

Touchstone Benchmark

Are We On The Right Way For Evaluating Medical Segmentation?

Pedro R. A. S. Bassi · Wenxuan Li · Yucheng Tang · Fabian Isensee · Zifu Wang · Jieneng Chen · Yu-Cheng Chou · Saikat Roy · Yannick Kirchoff · Maximilian Rokuss · Ziyang Huang · Jin Ye · Junjun He · Tassilo Wald · Constantin Ulrich · Michael Baumgartner · Klaus H. Maier-Hein · Paul Jaeger · Yiwen Ye · Yutong Xie · Jianpeng Zhang · Ziyang Chen · Yong Xia · Zhaohu Xing · Lei Zhu · Yousef Sadegheih · Afshin Bozorgpour · Pratibha Kumari · Reza Azad · Dorit Merhof · Pengcheng Shi · Ting Ma · Yuxin Du · Fan Bai · Tiejun Huang · Bo Zhao · Haonan Wang · Xiaomeng Li · Hanxue Gu · Haoyu Dong · Jichen Yang · Maciej A. Mazurowski · Saumya Gupta · Linshan Wu · Jiaxin Zhuang · Hao Chen · Holger Roth · Daguang Xu · Matthew B. Blaschko · Sergio Decherchi · Andrea Cavalli · Alan L. Yuille · Zongwei Zhou

These authors contributed equally to this work: Pedro R. A. S. Bassi (psalvad2@jh.edu) and Wenxuan Li (wli131@jh.edu)

Correspondence to: Alan L. Yuille (ayuille1@jhu.edu) and Zongwei Zhou (zzhou82@jh.edu)

This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research and the Patrick J. McGovern Foundation Award.



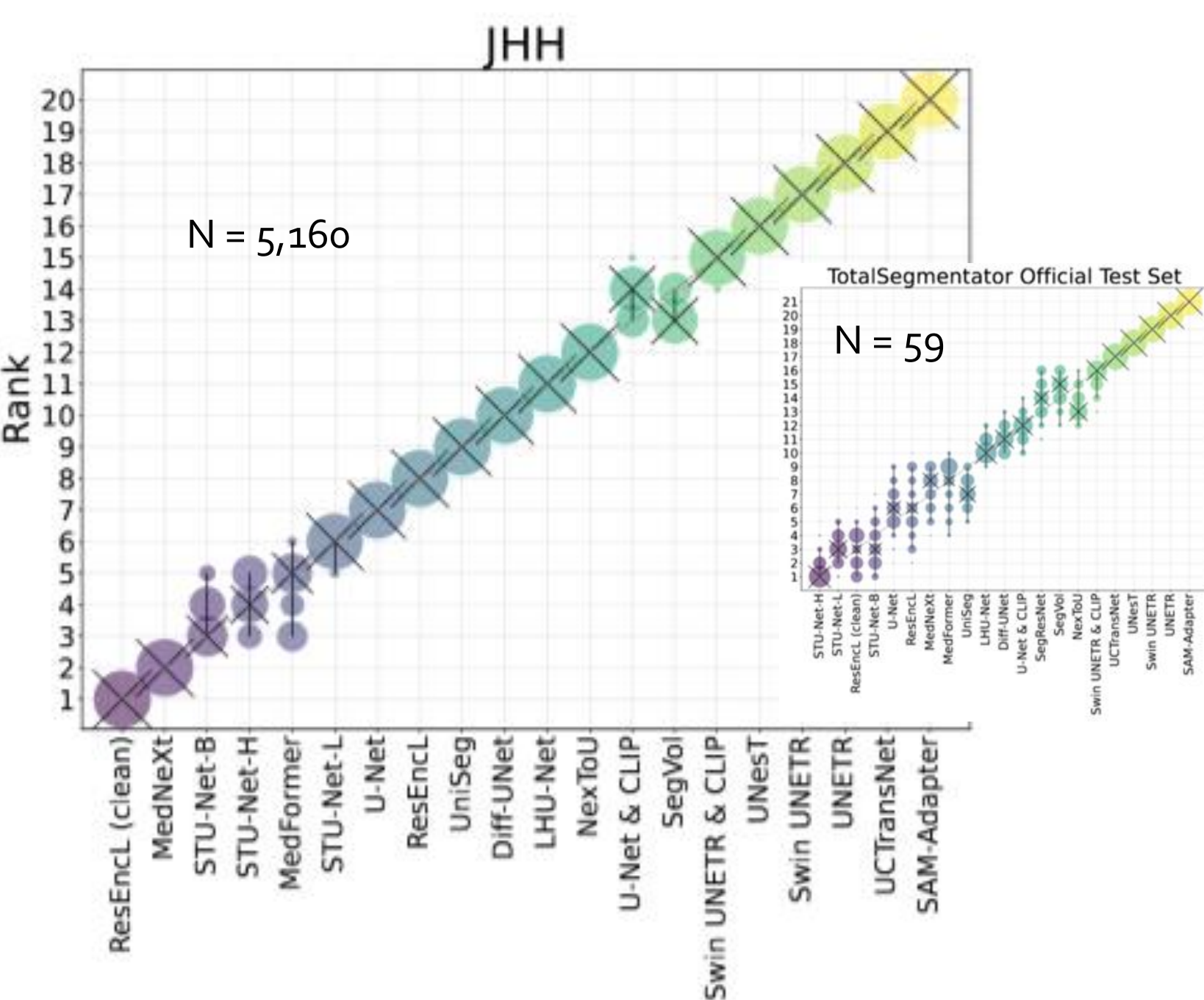
JHH results on 5,160 CT scans								TotalSegmentator results on 743 CT scans					
framework	architecture	param	spleen	kidneyR	kidneyL	gallbladder	liver	spleen	kidneyR	kidneyL	gallbladder	liver	
nnU-Net	UniSeg [†] [83]	31.0M	94.9±6.0	92.2±7.2	91.5±7.0	84.7±12.6	96.1±4.4	89.4±19.4	84.5±23.8	81.9±27.9	74.6±27.3	91.7±16.5	
	MedNeXt [64]	61.8M	95.2±6.3	92.6±7.4	91.8±7.3	85.3±12.9	96.3±4.5	91.6±18.2	85.5±24.7	86.0±23.8	75.8±28.4	93.0±15.8	
	NexToU [66]	81.9M	94.7±8.1	90.1±9.5	89.6±9.3	82.3±17.0	95.7±5.5	83.0±29.5	78.2±32.7	78.7±30.8	72.0±31.1	87.6±23.0	
	STU-Net-B [34]	58.3M	95.1±6.4	92.5±7.3	91.9±7.2	85.5±12.3	96.2±4.8	92.3±15.3	87.1±20.2	86.8±22.1	78.5±24.9	93.0±13.9	
	STU-Net-L [34]	440.3M	95.2±6.1	92.5±7.1	91.8±7.1	85.7±11.8	96.3±4.4	91.6±17.8	88.2±18.5	86.3±22.9	78.1±24.6	94.2±11.2	
	STU-Net-H [34]	1457.3M	95.2±5.9	92.6±6.9	91.9±7.1	86.0±11.6	96.3±4.4	92.4±14.6	88.9±16.2	86.5±23.4	77.7±25.3	94.0±11.4	
	U-Net [62]	31.1M	95.1±6.3	92.7±6.9	91.9±7.2	84.7±13.1	96.2±4.5	91.2±17.8	88.4±18.3	87.3±20.8	78.3±25.5	93.4±13.8	
	ResEncL [35, 37]	102.0M	95.2±6.3	92.6±7.0	91.9±6.9	84.9±13.0	96.3±4.5	91.8±17.5	88.9±18.0	88.2±20.5	78.0±25.1	91.7±18.4	
	ResEncL*	102.0M	95.1±6.2	92.7±6.9	91.9±7.1	84.9±12.8	96.3±4.5	92.0±16.7	89.9±15.3	89.5±18.3	78.0±24.7	92.4±17.4	
	Vision-Language	U-Net & CLIP [46]	19.1M	94.3±6.9	91.9±7.8	91.1±8.8	82.1±15.4	96.0±4.3	87.4±22.8	83.6±25.5	82.7±26.6	73.1±29.0	91.6±14.8
	Swin UNETR & CLIP [46]	62.2M	94.1±7.7	91.7±9.1	91.0±9.1	80.2±18.3	95.8±5.6	87.1±23.4	81.1±28.9	77.0±32.3	70.3±30.9	91.7±16.0	
MONAI	LHU-Net [65]	8.6M	94.9±6.3	92.5±7.0	91.8±7.4	83.9±14.5	96.2±4.3	86.0±25.7	81.8±29.3	82.4±26.9	71.3±32.0	87.7±22.9	
	UCTransNet [72]	68.0M	90.2±11.9	86.5±14.6	86.9±12.8	77.8±19.5	93.6±6.4	76.4±34.5	74.3±35.1	62.0±41.4	69.6±31.8	82.6±28.1	
	Swin UNETR [68]	72.8M	92.7±7.8	89.8±11.1	89.7±10.2	76.9±20.7	95.2±5.3	66.3±36.4	59.7±39.3	58.5±40.1	50.6±40.5	80.2±28.7	
	UNesT [85]	87.2M	90.9±8.1	90.9±8.1	90.1±8.2	75.1±21.2	95.3±5.0	79.5±26.6	70.9±32.3	72.0±33.8	50.3±39.9	87.6±20.8	
	UNETR [25]	101.8M	91.7±10.1	90.1±9.4	89.2±9.6	74.7±20.4	95.0±5.3	60.4±37.9	47.9±39.5	41.9±39.7	40.0±36.7	78.1±29.8	
	SegVol [†] [18]	181.0M	94.5±6.9	92.5±7.1	91.8±7.3	79.3±18.8	96.0±4.7	87.1±23.3	82.8±23.4	82.6±24.8	68.1±29.2	89.4±20.4	
n/a	SAM-Adapter [†] [23]	11.6M	90.5±8.8	90.4±7.9	87.3±9.6	49.4±22.9	94.1±5.3	53.5±33.0	8.5±11.1	19.9±22.0	11.5±17.5	66.4±35.4	
	MedFormer [19]	38.5M	95.5±6.1	92.8±7.3	91.9±7.4	85.3±13.6	96.4±4.4	90.7±15.0	85.5±18.4	84.0±21.5	74.1±26.7	92.8±12.4	
	Diff-UNET [81]	434.0M	95.0±6.9	92.8±7.4	91.9±7.5	83.8±14.8	96.2±4.7	88.3±23.5	81.3±27.9	81.0±28.3	71.8±29.9	92.4±14.8	
framework	architecture	param	stomach	aorta	postcava	pancreas	average	stomach	aorta	IVC [‡]	pancreas	average	
nnU-Net	UniSeg [†] [83]	31.0M	93.3±6.0	82.3±10.3	81.2±8.1	82.7±10.4	88.8±5.0	74.0±29.5	69.2±31.5	72.8±25.8	70.3±30.9	71.8±28.0	
	MedNeXt [64]	61.8M	93.5±6.0	83.1±10.2	81.3±8.3	83.3±11.0	89.2±5.1	77.2±28.7	71.9±30.1	75.2±23.5	71.6±31.4	73.9±27.3	
	NexToU [66]	81.9M	92.7±7.5	86.4±8.7	78.1±9.1	80.2±13.5	87.8±6.2	69.0±34.7	61.5±33.0	59.4±32.7	66.8±31.9	61.4±31.8	
	STU-Net-B [34]	58.3M	93.5±6.0	82.1±10.5	81.3±8.2	83.2±10.7	89.1±5.3	78.6±26.5	74.2±28.9	77.3±19.5	74.9±27.4	76.6±24.9	
	STU-Net-L [34]	440.3M	93.7±5.6	81.0±10.9	81.3±8.2	83.4±10.7	89.0±5.0	79.7±24.6	75.7±26.9	77.6±18.7	75.2±27.0	78.9±21.5	
	STU-Net-H [34]	1457.3M	93.7±5.7	81.1±10.9	81.1±8.2	83.4±10.7	89.1±5.0	78.5±25.5	74.7±28.0	76.9±19.0	74.5±27.5	77.6±23.8	
	U-Net [62]	31.1M	93.3±6.0	82.8±10.2	81.0±8.2	82.3±11.4	88.9±5.1	78.9±26.3	71.0±28.4	76.4±21.8	75.2±26.9	74.4±26.1	
	ResEncL [35, 37]	102.0M	93.4±6.0	81.4±11.1	81.5±8.8	82.9±10.8	88.8±5.1	78.9±25.3	73.8±25.9	76.4±20.1	76.3±25.8	77.8±21.8	
		ResEncL*	102.0M	93.5±5.9	88.0±7.3	80.5±8.7	82.8±11.1	89.5±7.8	80.9±23.0	84.2±20.5	76.3±20.0	77.3±24.9	84.5±20.1
	Vision-Language	U-Net & CLIP [46]	19.1M	92.4±6.8	77.1±12.7	78.5±9.6	80.8±11.5	87.2±5.0	77.7±26.7	59.0±32.8	65.8±27.2	74.6±25.7	67.7±28.4
	Swin UNETR & CLIP [46]	62.2M	92.2±8.3	78.1±12.6	76.8±11.0	80.2±12.5	86.7±6.3	71.2±30.6	58.6±34.5	63.6±27.3	70.3±28.8	64.6±30.7	
MONAI	LHU-Net [65]	8.6M	93.0±6.1	79.5±11.2	79.4±9.3	81.0±11.3	88.1±5.2	71.3±31.8	63.0±34.0	67.5±28.5	68.6±32.5	65.6±31.8	
	UCTransNet [72]	68.0M	81.9±12.9	86.5±8.0	68.1±15.8	59.0±21.6	81.2±8.6	61.6±36.1	49.7±34.8	49.3±36.4	59.0±35.1	48.5±34.4	
	Swin UNETR [68]	72.8M	90.5±8.6	77.2±15.1	75.4±11.8	75.6±14.5	84.9±7.1	52.2±35.1	54.5±36.9	38.1±34.6	42.3±34.4	45.4±31.1	
	UNesT [85]	87.2M	90.9±7.3	77.7±16.1	74.4±11.8	76.2±12.1	85.9±6.2	63.9±31.4	54.7±36.9	38.9±36.2	50.0±32.9	49.4±32.3	
	UNETR [25]	101.8M	88.8±8.4	76.5±16.4	71.5±12.8	72.3±14.5	83.4±7.0	42.1±32.0	41.0±31.3	41.3±32.3	28.2±29.1	37.3±27.9	
	SegVol [†] [18]	181.0M	92.5±7.0	80.2±11.3	77.8±9.7	79.1±12.4	87.2±5.6	71.6±29.8	60.8±29.8	63.0±24.3	66.3±28.0	66.8±26.2	
n/a	SAM-Adapter [†] [23]	11.6M	88.0±9.3	62.8±12.2	48.0±14.2	50.2±12.6	73.8±6.3	48.4±30.9	15.2±18.6	4.8±8.1	30.9±21.7	23.1±19.7	
	MedFormer [19]	38.5M	93.4±6.4	82.1±11.7	80.7±10.1	83.1±11.2	89.0±5.4	80.4±23.6	70.3±28.0	70.0±24.4	72.5±27.9	75.1±24.1	
	Diff-UNET [81]	434.0M	93.1±6.5	81.2±11.3	80.8±8.9	81.9±11.4	88.6±5.5	73.4±29.7	61.0±34.5	60.7±33.3	69.7±29.7	62.5±31.8	

- Performances given as DSC score (mean±s.d.). We bold the best-performing results and highlight the runners-up with no significant difference from the best results at p=0.05 level.
- [†]These architectures were pre-trained on external datasets. [‡]The class IVC (inferior vena cava) shares the same meaning as the class postcava in other datasets.
- *These architectures were trained on AbdomenAtlas 1.0 with enhanced label quality for the aorta and kidney classes.

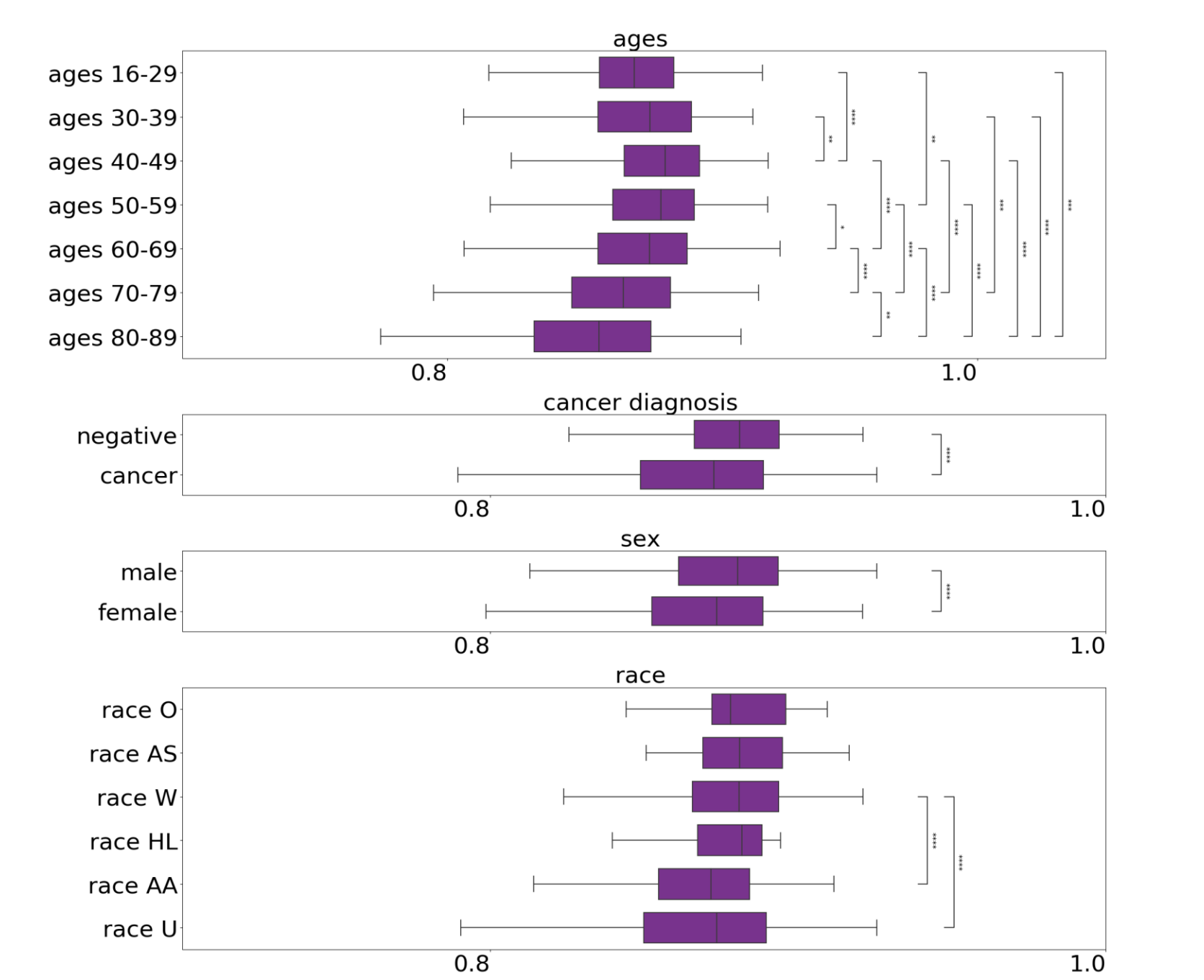
Benchmark Winners 🏆 Average on JHH: MedNeXt · STU-Net · MedFormer

- 🏆 Aorta: UCTransNet · NexToU
- 🏆 Gallbladder: STU-Net
- 🏆 KidneyL: Diff-UNET · MedFormer · STU-Net · U-Net · ResEncL · MedNeXt · LHU-Net · SegVol
- 🏆 KidneyR: MedFormer · Diff-UNET · U-Net · MedNeXt · STU-Net · ResEncL · SegVol
- 🏆 Liver: MedFormer · MedNeXt · STU-Net · ResEncL
- 🏆 Pancreas: STU-Net · MedNeXt · MedFormer
- 🏆 Postcava: STU-Net · MedNeXt · UniSeg · U-Net
- 🏆 Spleen: MedFormer
- 🏆 Stomach: STU-Net · MedNeXt

Observation 1 Larger Test Sets Create Trustworthy Rankings



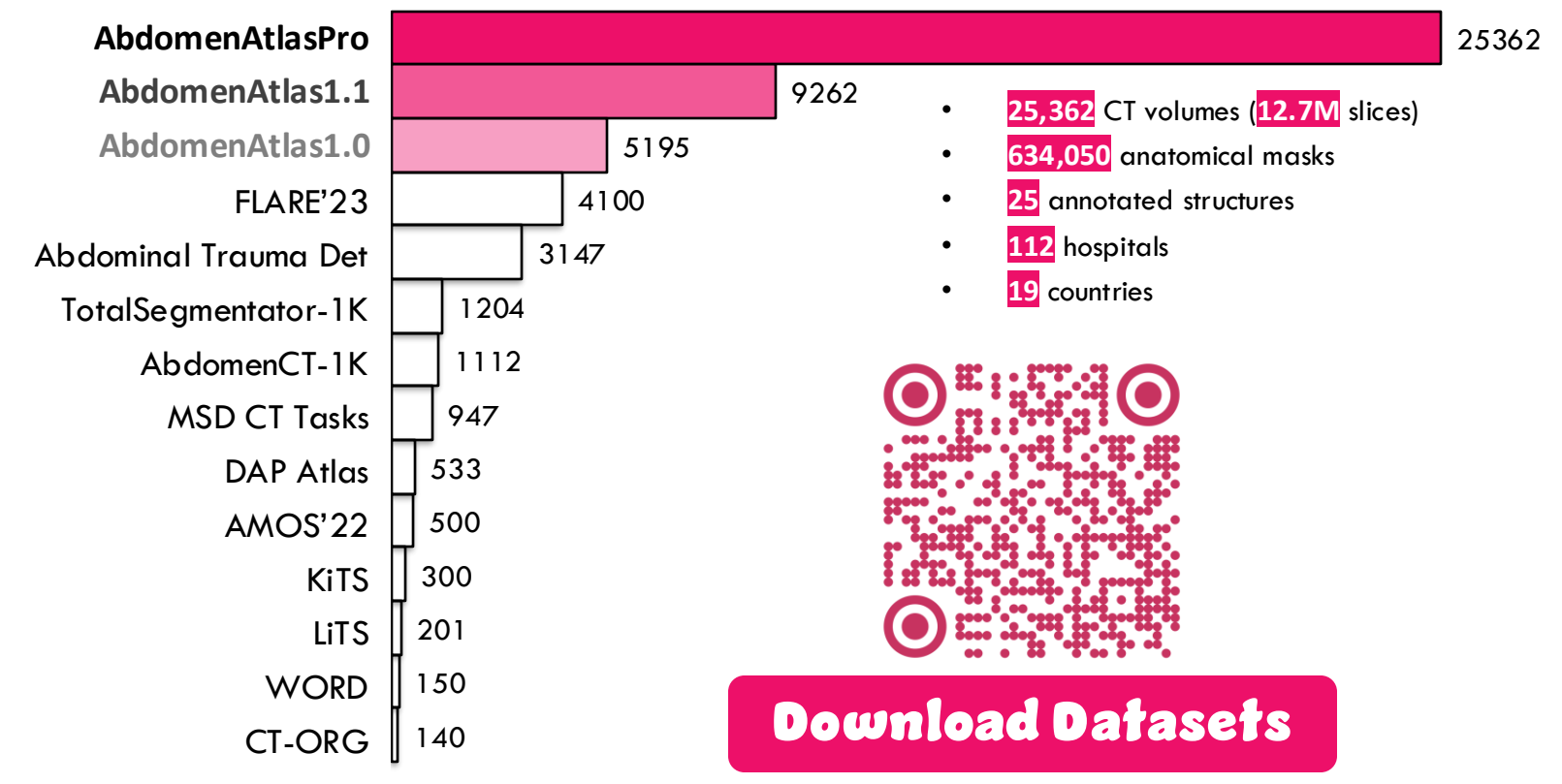
Observation 2 Confounders significantly impact AI performance



Training set AbdomenAtlas 1.0 (5,195 CT scans + 9 classes)

New AbdomenAtlas 1.1 (9,262 CT scans + 25 classes)

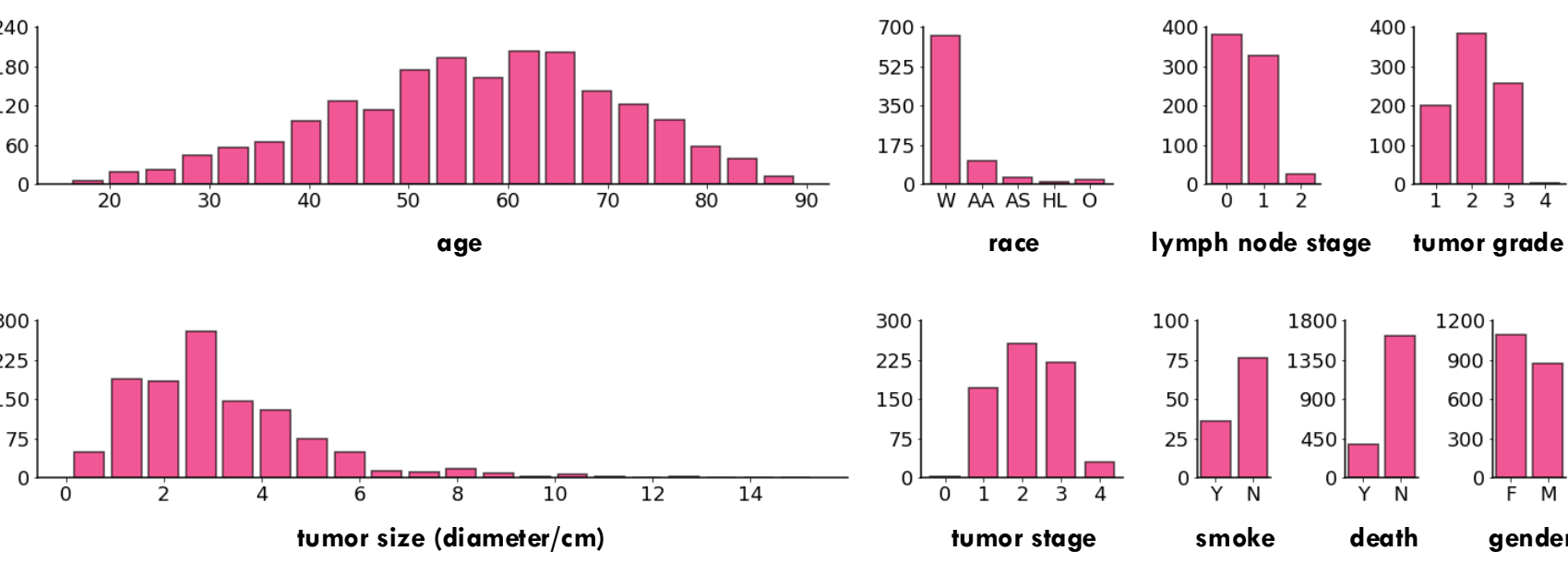
- AbdomenAtlasPro = 25K CT Volumes + 600K 3D Masks
- AbdomenAtlas1.1 = 9K CT Volumes + 225K 3D Masks
- AbdomenAtlas1.0 = 5K CT Volumes + 45K 3D Masks



Download Datasets

Test set JHH (5,160 CT scans) & TotalSegmentator (743 CT scans)

- ★ Evaluating on out-of-distribution data
 - ★ Providing a large test set (N = 5,903)
 - ★ Analyzing pros/cons from multiple perspectives
 - ★ Inviting inventors to train their own algorithms
 - ★ Evaluating new algorithms with long-term commitment
- 14 research teams from 29 institutions and 8 countries participated.



BodyMaps @bodymaps317 · Sep 17 Patient is experiencing colonic obstruction, leading to compression and displacement of multiple intra-abdominal organs.

BodyMaps @bodymaps317 · Sep 24 This patient has multiple cystic lesions in the right kidney and a mass in the left kidney.

BodyMaps @bodymaps317 · Nov 6 The patient has undergone bilateral femoral replacement, which is causing localized metal artifacts that affect the imaging clarity of the bladder and parts of the pelvic bowel loops. Additionally, the patient is post-cholecystectomy (gallbladder removal).

Follow @BodyMaps