparing the performance between less-experienced and experienced physician groups. Performance was evaluated using (a) volumetric Dice similarity coefficient (vDSC), (b) 95th percentile Hausdorff Distance (HD95), (c) surface DSC (sDSC), (d) Volumetric Revision Index (VRI), and (e) contouring time (less-experienced, n = 671 delineation sets; experienced, n = 322 delineation sets). Box plot definitions: center line, median; box limits, interquartile ranges (IQR); whiskers, 1.5×IQR, with precise sample sizes (n, number of OARs) for each box provided in the Source Data file. Statistical comparisons among methods were performed using two-sided Friedman tests followed by Dunn's multiple comparisons test, while comparisons between groups were analyzed using two-sided Mann–Whitney *U* tests. Statistical significance is denoted by asterisks: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$; ns, not significant ($p > 0.05$). Source data and exact p-values are provided as a Source Data file.

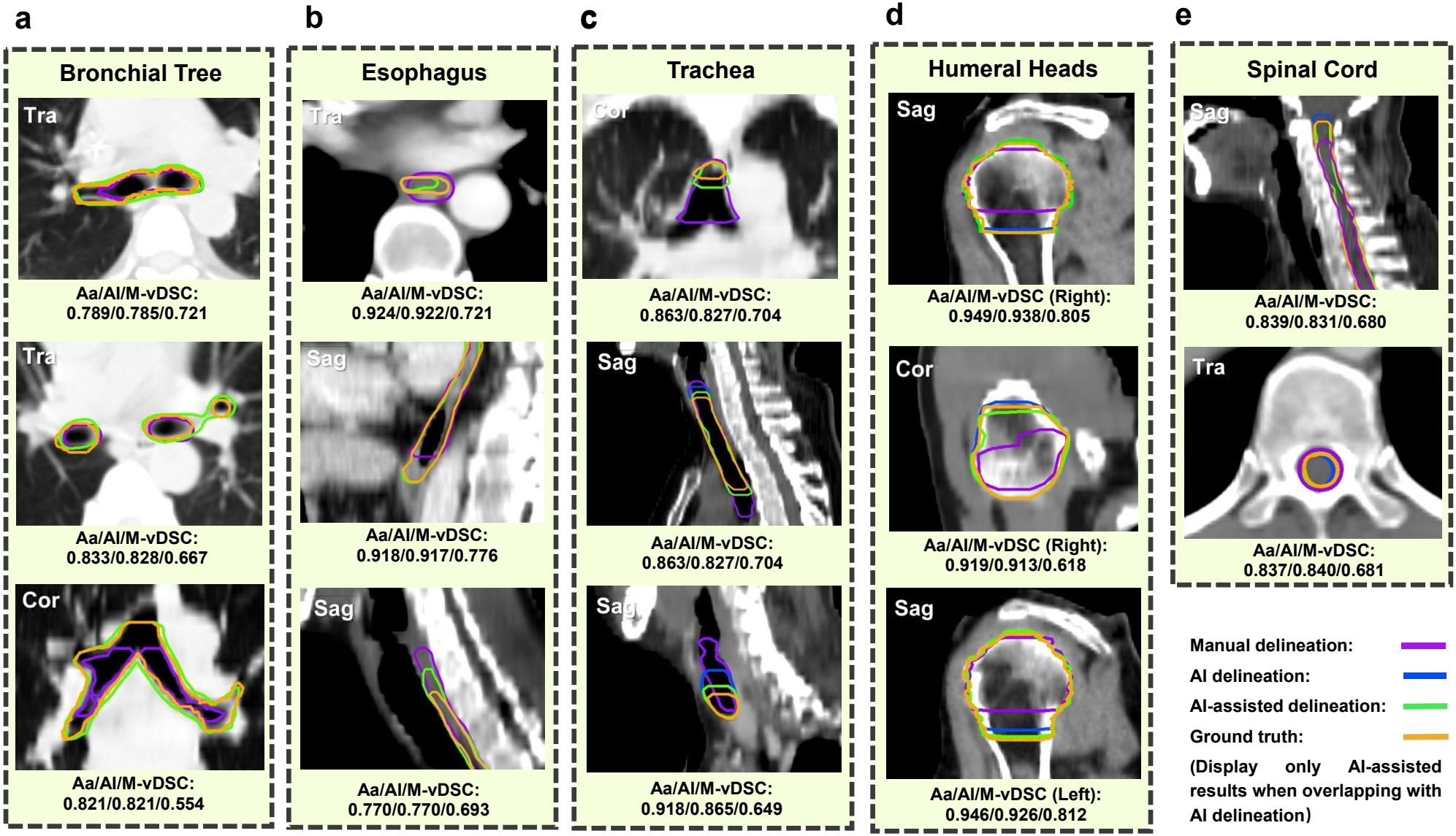


Supplementary Figure 17 | Impact of AI assistance on inter-physician performance and variability.
(a, b) Heatmaps comparing the delineation performance of **(a)** surface Dice similarity coefficient (sDSC) and **(b)** Volumetric Revision Index (VRI) between less-experienced (Less-exp, n = 276 delineation sets) and experienced (Exp, n = 276 delineation sets) physician groups, sho wn for both manual and AI-assisted methods. **(c-e)** Comparison of inter-physician variability, quantified as the absolute difference between the two experience groups (|Δ|). Heatmaps show the median variability in **(c)** sDSC and **(d)** VRI, and the box plots **(e)** shows the variability in contouring time (n = 276 delineation sets for each box). For these heatmaps, an asterisk indicates the method with significantly lower variability. Box plot definitions: center line, median; box limits, interquartile ranges (IQR); whiskers, 1.5×IQR. Abbreviations: HHD, humeral head; BT, bronchial tree. Statistical comparisons were performed using two-sided Wilcoxon signed-rank tests, with significance denoted by asterisks: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$; ns, not significant ($p > 0.05$). Source data and exact p-values are provided as a Source Data file.



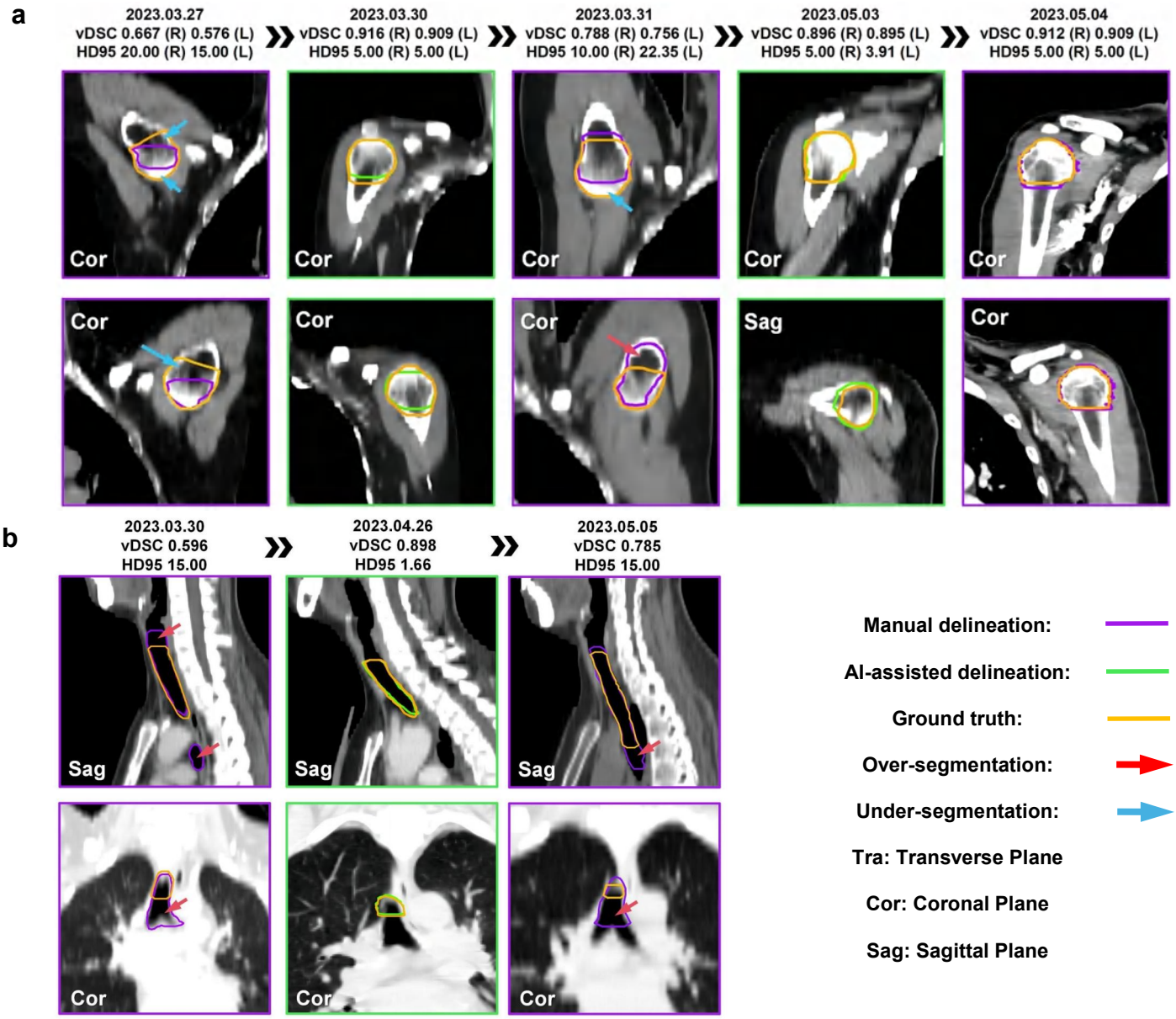
Supplementary Figure 18 | Relationship between manual delineation accuracy and the performance of AI and AI-assisted methods.

(a, b) Linear regression analysis of AI and AI-assisted (Aa) accuracy versus manual (M) accuracy for **(a)** volumetric Dice Similarity Coefficient (vDSC) and **(b)** 95th-percentile Hausdorff Distance (HD95). **(c, d)** Linear regression analysis of the performance improvement, measured as **(c)** vDSC improvement and **(d)** HD95 reduction, versus the initial manual accuracy. In all panels, solid lines represent the linear regression fit, and dotted lines indicate the 95% confidence interval (CI).

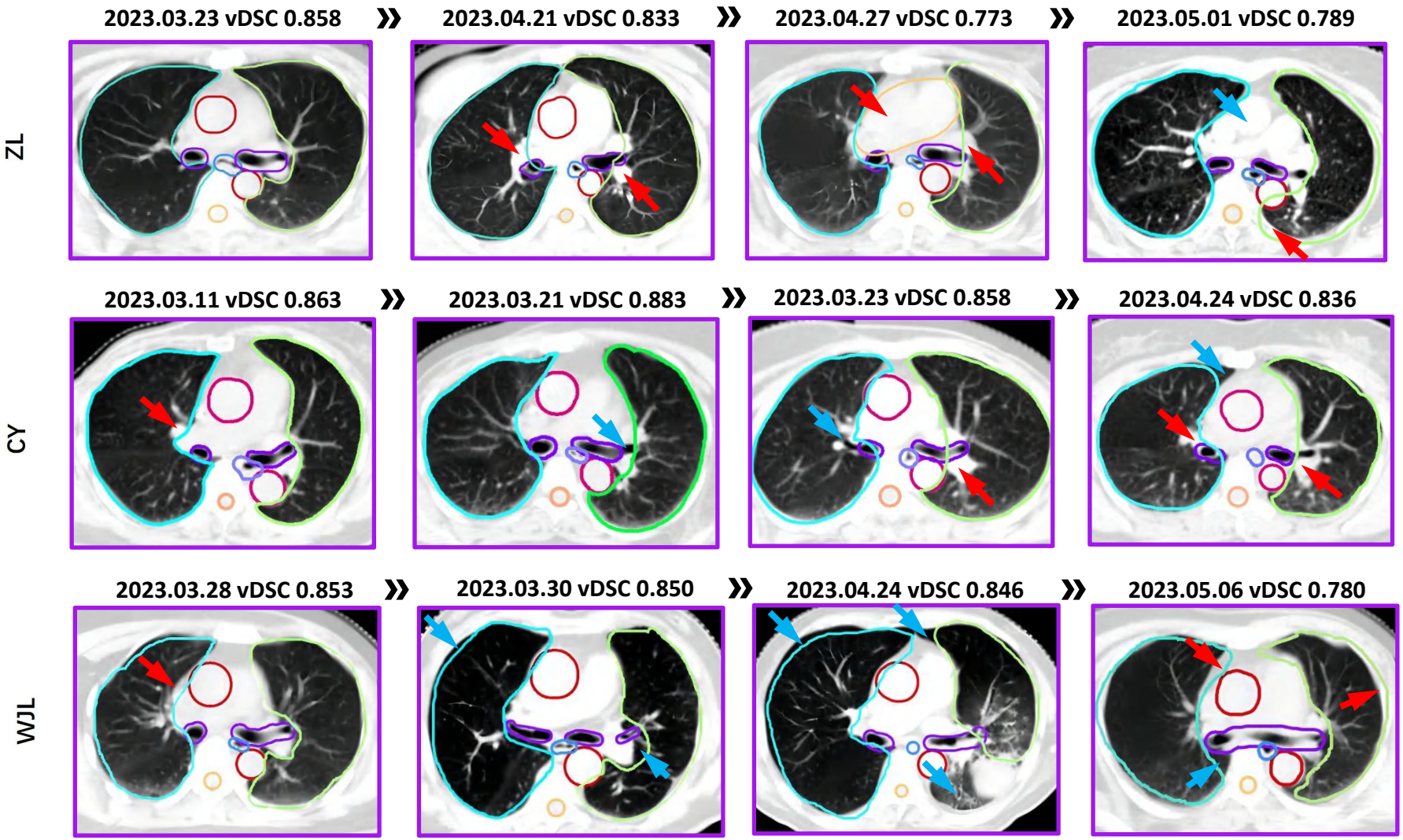


Supplementary Figure 19 | Performance of AI-assisted and AI delineation in enhancing manual delineation accuracy.

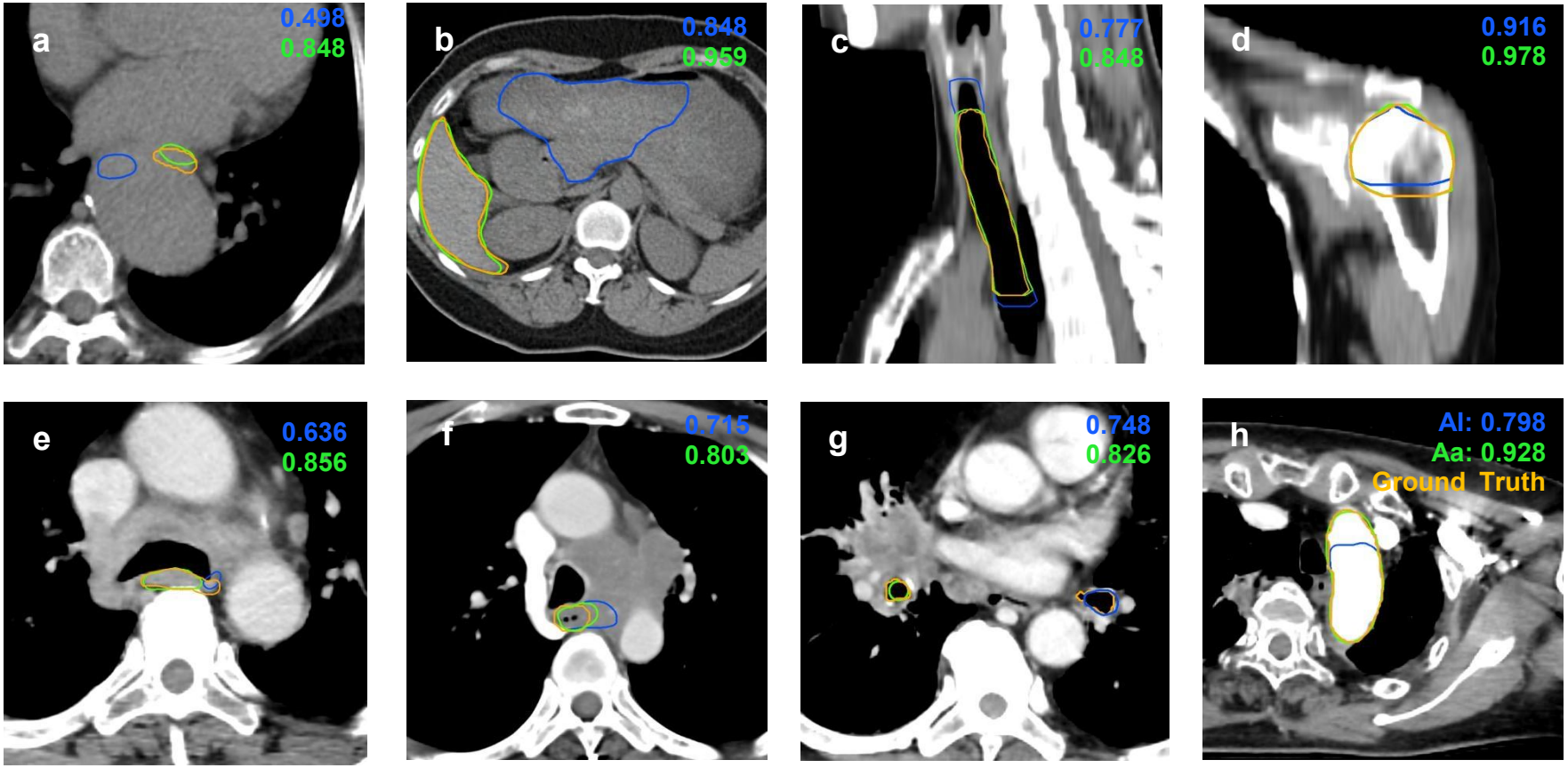
(a–e) For OARs with low M-vDSC (e.g., < 0.8), AI outperforms in boundary definition and recognizing adjacent anatomical structures, with specific improvements in recognizing **(a)** lobar bronchi and the upper boundary of the bronchial tree, **(b)** the upper and lower esophageal boundaries and the azygos vein, **(c)** the upper and lower tracheal boundaries, **(d)** the surgical neck of the humerus and covering the proximal cortical bone, and **(e)** the upper boundary of the spinal cord and the dural sac. Representative images are shown in the transverse (Tra), sagittal (Sag), and coronal (Cor) planes. vDSC, volumetric Dice Similarity Coefficient.



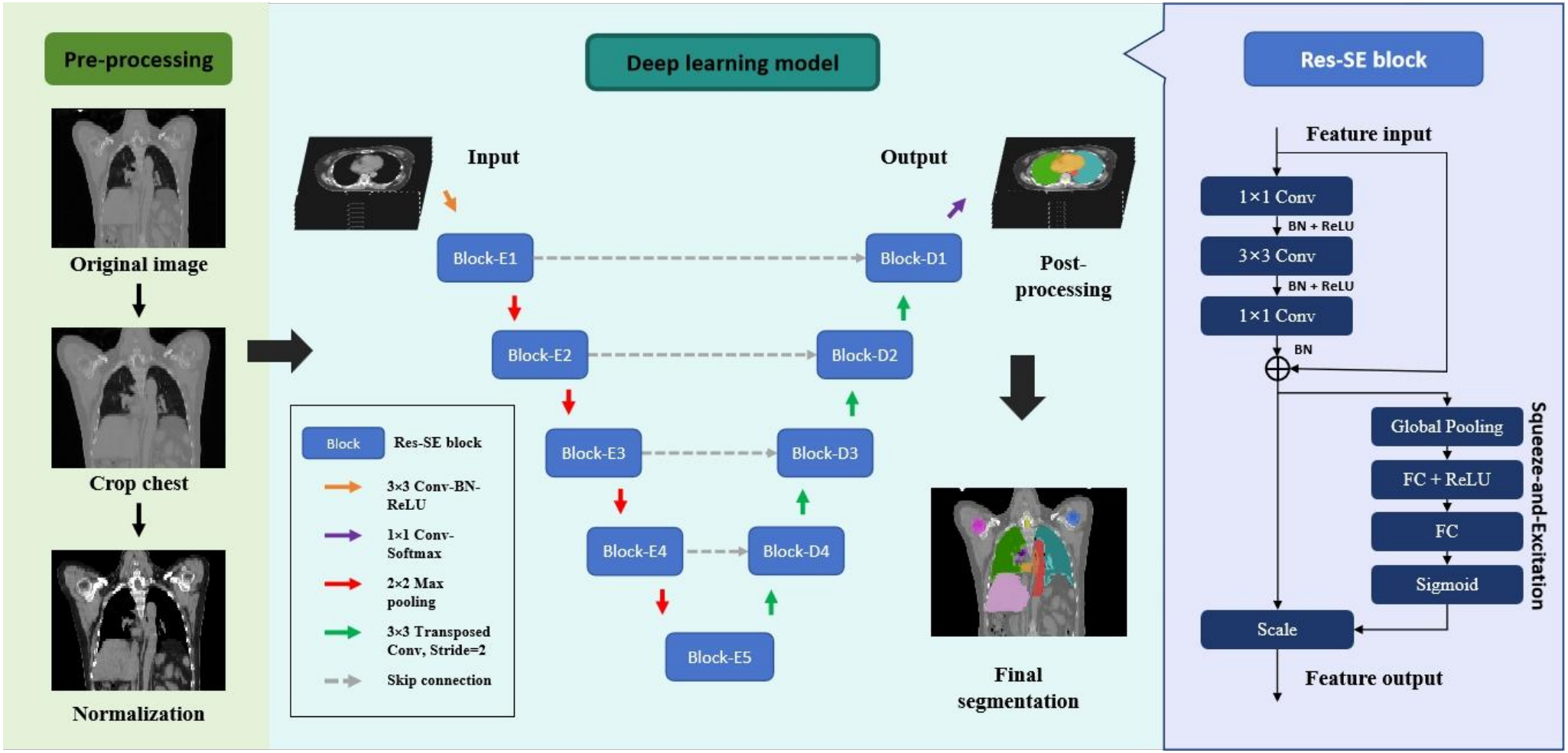
Supplementary Figure 20 | Performance of manual delineation improvement for the humeral head and trachea.
Each panel displays the manual (purple) and AI-assisted (green) delineations performed by a less-experienced physician, alongside the corresponding ground truth (gold), arranged in chronological order. **(a)** Improvement in humeral head delineation, characterized by enhanced recognition of the distal boundary (surgical neck) and improved precision in delineating the proximal cortical bone. **(b)** Improvement in trachea delineation, highlighting enhanced recognition of the upper and lower boundaries (inferior border of the cricoid cartilage and 2 cm above the carina, respectively). Abbreviations: Aa, AI-assisted; vDSC, volumetric Dice Similarity Coefficient; HD95, 95th percentile Hausdorff Distance.



Supplementary Figure 21 | Changes in manual delineation performance over time by three less-experienced physicians with declining vDSC.
The volumetric Dice similarity coefficients (vDSC) represent the mean value in manual delineation across 11 OARs. Red and blue arrows indicate areas of over- and under-segmentation, respectively.



Supplementary Figure 22 | Clinical scenarios where physicians adjusted auto-segmentation results.
Physicians improved auto-segmentation results by **(a-b)** mitigating the impact of low-quality images (e.g., non-contrast CT with thick slices) using anatomical knowledge and clinical experience; **(c-d)** conducting corrective modification based on the boundary definitions; **(e-g)** recognizing OAR structural anomalies caused by tumors or non-tumor factors; and **(h)** identifying model failure scenarios (e.g., over-enhanced OAR regions with adjacent significant artifacts). However, we are uncertain whether the under-segmentation **(g)** is due to model failure or because the training data do not classify tumor-invaded bronchi as OARs. The blue, green, and gold contours represent the AI, AI-assisted delineations, and ground truth, respectively. The blue and green numbers indicate the volumetric Dice similarity coefficients (vDSC) values for the AI and AI-assisted delineation, respectively.



Supplementary Figure 23 | CT image preprocessing workflow and schematic diagram of the Res-SE Net (iCurveE) architecture

Abbreviation: Conv, convolutional; Res, residual; SE, squeeze-and-excitation; BN, batch normalization; ReLU, rectified linear unit; FC, fully connected.

References

1. Wasserthal, J.*, et al.* TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiol Artif Intell* **5**, e230024 (2023).
2. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J. & Maier-Hein, K.H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* **18**, 203-211 (2021).

RESEARCH PROTOCOL

Prospective, Multicenter, Randomized Evaluation of the Performance and Clinical Applicability of AI-Assisted Radiotherapy Contouring Software for Thoracic Organs at Risk

Protocol version	1.0
Date	June 21, 2022
Clinicaltrials.gov-No.	NCT05787522
Principal investigator	Professor Zhiyong Yuan, Tianjin Medical University Cancer Institute & Hospital, Tianjin, China
Sponsor	Tianjin Medical University Cancer Institute & Hospital, Tianjin, China
Coordinating center	Wuhan Union Hospital, Wuhan, China
	Shanxi Province Cancer Hospital, Taiyuan, China
	The Cancer Center of The Fifth Affiliated Hospital, Zhuhai, China
	Nanxishan Hospital of Guangxi Zhuang Autonomous Region, Guilin, China

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AE	Adverse Event
AI	Artificial Intelligence
CI	Confidence Interval
CRA	Clinical Research Associate
CRC	Clinical Research Coordinator
CRF	Case Report Form
CT	Computed Tomography
DL	Deep Learning
DSC	Dice Similarity Coefficient
DV	Difference of Volume
eCRF	Electronic Case Report Form
EDC	Electronic Data Capture
ITD	Intention To Diagnostic
Max	Maximum
Min	Minimum
OAR	Organs At Risk
Q1	Lower Quartile
Q3	Upper Quartile
SD	Standard Deviation

1.SUMMARY

Title: Prospective, multicenter, randomized evaluation of the performance and clinical applicability of AI-assisted radiotherapy contouring software for thoracic organs at risk

Clinicaltrials.gov-No: NCT05787522

Primary objective: To evaluate the accuracy and efficiency of iCurveE software for artificial AI-assisted delineation of thoracic OARs compared to manual delineation.

Secondary objective:

1) Evaluate the impact of multicenter deployment of the software on inter-institutional contouring variability;
2) Evaluate the impact of multi-physician utilization of the software on contouring variability among physicians of varying experience levels;
3) Investigate the impact of physicians' subjective acceptance of the software on its clinical implementation;
4) Explore the respective strengths of AI and physicians in OAR delineation;
5) Explore the impact of AI use on physicians' manual delineation ability in clinical practice.

Trial design: Prospective, multicenter, randomized crossover trial.

Participants: We will enroll 500 patients aged 18 years or older with lung, esophageal, or breast cancer who are scheduled to receive thoracic radiotherapy.

Study procedure: This study adopted a crossover design, with anonymized computed tomography (CT) images randomized in a 1:1 ratio. Cohort 1 underwent initial manual delineation, followed by AI-assisted delineation after a washout period (>4 weeks), while Cohort 2 followed the opposite sequence. The AI delineation results were generated by the auto-segmentation software during the AI-assisted delineation process. Manual, AI, and AI-assisted delineation results will be compared with the ground truth established based on expert consensus and standardized atlases.

2. INTRODUCTION AND RATIONALE

2.1. INTRODUCTION AND BACKGROUND

The incidence of thoracic tumors has been gradually increasing in recent years. For patients with lung, breast, or esophageal cancer, the optimal utilization rate of radiotherapy is approximately 80%.[1] Accurate delineation of OARs is crucial in precision radiation therapy, as it helps protect normal tissues from both acute and late radiation-induced toxicity. However, the traditional method of manual contouring by radiation oncologists is time-consuming and labor-intensive.[2]

Deep learning (DL) algorithms have shown enhanced efficiency and consistency in the automatic segmentation of target volumes and organs at risk OARs.[3] However, the majority of existing research has primarily concentrated on software performance, often neglecting multicenter validation using diverse imaging data from cancer patients. [4,5] More importantly, due to the "black-box" nature of DL algorithms and the ethical implications surrounding AI, these auto-segmentation software cannot independently finalize delineations for treatment plans. This situation highlights the need for extensive end-user testing,[6] which requires physicians to review and revise the results.

While recent studies have conducted end-user testing based on AI-assisted delineation, they often involve small cohorts of images reviewed by a limited number of expert physicians.[7] However, considering the increasing demand for radiotherapy and significant regional disparities in resource availability,[8] less-experienced physicians may often review and modify AI-generated OARs in clinical practice. This variability in expertise may potentially affect both the revision workload and accuracy of final delineations. Furthermore, the inexplicability of DL algorithms complicates the understanding of how physicians and AI tools collaborate, divide tasks, and influence each other in clinical settings. Acquiring prospective, multi-institutional manual delineation data for comparison is resource-intensive. Therefore, despite the rapid development of model algorithms, the widespread clinical implementation of auto-segmentation tools may require more high-quality evidence.

2.2. RATIONALE FOR PERFORMING THE STUDY

This trial aims to evaluate the performance and clinical applicability of a deep learning-based auto-segmentation software for delineating thoracic OARs based on a large-scale multi-institutional validation and end-user testing. Additionally, by leveraging multicenter deployment and multi-user utilization of this software, we aim to investigate its role in promoting healthcare equity and the potential mechanisms underlying physician-AI interactions that enhance segmentation outcomes.

3. STUDY OBJECTIVES

3.1. PRIMARY OBJECTIVES

Evaluate the accuracy and efficiency of iCurveE software for AI-assisted delineation of thoracic OARs compared to the manual delineation.

3.2. SECONDARY OBJECTIVES

1) Evaluate the impact of multicenter deployment of the software on inter-institutional contouring variability;
2) Evaluate the impact of multi-physician utilization of the software on contouring variability among physicians of varying experience levels;
3) Investigate the impact of physicians' subjective acceptance of the software on its clinical implementation;
4) Explore the respective strengths of AI and physicians in OAR delineation;
5) Explore the impact of AI use on physicians' manual delineation ability in clinical practice.

4. STUDY DESIGN

4.1. OVERALL DESIGN

This is a prospective, multicenter, randomized crossover trial.

4.2. GROUPS

1) Independent manual delineation;
2) AI delineation;
3) AI-assisted delineation.

4.3. STUDY ORGANIZATION

1) Five academic clinical centers in China for recruitment Tianjin Medical University Cancer Institute & Hospital (TMUCIH), Wuhan Union Hospital (WHUH), Shanxi Cancer Hospital (SXCH), The Fifth Affiliated Hospital of Sun Yat-sen University (SUH), and Nanxishan Hospital of Guangxi Zhuang Autonomous Region (NXSH);
2) Total recruitment of 500 patients eligible for thoracic radiotherapy;
3) Thirty-seven radiation oncologists generate manual, AI, and AI-assisted delineation results;
4) Sixteen senior radiation oncologists with >10 years of contouring experience establish the ground truth, independent of 3).

5. STUDY POPULATION

This trial will enroll 500 patients aged 18 or older with lung, esophageal, or breast cancer, who are scheduled to receive thoracic radiotherapy across five clinical cancer institutes.

5.1. INCLUSION CRITERIA

1) Subjects aged 18 years or older, with no gender restrictions;
2) Patients diagnosed with breast cancer, lung cancer, or esophageal cancer, who are scheduled for chest CT scanning followed by thoracic radiotherapy;
3) CT slice thickness of 5 mm or less;
4) Able to understand the study objectives, willing to participate voluntarily, and able to provide informed consent.

5.2. EXCLUSION CRITERIA

1) Congenital malformations or abnormal anatomical structures resulting from non-tumor

factors in the scan area;

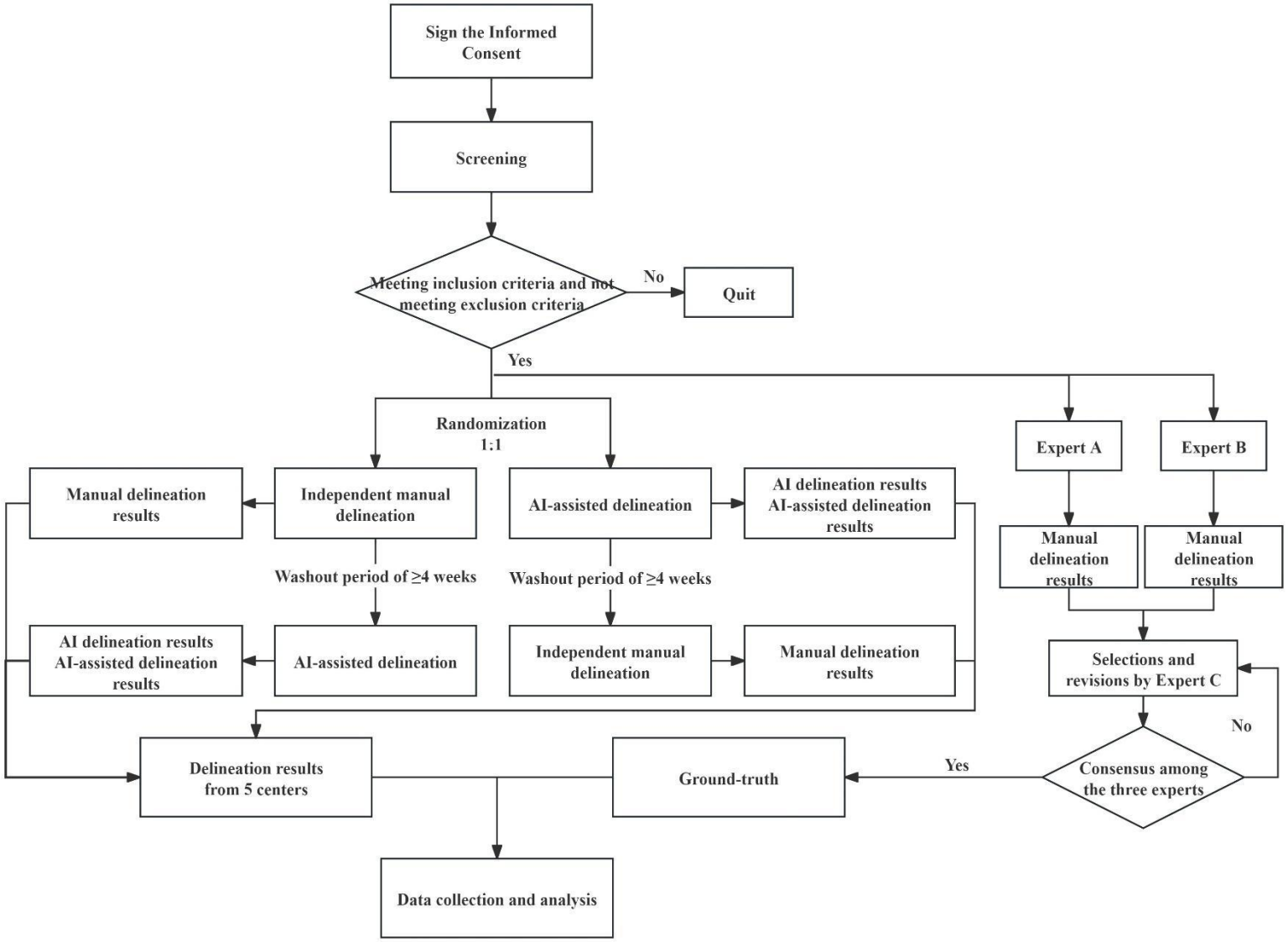
2) Poorly defined anatomical structures due to the presence of artifacts, prosthetics, or implants;
3) Imaging not compliant with DICOM standards;
4) Deemed unsuitable by the investigators.

5.3. WITHDRAWAL CRITERIA

1) Subject voluntarily withdraws or requires alternative treatment;
2) Investigator determines discontinuation is necessary due to a significant safety concern.

6. STUDY PROCEDURE

6.1. STUDY FLOW CHART



6.2. CONSENT AND SCREENING PROCEDURES

6.2.1. SUBJECT SOURSES

Subjects are recruited from outpatient and inpatient populations scheduled for thoracic radiotherapy, with screening conducted at five clinical centers.

6.2.2. INFORMED CONSENT

According to the Declaration of Helsinki, patients must provide informed consent before participating in the study. The clinical oncologist will explain the purpose, potential benefits, and risks of the clinical trial to the patients. After fully understanding the information, patients will agree to participate and voluntarily sign the informed consent form. The CRC will assist the investigator in conducting preliminary screening according to inclusion and exclusion criteria, collecting copies of patients' identification documents, and recording the subject screening and enrollment form. Eligible patients will be enrolled in the trial. The subject screening and enrollment form includes information such as screening number, name initials, Patient ID, consent date, chest CT examination date, imaging number, screening success status, enrollment number, and reason for screening failure. The investigator should only record the first identified reason for exclusion in the cases of screening failure.

6.2.3. PHYSICIAN REQUIREMENTS

The principal investigator at each participating center shall designate several radiation or oncology physicians to serve as investigators responsible for subject screening, case selection, and performing contouring (manual, AI, and AI-assisted). These physicians shall not be involved in establishing the ground truth. For each subject and corresponding imaging, the roles of screening physician and contouring physician must remain independent. All designated physicians must have a minimum of one year of experience in radiotherapy.

6.3. CT IMAGING

6.3.1. CT IMAGE ACQUISITION

1) The scanning range must include at least from the cricoid cartilage level to the lower edge of the liver, with both non-contrast and contrast-enhanced scans acceptable;
2) Patient positioning should be consistent with the requirements of the subsequent radiotherapy plan, with no specific restrictions on supine/prone positioning or arm placement;
3) If there is minor edge loss in the scanning range, the principal investigator at each site may determine case eligibility based on the actual situation.

6.3.2. IMAGING DATA ANONYMIZATION

This trial will collect only CT imaging data, and contouring results will not be used for subsequent radiotherapy planning. At each participating center, CT images (in DICOM format) will be anonymized using a third-party anonymization tool (Dicom Tag Erase) to remove personally identifiable information, including patient ID, patient name, institution name, institution address, attending physician's name, operator's name, recorder's name, and patient address. Demographic and clinical information such as gender, age, cancer type, and scan parameters will be retained, and each dataset will be assigned a specific anonymized image code.

6.3.3. INTER-CENTER IMAGING TRANSFER

Each participating center will utilize a rotational contouring methodology. In addition to contouring images from their own center, investigators (excluding those involved in generating the ground truth) will contour images from the preceding center in the sequence. Specifically:

- Investigators from Center 01 will contour images from Center 01 and Center 05;
- Investigators from Center 02 will contour images from Center 02 and Center 01;
- Investigators from Center 03 will contour images from Center 03 and Center 02;
- Investigators from Center 04 will contour images from Center 04 and Center 03;
- Investigators from Center 05 will contour images from Center 05 and Center 04.

6.4. RANDOMIZATION

To minimize bias from the delineation sequences, this study adopts a crossover design, with anonymized CT images randomized in a 1:1 ratio. Group 1 underwent initial manual delineation, followed by AI-assisted delineation after a washout period (>4 weeks), while Group 2 followed the opposite sequence. The delineation sequence was determined using SAS version 9.4 (SAS Institute, Cary, NC), with block randomization (block size of ten) employed to generate a computerized randomization schedule. Sealed randomization cards were prepared by an independent statistician based on this schedule and distributed to each center. Contouring physicians will determine the sequence of manual and AI-assisted delineation according to the corresponding randomization card assigned to each image.

6.5. IMAGING DELINEATION

6.5.1. OARS FOR DELINEATION

The thoracic OARs that need to be delineated include the left and right lungs, heart, spinal cord, left and right humeral heads, aorta, trachea, bronchial tree, esophagus, and liver. The OAR boundary definitions are based on the RTOG 1106 guidelines.[9]

6.5.2. MANUAL DELINEATION

Manual delineation refers to physicians using the brush tool on the contouring platform to segment OAR manually, without the use of auto-segmentation tools. Manual delineation will be performed on the PVmed Contouring version 1.0 platform. This software is registered with the National Medical Products Administration (NMPA) in China (Registration No. 20203210802).

6.5.3. AI AND AI-ASSISTED DELINEATION

AI delineation (auto-segmentation results) for subject imaging is performed using Res-SE Net (integrated into the iCurveE software). The contouring physicians import the AI-generated results into the PVmed Contouring platform and perform subsequent manual modifications, which generates AI-assisted delineation results.

6.5.4. GROUND TRUTH GENERATION

Ground truth for each image is generated by three senior radiation oncologists (A, B, and C), each with over 10 years of radiotherapy experience, following standard anatomical atlases. [9] Each center has its corresponding experts A and B, who are independent from the physicians responsible for generating the manual and AI-assisted delineation results. Expert C is independent of the five centers participating in this study. Experts A and B independently delineate each CT image, producing two sets of OARs, which are subsequently reviewed by expert C. Expert C compares the delineations from experts A and B, selecting the most accurate delineation for each OAR (e.g., the right lung from expert A and the spinal cord from expert B), and makes minor adjustments if necessary. To minimize bias, the OARs finalized by expert C are established as the ground truth only if consensus is reached among all three experts; otherwise, further modifications are conducted.

6.6. EVALUATION METRICS

Metrics		**Definition**
Primary Endpoints	DICE similarity coefficient, DSC	DSC= 2 × (A∩B) / (A+B)
	Contouring time (min)	Manual delineation time is recorded from the time the CT is loaded on the contouring platform to the completion of contouring. AI-assisted delineation time is defined as the sum of the auto-segmentation model runtime, the transfer to the contouring platform, and the subsequent manual modification.
Secondary Endpoints	Difference of volume, DV	DV = | B-A| / A
	Recall, Rec	Rec = | A ∩ B | / A
	Precision, Pre	PR= | A ∩ B | / B
	Time efficiency improvement rate (%)	Time efficiency improvement rate = (manual delineation time - AI-assisted delineation time)/ manual delineation time*100%
	Subjective assessment of AI delineation using a Likert scale	Evaluated on a 1-5 Likert scale: 1 - Strongly dissatisfied, 2 - Dissatisfied, 3 - Neutral, 4 - Satisfied, 5 - Strongly satisfied.

A refers to the volume of ground truth, while B refers to the volume of manual or AI-assisted delineation.

7. STATISTICAL ANALYSIS PLAN

7.1. RESEARCH HYPOTHESES

The primary endpoint is the DSC. To ensure that AI-assisted delineation meets clinical application requirements, a non-inferiority comparison will be employed to evaluate the DSC between AI-assisted and manual delineation for OARs, including the heart, liver, left lung, right lung, spinal cord, left humeral head, right humeral head, aorta, trachea, bronchial tree, and esophagus. The corresponding statistical hypothesis test is as follows:

$$H0: P_T - P_C \leq \Delta$$

$$H1: P_T - P_C > \Delta$$

Where PT represents the DSC in the experimental group (i.e., AI-assisted delineation), PC represents the DSC in the control group (i.e., manual delineation), and Δ is the non-inferiority margin. If non-inferiority is established, superiority testing will be conducted.

7.2. SAMPLE SIZE CALCULATION

The calculations are based on different parameters, including significance level, power, expected dropout rate, non-inferiority margin. Based on prior experimental results and published literature, this study will compare the DSC of AI-assisted delineation with manual delineation for the following OARs: heart, liver, left and right lungs, spinal cord, left and right humeral heads, aorta, trachea, bronchial tree, and esophagus.

For the non-inferiority sample size calculation, α is set at 0.025 (one-sided test), with 1 - β equals 0.9, and an anticipated dropout rate of 10%. The calculated results are summarized as follows.

$$\text{Sample size: } n_T = n_C = 2(Z_{1-\alpha} + Z_{1-\beta})^2 \sigma^2 / (D - \Delta)^2$$

D represents the difference in DSC between the two groups, $D = \mu_T - \mu_C$ (if $\mu_T > \mu_C$, D = 0; if $\mu_T < \mu_C$, D = the difference value), σ represents the SD for each group, and Δ is the non-inferiority margin, set at -0.05.

Based on the results of the sample size calculation, a minimum of 474 participants is required. Therefore, this trial aims to recruit 500 participants.

OARs	α (one -sided)	Power	DSC (mean)		DSC (SD)			**D**	**Δ**	Calculated sample size	Anticipated sample size (considering 10% dropout)
			Manual	AI -assisted	Manual	AI -assisted	**σ**				
Aorta	0.025	0.9	0.872	0.926	0.040	0.029	**0.040**	**0**	**-0.05**	15	17
Bronchial Tree			0.751	0.816	0.085	0.065	**0.085**			62	69
Esophagus			0.732	0.794	0.055	0.028	**0.055**			27	30
Left Humeral Head			0.898	0.945	0.037	0.026	**0.037**			13	15
Right Humeral Head			0.769	0.873	0.225	0.091	**0.225**			426	474
Spinal Cord			0.778	0.880	0.215	0.074	**0.215**			389	433
Heart			0.925	0.967	0.018	0.019	**0.018**			4	5
Liver			0.914	0.971	0.022	0.012	**0.022**			6	7
Left Lung			0.922	0.979	0.020	0.008	**0.020**			5	6
Right Lung			0.798	0.848	0.115	0.024	**0.115**			113	126
Trachea			0.813	0.869	0.089	0.030	**0.089**			68	76

7.3. STATISTICAL ANALYSIS METHODS

Continuous variables will be summarized by the number of subjects, mean, standard deviation, median, interquartile range, minimum, and range (minimum and maximum). Categorical variables will be described using frequencies and percentages. For superiority testing, if the lower limit of the 95% confidence interval exceeds 0, efficacy is confirmed. For non-inferiority testing, if the lower limit of the 95% confidence interval exceeds the non-inferiority margin, non-inferiority is established. Baseline demographic data will be primarily analyzed using descriptive statistics, with potential subgroup comparisons for categorical variables conducted using the likelihood ratio χ^2 test. Paired t-tests will be used for parametric comparisons of paired samples. For non-parametric data, the Wilcoxon signed-rank test and the Mann-Whitney U test will be applied for paired and independent samples, respectively. Differences among multiple groups of quantitative data will be assessed using analysis of variance (ANOVA) and the F-statistic. A two-sided

p-value less than 0.05 is considered statistically significant.

8. ADVERSE EVENTS

8.1. DEFINITIONS

1) **Adverse Events (AEs):** Adverse events are defined as any unfavorable medical occurrences during the clinical trial of a medical device, regardless of their relationship to the investigational device.
2) **Serious Adverse Events (SAEs):** Serious adverse events are defined as occurrences during the clinical trial that result in death or significant deterioration of health, including life-threatening illnesses or injuries, permanent impairment of body structure or function, hospitalization or prolongation of existing hospitalization, or the need for medical intervention to prevent permanent impairment of body structure or function.

8.2. DOCUMENTATION OF AEs AND SAEs

This study does not involve any patient intervention, and the delineation results will not be used for radiotherapy planning. Therefore, only adverse events occurring during the CT imaging process will be recorded in this trial.

9. DATA RECORDING AND MANAGEMENT

9.1. DATA RECORDING

This trial uses an Electronic Data Capture (EDC) system to collect information required for case report forms.

1) **EDC Completion Guidelines:** Before completing the EDC, investigators should thoroughly review the EDC completion manual. All observations and findings during the clinical trial must be recorded in the EDC accurately, comprehensively, and promptly. It is imperative that the entered data are traceable, supported by source documents, and consistent with them.
2) **Timely EDC Entry:** The EDC must be completed promptly and not deferred until all cases have concluded. During the recording process, it is crucial to ensure timely

communication for the investigator to complete the signature process. An Epidata database will be established for data entry. Data entry will be performed by two independent data entry personnel, and logic validation rules will be set up to check the accuracy of the data entered automatically by the software. After data entry is completed, a double-check comparison will be conducted. Any inconsistencies found will be corrected based on the CRF.

3) **Monitoring and Verification:** During the trial, the CRA will regularly visit research centers to verify the data entered in the EDC. CRA should cross-check the data recorded in the EDC with the original medical records to ensure consistency. They are required to verify the EDC page by page to ensure authenticity, consistency, and completeness.
4) **Source Data Documentation:** For information on which original documents correspond to the data entered in the EDC, refer to the Source Data Attribution (SDA) protocol.

9.2. DATA MANAGEMENT

Once the CRF is completed and signed off, it will be submitted to the clinical trial management bodies at each participating center for quality control. In case of any queries, a query form will be used to address them, and the investigator will provide a response upon confirmation. All queries and responses, as well as data updates and modifications, will be recorded and properly maintained. The data administrator, investigator, and statistical analyst will complete the final definition and determination of the analysis population. The data can be locked once the following conditions are met: 1) All data have been entered into the database; 2) All queries have been resolved; and 3) The analysis population has been defined and finalized.

10. ETHICAL CONSIDERATIONS

10.1. ETHICS COMMITTEE APPROVAL

Ethics Committee approval is required for this study at the research center. Before the start of the clinical trial, the protocol must be approved by the Ethics Committee of the

research center. During the trial, any modifications to the protocol that may impact the rights, safety, or health of participants, or that may affect the trial's objectives or endpoints, must also be approved by the Ethics Committee. All substantial amendments must be submitted to the Ethics Committee for review or filing.

10.2. INFORMED CONSENT REQUIREMENTS

The investigator must follow the ethical principles of the Declaration of Helsinki and meet the following requirements:

1) Use the latest version of the informed consent form approved by the Ethics Committee.
2) Before participation, the investigator must explain the trial, including the medical device, procedures, potential benefits, risks, and discomforts, after which the participant signs and dates the consent form. The investigator also signs and dates it.
3) If the participant lacks capacity, written consent must be obtained from their guardian. If the participant cannot read, an impartial witness must be present, and both the witness and the participant must sign and date the form.
4) Participants must not be coerced or unduly influenced to join the trial.
5) If the informed consent form is updated and approved by the Ethics Committee, all affected participants must be informed and re-sign the updated form.

10.3. PATIENT CONFIDENTIALITY

The investigator must ensure that the privacy of clinical trial participants is protected. In all documents submitted to the sponsor, only subject numbers and initials should be used to identify participants, and full names must not be disclosed. The investigator must maintain a subject identification code list, which is kept strictly confidential by the investigator.

11. REFERENCES

1. Atun R, Jaffray DA, Barton MB, et al. Expanding global access to radiotherapy. The Lancet Oncology 2015; 16(10): 1153-86.
2. Vinod SK, Jameson MG, Min M, Holloway LC. Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. Radiotherapy and oncology: journal of the European Society for Therapeutic Radiology and Oncology 2016; 121(2): 169-79.
3. Feng X, Qing K, Tustison NJ, Meyer CH, Chen Q. Deep convolutional neural network for segmentation of thoracic organs-at-risk using cropped 3D images. Medical physics 2019; 46(5): 2169-80.
4. Shi F, Hu W, Wu J, et al. Deep learning empowered volume delineation of whole-body organs-at-risk for accelerated radiotherapy. Nature communications 2022; 13(1): 6566.
5. Chen X, Sun S, Bai N, et al. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. Radiotherapy and oncology: journal of the European Society for Therapeutic Radiology and Oncology 2021; 160: 175-84.
6. Huynh E, Hosny A, Guthier C, et al. Artificial intelligence in radiation oncology. Nature reviews Clinical oncology 2020; 17(12): 771-81.
7. Lu SL, Xiao FR, Cheng JC, et al. Randomized multi-reader evaluation of automated detection and segmentation of brain tumors in stereotactic radiosurgery with deep neural networks. Neuro-oncology 2021; 23(9): 1560-8.
8. Zubizarreta E, Van Dyk J, Lievens Y. Analysis of Global Radiotherapy Needs and Costs by Geographic Region and Income Level. Clinical oncology (Royal College of Radiologists (Great Britain)) 2017; 29(2): 84-92.
9. Kong FM, Ritter T, Quint DJ, et al. Consideration of dose limits for organs at risk of thoracic radiotherapy: atlas for lung, proximal bronchial tree, esophagus, spinal cord, ribs, and brachial plexus. International journal of radiation oncology, biology, physics 2011; 81(5): 1442-57.

STATISTICAL ANALYSIS PLAN (SAP) ADDENDUM

(Amendment for Inclusion of Geometric and Efficiency Metrics)

Clinicaltrials.gov No.: NCT05787522

Date: September 24, 2025

1. Rationale for Amendment

Subsequent to the trial initiation, the NRG Oncology working group published consensus recommendations emphasizing that volume-based metrics alone are insufficient for standardized validation.[1] Specifically, this consensus mandates the inclusion of distance-based metrics (e.g., HD95) as a minimum requirement for AI auto-segmentation assessment. To align with these evolving community standards and address peer review feedback, this addendum formally incorporates HD95, sDSC, and efficiency metrics (e.g., VRI). These metrics have been synchronized with the updated ClinicalTrials.gov record and the final manuscript to ensure comprehensive reporting of geometric accuracy and clinical applicability.

2. Definitions of Metrics

The detailed definitions and calculation formulas for the additional metrics are listed below:

Metric		**Definition**
Primary Endpoints	Volumetric Dice similarity coefficient, vDSC	vDSC= $2 \times V_{A\cap B} / (V_A+V_B)$
	Contouring time (min)	Manual delineation time is recorded from the moment the CT image is displayed on the contouring platform. The AI-assisted delineation time includes the runtime of the auto-segmentation model, the transfer of auto-segmentation results to the contouring platform, and the subsequent manual modification period.
Secondary Endpoints	95th percentile Hausdorff Distance, HD95	HD95(A, B) = max (h95(A, B), h95(B, A)), where h95(A, B) is the 95th percentile of the shortest distances from all points on surface A to surface B, and vice-versa for h95(B, A).
	Surface Dice similarity coefficient, sDSC	sDSC = ($\|S(A) \cap S(B)\tau\| + \|S(B) \cap S(A)\tau\|$) / ($\|S(A)\| + \|S(B)\|$), where S(A) and S(B) are the sets of points on the surfaces of A and B, S(B)τ represents the points on surface B that are within the tolerance τ of surface A, and S(A)τ represents the points on surface A that are within the tolerance τ of surface B.
	Volumetric revision index, VRI*	$[(V_A- V_{A\cap B}) + (V_B-V_{A\cap B})] /V_A$
	Time efficiency improvement rate (%)	Time efficiency improvement rate = (manual delineation time - AI-assisted delineation time)/ manual delineation time*100%

	Recall, Rec	Rec = $V_{A\cap B}$ / V_A
	Precision, Pre	Pre=$V_{A\cap B}$ / V_B
	Relative volume difference, RVD	RVD= $\left\lvert V_A - V_B \right\rvert$ / V_A
	Subjective assessment of AI delineation using a Likert scale	Evaluated on a 1-5 Likert scale: 1 - Strongly dissatisfied, 2 - Dissatisfied, 3 - Neutral, 4 - Satisfied, 5 - Strongly satisfied.

In the formulas, A represents the ground truth and B represents the manual, AI or AI-assisted delineation, while V_A and V_B represent their respective volumes. A tolerance of τ=2 mm was used in this study. *When the delineated volume by the physician closely matches the ground truth in size but differs significantly in shape, RVD may not accurately reflect the expert's modification effort. To address this, we introduced the VRI parameter, which considers both the added and removed volumes during modifications.

3. Classification of Outcomes

To ensure consistency with the updated ClinicalTrials.gov registration:

3.1 Primary Outcomes:

- vDSC and contouring time remain the primary endpoints, consistent with the original protocol.
- Clarification: The term "vDSC" adopted in this addendum represents the exact same metric as "DSC" defined in the original protocol. The definition and calculation method remain completely identical.

3.2 Secondary Outcomes:

- 95th Percentile Hausdorff Distance (HD95), Surface Dice Similarity Coefficient (sDSC), and Volumetric Revision Index (VRI) are newly calculated and supplemented as key secondary endpoints to support the multidimensional assessment of the software's performance.
- Retained Metrics: The original secondary metrics, including Recall, Precision, and Relative Volume Difference (RVD) (previously termed "Difference of Volume (DV)" in the original protocol), are fully retained.

4. Statistical Methods

Statistical analyses will follow the methodology in Section 7.3 of the original protocol, with the following specific additions for the new metrics and multi-group comparisons:

Two-Group Comparisons: Differences between two paired groups (e.g., AI-assisted vs. Manual) will be assessed using Paired t-tests (parametric) or Wilcoxon signed-rank tests (non-parametric).

Multi-Group Comparisons: For comparisons across three paired groups (Manual vs. AI vs. AI-assisted), the Friedman test will be employed. If the Friedman test indicates statistical significance, Dunn's post-hoc test with Bonferroni correction will be performed for pairwise comparisons.

5. Data Source and Ethical Compliance

This supplementary analysis is strictly a post-hoc assessment of data already acquired during the trial; it involves no new patient enrollment, additional clinical interventions, or modifications to the patient care workflow. Therefore, it does not alter the risk-benefit profile of the study or require additional ethical approval for patient participation.

6. Reference

1. Rong, Y.*, et al.* NRG Oncology Assessment of Artificial Intelligence Deep Learning-Based Auto-segmentation for Radiation Therapy: Current Developments, Clinical Considerations, and Future Directions. *International journal of radiation oncology, biology, physics* **119**, 261-280 (2024).