SIIM CAIMI26 Conference on Artificial Intelligence in Medical Imaging
OCT 26-27 | University of Pennsylvania, Philadelphia, PA

AI-Assisted Radiologist Curation of PANORAMA for PDAC Tumor Segmentation Benchmarking

Khurram Khaliq Bhinder, MBBS, Mayo Clinic; Sovanlal Mukherjee, PhD; Angela Ammirabile, MD, PhD; Armin Zarrintan, MD; Takeru Yamaguchi, MD, PhD; Baloy J. Talukdar, MD; Subhosree Dey, MBBS, MD; Ahmed K. Jadoon, MBBS; Ajit H. Goenka, MD

Introduction/Background
Public imaging datasets can support reproducible AI evaluation when their annotations provide reliable reference standards for the intended task. PANORAMA is a recent international multi-institutional abdominal CT dataset for pancreatic ductal adenocarcinoma (PDAC) detection whose distributed annotations could also support tumor segmentation and imaging biomarker research. We evaluated AI-assisted expert curation as a strategy to establish PANORAMA as a public reference for reproducible tumor segmentation benchmarking.

Methods/Intervention
Radiologists classified PANORAMA PDAC segmentation-CT pairs as evaluable when the tumor was assessable. Evaluable pairs were assessed for image quality sufficient for precise boundary delineation, percentage tumor-volume capture, and over- or under-segmentation subtype. Model-WS, a recent whole-CT nnU-Net, generated draft segmentations for evaluable CTs with adequate image quality, which radiologists corrected to create the reference standard. Two validated tumor segmentation models, Model-BB, a pancreas-localized nnU-Net, and Swin UNETR, a 3D vision transformer, were evaluated against the original segmentations and corrected reference standard. Dice similarity coefficient (DSC) was compared using paired Wilcoxon signed-rank tests.

Results/Outcome
Of 679 segmentation-CT pairs, 614 (90.4%) were evaluable; 65 were non-evaluable due to overlapping reasons: no detectable tumor despite a segmentation (28), stent artifacts (27), chest CT acquisition (7), and other technical limitations (9). Image quality was adequate in 590 evaluable pairs (96.1%); suboptimal contrast timing accounted for 19 of 24 inadequate pairs (79.2%). Among the 590 adequate pairs, 335 segmentations (56.8%) captured ≥90% of the tumor. Among the remaining 255, overlapping deficiencies included lateral under-segmentation (185, 72.5%), craniocaudal truncation (139, 54.5%), and over-segmentation (29, 11.4%). Relative to original segmentations, the corrected reference standard yielded higher DSC for Model-BB (0.68 vs 0.56; difference, 0.12; 95%CI, 0.10-0.13; $P < .001$) and Swin UNETR (0.61 vs 0.52; difference, 0.09; 95%CI, 0.08-0.11; $P < .001$).

Conclusion
AI-assisted expert curation established a radiologist-corrected reference standard for PDAC tumor segmentation benchmarking and yielded higher performance estimates for two independent models than the distributed segmentations. Public release of the curated segmentations and metadata will provide a reproducible benchmark for PDAC segmentation and support secondary imaging-informatics research.

Statement of Impact

This work provides a radiologist-corrected public reference standard that extends PANORAMA from PDAC detection toward reproducible CT tumor-segmentation benchmarking.

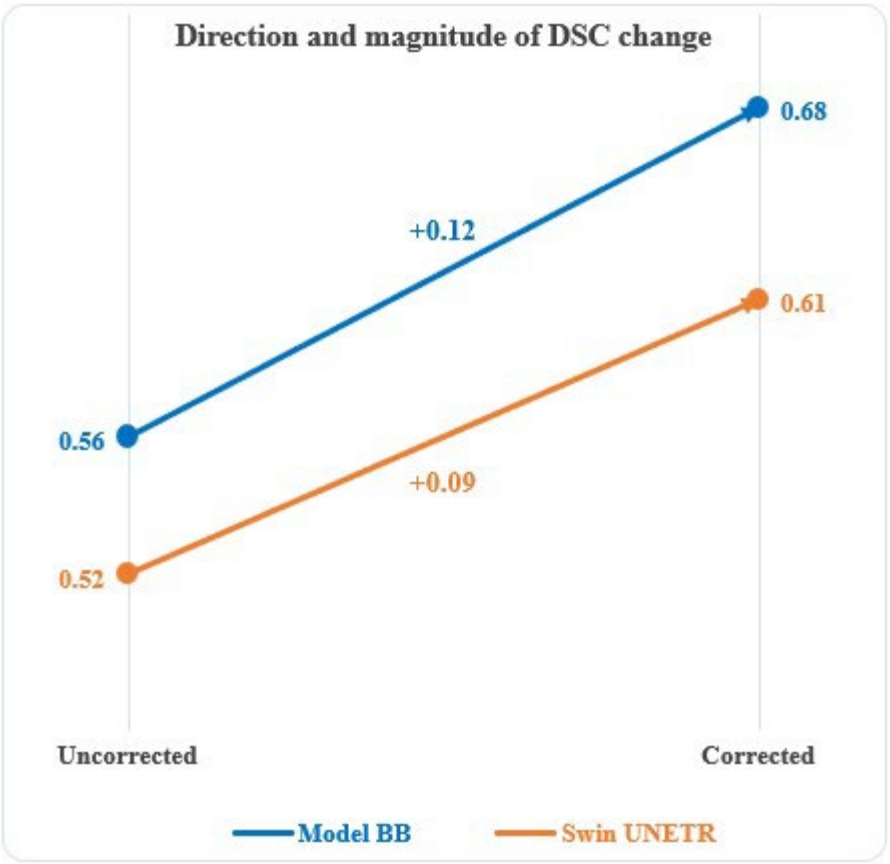

Figure 1. Change in mean Dice Similarity Coefficient (DSC) before and after label correction. Mean DSC increased from 0.56 to 0.68 (+0.12) for Model BB and from 0.52 to 0.61 (+0.09) for Swin UNETR, demonstrating improved segmentation accuracy following correction of annotation errors.

Figure 2. Representative example demonstrating correction of an erroneous tumor segmentation. (A) Original Axial CT image without any segmentation overlay. (B) CT image with fully filled segmentation masks, showing the uncorrected segmentation (green) that incorrectly delineated a vessel thrombus instead of the tumor, and the corrected segmentation (red) accurately identifying the pancreatic tail tumor. (C) CT image with color-coded outlined masks only, highlighting the discrepancy between the uncorrected (green) and corrected (red) segmentations.

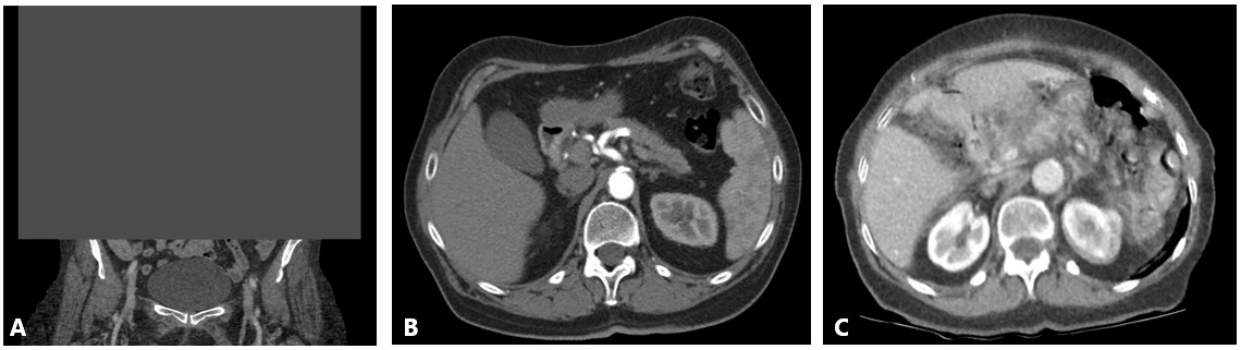

Figure 3. Examples of non-evaluable cases. (A) Corrupted CT image with loss of image data causing non-visualization of the upper abdomen. (B) Arterial-phase CT obtained as part of a lower-extremity CT angiography study, demonstrating opacification of the celiac axis and branches with heterogeneous splenic enhancement and absent portal venous and inferior vena cava opacification. (C) CT degraded by severe motion artifact, resulting in distortion of the left kidney and right-sided ribs, impaired pancreatic tumor localization, and unreliable tumor segmentation.

Keywords

Pancreatic Ductal Adenocarcinoma (PDAC); Tumor Segmentation; Artificial Intelligence; Computed Tomography (CT); Reference Standard Curation; Benchmark Dataset