scientific reports

Check for updates

OPEN

Automated, anatomy-based, heuristic post-processing reduces false positives and improves interpretability of deep learning intracranial aneurysm detection models

Jisoo Kim[1,2,7], Alberto Ceballos-Arroyo[1,3,7], Chu-Hsuan Lin[1], Ping Liu[1,4], Huaizu Jiang[1,3], Shrikanth Yadav[1,5], Qi Wan[1,6], Lei Qin[2,4,8] & Geoffrey S. Young[2,4,8]✉

Deep learning (DL) models can help detect intracranial aneurysms on CTA, but high false positive (FP) rates remain a barrier to clinical translation. We developed a fully automated method to reduce FP by integrating automated in-scan anatomic segmentation and removal of background and venous voxels into hybrid heuristic-DL pipelines. Two DL models, CPM-Net, and a deformable 3D convolutional neural network-transformer hybrid (3D-CNN-TR), trained with 1,186 open-source CTAs (1,373 annotated aneurysms) were integrated into an automated pipeline with heuristic post-processing modules comprising 5 combinations of an open-source brain mask and novel DL-based artery-vein separation modules which create artery, vein, and cavernous venous sinus (CVS) segmentation masks from unlabeled CTA data. FPs that do not overlap with the brain mask and/or overlap with vein or vein-more-than-artery masks were eliminated. Each pipeline was tested on 143 held-out private and 843 publicly available CTAs with 218 and 1027 annotated aneurysms, respectively. On our private dataset, CPM-Net yielded 139 true-positives (TP), 79 false-negative (FN), 126 FP, while 3D-CNN-TR yielded 179 TP, 39 FN, 182 FP. FPs were commonly extracranial (CPM-Net 27.3%; 3D-CNN-TR 42.3%), venous (CPM-Net 56.3%; 3D-CNN-TR 29.1%), arterial (CPM-Net 11.9%; 3D-CNN-TR 53.3%), and non-vascular (CPM-Net 25.4%; 3D-CNN-TR 9.3%) structures. Method 5 (combination of brain and vein-more-than-artery mask) performed best, reducing FP by 70.6% (89/126) and 51.6% (94/182) without reducing TP, lowering the FPR from 0.88 to 0.26, and from 1.27 to 0.62 for CPM-Net and 3D-CNN-TR, respectively. On the public RSNA dataset, CPM-Net yielded 791 TP, 236 FN, 748 FP, 0.89 FPR; while 3D-CNN-TR yielded 940 TP, 87 FN, 1552 FP, 1.84 FPR. Method 1 (enhanced brain mask) performed best at preserving TP, removing 1 and 4 TP for CPM-Net and 3D-CNN-TR, respectively, while removing 31.4% (235/748) and 33.8% (524/1552) of FP. Method 5 eliminated 57.9% and 43.8% of FP, but removed 24 and 25 TP from CPM-Net and 3D-CNN-TR output, respectively. Integration of interpretable, anatomy-based, background and vein removal modules into a fully automated DL-based aneurysm detection pipeline improved model performance on two external test datasets. This suggests that heuristic-DL hybrid pipelines created by integrating in-pipeline domain-informed heuristic post-processing with DL may increase performance and clinical acceptance of radiology domain AI.

[1]Dept of Radiology, Mass General Brigham, Brigham and Women's Hospital, 75 Francis Street, Boston, MA 02115, US. [2]Department of Radiology, Harvard Medical School, Boston, MA 02115, US. [3]Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115, US. [4]Dept of Imaging, Dana-Farber Cancer Institute, Boston, MA 02115, US. [5]Program in Imaging sciences, Department of Biomedical Engineering, Washington University in Saint Louis, St. Louis, MO, USA. [6]First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China. [7]These co-first authors contributed equally: Jisoo Kim and Alberto Ceballos-Arroyo . [8]These co-senior authors provided equal leadership and supervision: Lei Qin and Geoffrey S. Young . ✉email: gsyoung@bwh.harvard.edu

Abbreviations

DL	Deep learning
CNN	Convolutional neural network
CTA	Computed Tomography Angiography
TP	True positive
FP	False positive
FPR	False positive rate
TN	True negative
FN	False negative
CVS	cavernous venous sinus

Intracranial aneurysms are abnormal protrusions from cerebral arteries caused by the weakening of the vessel wall. Subarachnoid hemorrhage due to aneurysm rupture has a 40–50% 30-day mortality[1]. Visual inspection of hundreds of images in each CTA or MRA is time-consuming, tiring, and error-prone, even for expert radiologists. Deep learning (DL) models developed to assist with aneurysm detection have achieved sensitivity over 90%[2,3], but clinical translation is hindered by high false positive (FP) rates (FPR) that can reduce efficiency, overwhelm clinicians and prevent DL model acceptance. A recent meta-analysis of 43 studies reported FPs rates ranging from 0.13 to 31.8 per scan with a pooled FPR of 16.5%[4], underscoring the need for robust FP reduction strategies.

In recently reported FP reduction methods, adjusting confidence detection thresholds reduces the number of FPs per case, but at the cost of lowering sensitivity[5,6]. A dedicated multi-dimensional convolutional neural network (CNN) was reported that reduced the FP/case rate to 5.0 during external testing, but at a relatively low fixed sensitivity of 80%[7]. These are important incremental improvements, but manual parameter optimization specific to each model, and the intrinsically 'black-box' nature of dedicated FP reduction models, limits the generalizability of these approaches. Furthermore, these approaches provide no insight into the causes of FP; therefore, they cannot properly aid expert clinicians in verifying predictions during clinical application, nor do they support researchers in systematically reducing FPR. An alternative heuristic-hybrid post-processing approach has recently produced promising results in a very different medical imaging AI domain, improving performance of AI models designed to detect seizure onset zones in resting state functional MRI data from patients with drug-resistant epilepsy[8,9]. Developing a similar strategy for CTA aneurysm detection requires first identifying anatomic characteristics predictive of AI model FP aneurysm detections. Since no study has systematically analyzed the location, characteristics, and vascular anatomic relationships of FP and true positive (TP) aneurysm detections to date, we performed such an analysis, and used this knowledge to develop systematic FP reduction methods grounded in domain-specific knowledge of aneurysm features and relevant anatomy, and finally integrated the methods into a fully automated anatomy-based, heuristic-DL hybrid post-processing pipeline. This pipeline is designed to be flexible so that any existing or future DL detector can be incorporated by simply substituting the core DL module. To illustrate this flexibility, we used two distinct DL models with different architectures in the pipeline. Our end-to-end pipelines require no human input during inference. In contrast, it is inherently more generalizable and interpretable, because it reduces FP *after* detection. To determine which of several possible pipeline post-processing configurations produces the best performance, we implemented 5 variations of the post-processing module for each DL model (10 pipeline instances total). Each integrated heuristic-DL pipeline comprises: a vascular segmentation module that generates individual patient artery, vein, and/or cavernous venous sinus (CVS) segmentations from the individual inference CTA scan images alone (without human operator interaction or retrieval of pre-produced masks), used to eliminate venous voxels; a brain masking module adapted from open-source code to generate patient-specific brain masks, which is used to eliminate background voxels; and one of two core DL aneurysm detection modules. This hybrid design allows systematic evaluation of how anatomical heuristics can complement DL-based detections to reduce FPs in a fully automated manner.

We demonstrate marked reduction of FPR with preserved sensitivity on 2 separate held-out test datasets. By decoupling FP reduction from the DL model and grounding the FP reduction strategy in relevant anatomic knowledge, this approach also increases transparency and generalizability, which will be critical to increase acceptance of AI models in radiology. We emphasize that although the heuristic modules are external to the DL models, they are incorporated in an integrated pipeline that achieves fully automated end-to-end performance at the time of inference: the pipeline requires only a single patient un-annotated CTA image-set as input, and no human interaction, to output the final aneurysm detection region of interest overlayed on the CTA scans.

Methods

To reduce FP aneurysm detections, we developed a fully automated hybrid pipeline that integrates DL models with anatomically informed heuristic suppression modules. The pipeline first applies a DL-based aneurysm detector to predict candidate bounding boxes in the CTA image. These predictions are then refined using patient-specific anatomical masks, including brain, artery, vein, and CVS masks, to suppress likely FPs based on spatial and vascular context. A flow chart describing our approach is shown in Fig. 1.

DL models

In order to investigate whether our heuristic-DL pipeline reduces FP aneurysm detections independent of the specific DL model, we implemented two previously published unmodified high performing DL CTA models based on significantly distinct DL architectures: (1) CPM-Net, a CNN-based 3D detector[10]; and (2) 3D-CNN-TR, a deformable 3D CNN-Transformer hybrid leveraging artery segmentation data as an auxiliary input[2]. Each model was trained for 45 epochs using the PyTorch library using the AdamW optimizer with a steadily

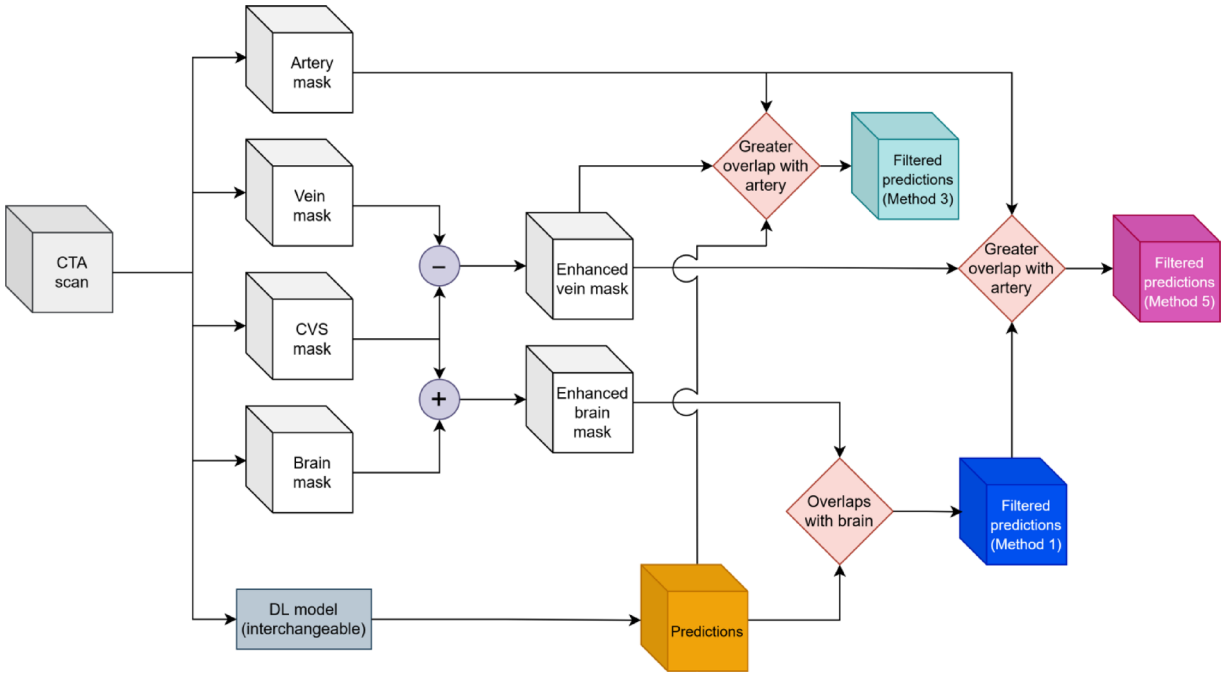

Fig. 1. Flow chart depicting the structure of the fully automated pipeline corresponding to the three best performing methods: method 1 only retains detections that overlap with the enhanced brain mask resulting from adding the CVS region to a mask of the brain; method 3 only retains detections that have greater overlap with the artery mask than with the vein mask; method 5 only retains detections that meet both criteria from methods 1 and 3.

decreasing learning rate, from 0.0001 to 0.00001. Since both models were detectors, we represented aneurysms as their minimally bounding 3D boxes. Specifically, the models were trained to output each aneurysm's location and size (height, width, depth). At inference, we cropped a full volume into a set of contiguous sub-volumes to predict the location of potential aneurysms across the patient's CT scan, which we then stitched together for evaluation[2].

Artery and vein mask

To suppress FPs from the DL models, we applied anatomically informed fully automated post-processing using patient-specific artery and vein segmentation masks. These masks were used in the heuristic venous voxel suppression module to identify and remove bounding boxes that overlapped with venous structures or lacked sufficient overlap with arterial anatomy.

The artery and vein masks were produced from each inference (test) CTA using our publicly available intracranial artery-vein segmentation algorithm (reported separately)[2,13,17]. This algorithm was trained using a nnUNet model on 4D dynamic CTA data, with cerebral veins and arteries annotated via a semi-automatic MRA-based annotation tool[13] (Fig. 2A). Since CVS, an anatomically unique venous structure located at the base of the brain, overlaps spatially with a number of common arterial aneurysm locations, we produced the CVS mask in a fully automated manner using the outputs of the vascular segmentation model, as described below, and subtracted it from the vein mask.

Cavernous venous sinus mask

The patient-specific CVS segmentations were produced within the pipeline from each inference (test) CTA. To localize the CVS region, a bounding box derived from our CTA vascular atlas[13] was aligned to each inference CTA image using affine registration via the Advanced Normalize Tool (ANT) framework. This box was then expanded by 3.2 mm (8 pixels) in 3 dimensions to ensure inclusion of all surrounding anatomical structures (Fig. 2B). The final CVS mask was produced by considering only voxels contained in both this expanded CVS region box and the venous segmentation mask (Fig. 2C and D).

Brain mask

To suppress detections outside the intracranial region, we incorporated patient-specific brain masks into the heuristic post-processing module. These masks were used to remove bounding boxes located outside the brain. Each brain mask was generated from the inference (test) CTA using the TotalSegmentator library (Fig. 3A)[14]. This mask was dilated 3.6 mm (9 pixels) in 3 dimensions to ensure inclusion of all intracranial structures and then combined with the CVS region bounding box to ensure inclusion of appropriate skull base structures (Fig. 3B).

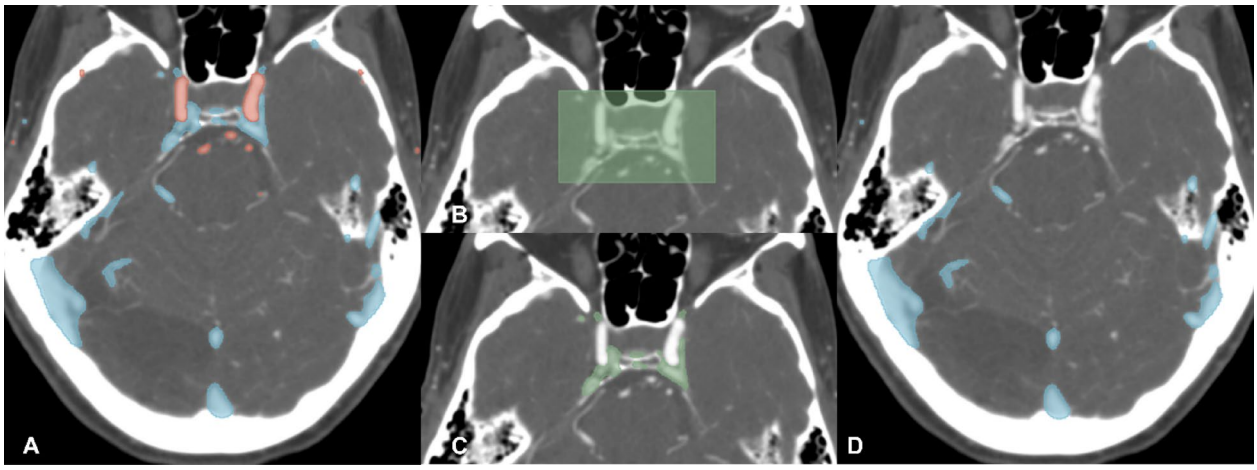

Fig. 2. Arterial (red) and venous (blue) segmentation masks of intra and extracranial structures (**A**). The cavernous venous sinus (CVS) region box derived from an atlas, registered to the same CTA using the ANT registration framework (**B**). The CVS mask defined by overlap between the CVS region box and the venous mask (**C**). Vein mask with CVS mask subtracted (**D**).

Fig. 3. Brain segmentation mask using TotalSegmentator (**A**), dilated 3.6 mm in each direction (yellow in B). The cavernous venous sinus (CVS) region bounding box from Fig. 2B (green) was added to the brain mask to ensure inclusion of the appropriate skull base structures (**B**).

Post-processing

To determine the most efficient combination of these heuristic extracranial and venous voxel suppression modules, 5 different instances of the fully automated pipeline were produced for each DL model (10 pipelines total). In method 1, the heuristic module only keeps bounding boxes located inside of the brain mask; in method 2, the module removed bounding boxes having any overlap with the vein mask; in method 3 the module only keeps bounding boxes that overlapped with the artery mask more than the vein mask; in method 4, the modules from methods 1 and 2 were combined; in method 5, the modules from methods 1 and 3 were combined (Table 1).

Data

The experimental protocol of this retrospective medical records study complies with all relevant regulations and ethical guidelines, including particularly HIPAA regulations and the Declaration of Helsinki, and was approved by the Mass General Brigham (MGB) Human Research Committee (HRC) Institutional Review Board (IRB) with a waiver of the requirement for informed consent (MGB HRC IRB Protocol #2022P000792). The training dataset for each model consisted of 1,186 CTAs with 1,373 annotated aneurysms, available online in the Large IA Segmentation dataset repository, https://zenodo.org/records/6801398[11]. Held-out private evaluation data were

	Heuristic modules KEEP bounding boxes from DL output that were:
Method 1	Located inside the brain mask
Method 2	Not overlapping with any voxel of the vein mask
Method 3	Overlapping with more voxels of the artery mask than the vein mask
Method 4	Kept by both methods 1 and 2
Method 5	Kept by both methods 1 and 3

Table 1. Post-processing methods (1–5) for filtering DL model outputs.

gathered through an internal medical record search tool that queried for the diagnosis of "cerebral aneurysm" in the records of 9 hospitals within our system from 2005 to 2016, and are available from the corresponding author on reasonable request. A randomly selected subset of the CTAs and corresponding radiology reports were reviewed, and the aneurysms were annotated as bounding boxes on axial sub-millimeter source images by a radiologist with 10 years of experience (QW) on an internal research-based DICOM viewer. For additional validation, these were reviewed and edited by a subspecialty-trained neuroradiologist with 4 years of subspecialty experience (JK) on 3D Slicer (http://www.slicer.org)[12]. For any discrepancies, a subspecialty-trained neuroradiologist of 25 years of subspecialty experience (GY) made the final decision on the presence and location of aneurysms. Part of this evaluation data was also used for a previously reported paper on 3D-CNN-TR[2].

Analysis

Ground truth annotation for analysis was produced by a research assistant (CL) and a subspecialty-trained neuroradiologist with 5 years of subspecialty experience (PL) who independently reviewed parts of the model outputs on 3D Slicer. All model outputs were subsequently reviewed by a subspecialty-trained neuroradiologist with 4 years of subspecialty experience (JK). An output was considered a TP if the bounding box was centered around the annotated aneurysm with the appropriate box size. Otherwise, output was recorded as a FP. Location and description of each FP was recorded. Outputs noted as "caliber change" included stenoses or branchpoints. Outputs noted as "vessel" included portions of small vessels that were difficult to identify as arterial or venous.

Test data

Private data

Medical record search for cerebral aneurysms yielded 7,749 CTAs, of which 5,136 were successfully uploaded to the internal research-based DICOM viewer. A total of 143 CTAs were randomly selected. These contained 34 males and 109 females with an average age of 60.9 years. A total of 218 aneurysms were annotated, including 123 internal carotid artery (ICA), 40 middle cerebral artery (MCA), 31 anterior communicating artery (Acomm), 8 basilar, 7 superior cerebellar artery (SCA), 3 anterior cerebral artery (ACA) distal to Acomm, 2 posterior inferior cerebellar artery (PICA), and 1 posterior cerebral artery (PCA) aneurysms.

RSNA data

The Radiological Society of North America (RSNA) released a collection of 4348 scans comprising CTA and MRA for its Intracranial Aneurysm Detection AI Challenge (https://www.kaggle.com/competitions/rsna-intracranial-aneurysm-detection/)[17], with coordinate annotations for 2,254 aneurysms. For our study, we filtered the dataset to include only CTA volumes with at least 100 slices, resulting in a subset of 843 scans with 1027 aneurysms. Since our evaluation approach requires bounding boxes rather than point coordinates, we used an ensemble of three nnDetection[18] models (trained on the Large IA Segmentation Dataset[11] using 3-fold cross-validation) to generate 3D bounding box annotates. For each aneurysm, we accepted a bounding box if all three models agreed with confidence > 0.5, and the predicted box sizes were within 1 mm of the ensemble mean across all dimensions. Additionally, the predicted bounding boxes had to encompass the ground truth coordinate. Cases where the models failed to reach consensus were manually annotated by a subspecialty-trained neuroradiologist of 25 years of subspecialty experience (GY) using an in-house developed viewer.

Results

Of the 218 ground truth annotations in our private test set, using a confidence threshold of 0.8, CPM-Net detected 139 TP, 79 FN, and 126 FP, for a FPRof 0.88. 3D-CNN-TR detected 179 TP, 39 FN, 182 FP, for a FPR of 1.27 (Table 2). Most of the FPs from CPM-Net were in venous structures (71/126, 56.3%) while those from 3D-CNN-TR were in arterial structures (97/182, 53.3%) (Tables 3 and 4). 27/99 (27.3%) CPM-Net FP and 77/182 (42.3%) of 3D-CNN-TR FP were extracranial.

In the RSNA dataset, using a confidence threshold of 0.8, CPM-Net detected 791 TP, 236 FN, and 748 FP, for a FPR of 0.89; 3D-CNN-TR detected 940 TP, 87 FN, and 1,552 FP, for a FPR of 1.84. While a similarly detailed analysis of FP locations was not possible due to the scale of the dataset, we note that the brain mask removes 235/748 (31.4%) FP for CPM-Net and 524/1552 (33.8%) for the 3D-CNN-TR model.

In our private dataset, the most common venous structure identified as FP were the vein of Galen (33/71, 46.4%) on CPM-Net, and extracranial veins (35/53, 66.0%) on 3D-CNN-TR. Arterial structures accounted for 15 FP on CPM-Net, and 97 FP on 3D-CNN-TR. 3D-CNN-TR identified a wider variety of normal or abnormal structures as FP, compared to CPM-Net. Structures detected as FP by 3D-CNN-TR and not by CPM-Net include basilar tip confluence (8), vertebral artery caliber change (2), PCA caliber change (1), daughter aneurysm (1), and artifacts such as aneurysm clips or calcifications (4). Structures detected as FP more commonly on 3D-CNN-TR

			None	Method 1	Method 2	Method 3	Method 4	Method 5
Private	CPM-Net	TP	139	139	129	139	129	139
		FP *(FP/case)*	126 *(0.88)*	98 *(0.69)*	43 *(0.30)*	48 *(0.34)*	33 *(0.23)*	37 *(0.26)*
		FN	79	79	89	79	79	79
	3D-CNN-TR	TP	179	179	169	179	169	179
		FP *(FP/case)*	182 *(1.27)*	104 *(0.73)*	98 *(0.69)*	116 *(0.81)*	79 *(0.55)*	88 *(0.62)*
		FN	39	39	49	39	39	39
RSNA	CPM-Net	TP	791	790	691	767	690	766
		FP *(FP/case)*	748 (0.89)	513 (0.61)	499 (0.59)	517 (0.61)	304 (0.36)	315 (0.37)
		FN	236	237	366	260	337	261
	3D-CNN-TR	TP	940	936	829	917	827	914
		FP *(FP/case)*	1,552 (1.84)	1,028 (1.22)	979 (1.16)	1,178 (1.40)	799 (0.95)	872 (1.03)
		FN	87	91	198	110	200	113

Table 2. Performance of the 10 CPM-Net and 3D-CNN-TR based pipelines (with a fixed confidence threshold of 0.8) and post-processing methods 1–5, for the private and RSNA datasets.

		FP removed by				
	All FP	Method 1	Method 2	Method 3	Method 4	Method 5
Vein	71	11	67	67	67	67
Vein of Galen	*33*	*0*	*33*	*33*	*33*	*33*
Venous sinus	*8*	*0*	*8*	*8*	*8*	*8*
Cavernous venous sinus	*3*	*0*	*0*	*0*	*0*	*0*
Other intracranial veins	*17*	*1**	*16*	*16*	*16*	*16*
Extracranial veins	*10*	*10*	*10*	*10*	*10*	*10*
Artery	15	2	6	1	6	2
Cervical artery	*2*	*2*	*2*	*1*	*2*	*2*
ICA infundibulum/branch points	*2*	*0*	*0*	*0*	*0*	*0*
MCA branch points	*3*	*0*	*1*	*0*	*1*	*0*
Caliber change at A1/A2 or ACA	*4*	*0*	*1*	*0*	*1*	*0*
Vertebrobasilar confluence	*2*	*0*	*2*	*0*	*2*	*0*
ICA occlusion	*1*	*0*	*0*	*0*	*0*	*0*
Ectatic basilar artery	*1*	*0*	*0*	*0*	*0*	*0*
Vessel	8	7	5	5	8	8
Intracranial	*1*	*0*	*1*	*1*	*1*	*1*
Extracranial	*7*	*7*	*4*	*4*	*7*	*7*
Tissue	32	8	5	5	12	12
Choroid plexus	*19*	*0*	*2*	*2*	*2*	*2*
Pineal gland	*4*	*0*	*2*	*2*	*2*	*2*
Other intracranial	*1*	*0*	*0*	*0*	*0*	*0*
Extracranial	*8*	*8*	*1*	*1*	*8*	*8*
Total	126	28	83	78	93	89
Intracranial	*99*	*1*	*66*	*62*	*66*	*62*
Extracranial	*27*	*27*	*17*	*16*	*27*	*27*

Table 3. Analysis of the false positives predicted by the CPM-Net model output before and after post-processing, for our private dataset (*: foramen magnum, ICA: internal carotid artery, A1: A1 segment of anterior cerebral artery, A2: A2 segment of anterior cerebral artery, ACA: anterior cerebral artery).

than CPM-Net included: more cervical arteries on 3D-CNN-TR (22) than CPM-Net (2), more ICA infundibula or branchpoints on 3D-CNN-TR (37) than CPM-Net (2), more MCA branchpoints on 3D-CNN-TR (10) than CPM-Net (3), and more IAC occlusions on 3D-CNN-TR (5) than CPM-Net (1). Caliber changes at the junction of A1 and A2 segments of the ACA or other areas of ACA were identified at similar rates by 3D-CNN-TR (5) and CPM-Net (4). The two models' FP included the same number of vertebrobasilar confluences (2) and ectatic

	All FPs	FPs removed by				
		Method 1	Method 2	Method 3	Method 4	Method 5
Vein	53	36	49	48	49	49
Vein of Galen	*12*	*0*	*12*	*12*	*12*	*12*
Venous sinus	*0*	*0*	*0*	*0*	*0*	*0*
Cavernous venous sinus	*4*	*0*	*0*	*0*	*0*	*0*
Other intracranial veins	*2*	*1**	*2*	*2*	*2*	*2*
Extracranial veins	*35*	*35*	*35*	*34*	*35*	*35*
Artery	97	22	22	7	30	22
Cervical artery	*22*	*22*	*14*	*7*	*22*	*22*
ICA infundibulum/branch points	*37*	*0*	*1*	*0*	*1*	*0*
MCA branch points	*10*	*0*	*1*	*0*	*1*	*0*
Caliber change at A1/A2 or ACA	*5*	*0*	*1*	*0*	*1*	*0*
Vertebrobasilar confluence	*2*	*0*	*2*	*0*	*2*	*0*
ICA occlusion	*4*	*0*	*0*	*0*	*0*	*0*
Ectatic basilar artery	*1*	*0*	*0*	*0*	*0*	*0*
Basilar tip confluence	*8*	*0*	*1*	*0*	*1*	*0*
Vertebral artery caliber change	*2*	*0*	*0*	*0*	*0*	*0*
PCA caliber change	*1*	*0*	*1*	*0*	*1*	*0*
Daughter aneurysm	*1*	*0*	*0*	*0*	*0*	*0*
Artifact (e.g. clip, calc)	*4*	*0*	*1*	*0*	*1*	*0*
Vessel	15	12	6	5	14	13
Intracranial	*3*	*0*	*2*	*1*	*2*	*1*
Extracranial	*12*	*12*	*4*	*4*	*12*	*12*
Tissue	17	8	7	6	10	10
Choroid plexus	*4*	*0*	*2*	*2*	*2*	*2*
Pineal gland	*0*	*0*	*0*	*0*	*0*	*0*
Posterior clinoid process	*3*	*0*	*0*	*0*	*0*	*0*
Other intracranial	*2*	*0*	*0*	*0*	*0*	*0*
Extracranial	*8*	*8*	*5*	*4*	*8*	*8*
Total	182	78	84	66	103	94
Intracranial	*105*	*1*	*26*	*17*	*26*	*17*
Extracranial	*77*	*77*	*58*	*49*	*77*	*77*

Table 4. Analysis of the false positives predicted the 3D-CNN-TR model output before and after post-processing, for our private dataset (*: foramen magnum; ICA: internal carotid artery, A1: A1 segment of anterior cerebral artery, A2: A2 segment of anterior cerebral artery, ACA: anterior cerebral artery, PCA: posterior cerebral artery).

basilar arteries (1). Small vessels (intracranial and extracranial) were more commonly identified as aneurysms with 3D-CNN-TR (15) than CPM-Net (8). CPM-Net and 3D-CNN-TR both detected the choroid plexus as FP (19 and 4). Only CPM-Net falsely identified the pineal gland (4), while only 3D-CNN-TR falsely detected the posterior clinoid process (3).

Integration in the fully automated pipeline employing the brain mask-based background suppression module only (method 1) did not remove any TP on the CPM-Net or 3D-CNN-TR outputs for our private dataset, and removed < 0.5% of the TP detected by both models in the RSNA dataset. The pipelines employing venous voxel suppression using the vein mask without the artery mask (method 2) removed ~ 7% of the TP detected in the private dataset, and ~ 12% of the TP detected in the RSNA dataset. The pipeline employing venous voxel suppression using vein and artery masks (method 3) did not remove any TP in the private dataset, although it removed 2.5%–3% of TP predicted by both models in the RSNA dataset. Pipelines combining the post-processing approaches (methods 4, 5) aggregated the effects of each method.

In Fig. 4, we report the model response to all possible confidence thresholds with a Sensitivity vs. FPR curve. We observe that, for any fixed tolerance for FP, method 1 significantly improves upon the baseline provided by each model on both datasets, and methods 3 and 5 surpass the baselines in all scenarios except for the 3D-CNN-TR model on the RSNA dataset. Moreover, method 5 consistently results in greater sensitivity for all models and datasets in the low FP tolerance regime (0 to 0.5 FPR). Method 1, which uses only the enhanced brain mask (Fig. 2C) surpasses method 5 in the medium to high FP tolerance regime (0.5 to 8 FPR) when applied to the 3D-CNN-TR model predictions for the RSNA dataset. Taking these results into account, we recommend using either method 1 or method 5 based on the time constraints of the radiologist that would review the pipeline's outputs, with method 1 providing slightly greater coverage of TP detections and method 5 further reducing the number of FP that need to be reviewed.

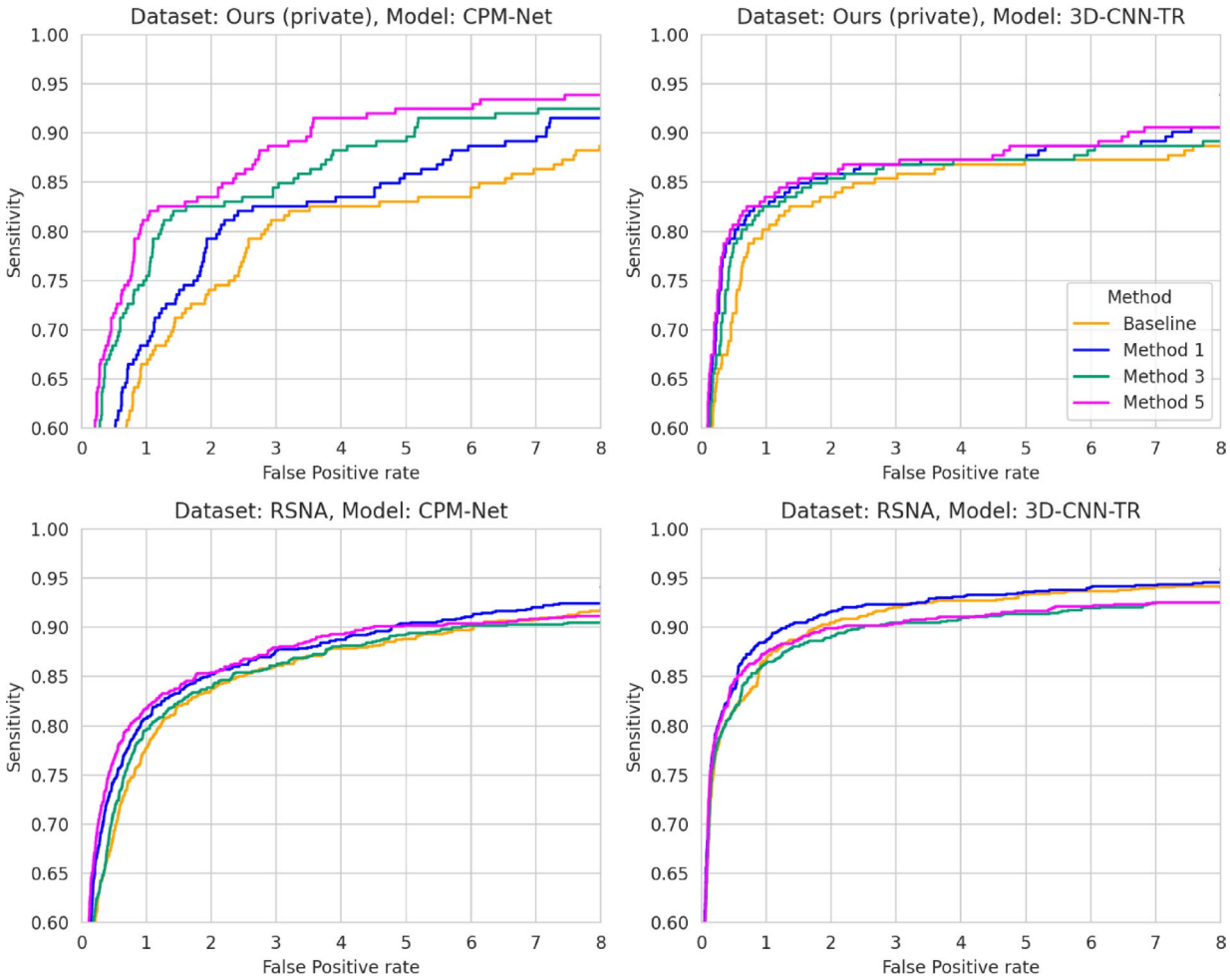

Fig. 4. AUROC (Sensitivity vs. False Positive rate) curves for the three best performing methods, compared with the baseline (model outputs without post-processing), on two external evaluation datasets. We omit methods 2 and 4 due to their lower performance and to enhance readability.

Discussion

Reducing the FPR of AI aneurysm detection pipelines will be critical for successful clinical translation of CTA aneurysm detection models, and other radiology domain AI models, because FPs waste time, cause 'alarm fatigue', deter radiologists from using or trusting the models, and lead to unnecessary follow-up imaging and clinic visits. Improving the balance between sensitivity and FPR may be achieved in specific instances by additional training or adjusting model parameters, but the generalizability, robustness, and stability of such interventions are uncertain. Equally importantly, these approaches do not assist with and may complicate detection and understanding of the decline in machine learning model performance over time ('drift'), which remains a critical barrier to clinical AI translation[15].

In contrast, methods that integrate domain-specific prior knowledge have shown promise in another medical imaging domain[8,9] and promise to increase interpretability, verifiability, and generalizability. To achieve similar advances in the domain of AI CTA aneurysm detection required three distinct steps: first, developing a systematic understanding of the causes of AI model FP by performing a thorough analysis of all FP cases from two of the highest-performing intracranial aneurysm detection models. Second, applying this knowledge to create automated vascular and brain anatomy-based heuristic post-processing software modules by exploiting vascular, bone, and soft-tissue masks. Third, realizing the benefits of these methods without impairing the end-to-end usability essential to successful clinical translation of AI software, required integrating these modules along with a selected AI detection model in a fully automated pipeline.

We validated that the resulting integrated and fully automated heuristic-DL pipelines, deployed with 2 distinct AI models on 2 separate held-out test datasets, achieve highly accurate aneurysm detection at inference from CTA inputs, and that the post-processing modules in the pipeline decreased FPR with minimal reduction in Sensitivity. Crucially, although it was developed based on insights obtained from our smaller, private dataset, our approach (and, particularly, method 1) generalized very well to a much larger publicly available dataset without any adjustment, modification, or retraining. With CPM-Net and 3D-CNN-TR respectively, 27.3 and 42.3% of FP from the private dataset and 31.4 and 33.8% from the RSNA dataset occurred outside the brain.

Many were associated with cervical arteries, facial veins, or other vessels. FP resulting from lens, cartilaginous or other non-vascular structures, may be due to the high density of these structures on CT. By augmenting the original open-source brain mask with addition of the CVS region box, we ensured inclusion of TP aneurysms near to or within the CVS, and made possible automated removal of all extracranial FP with minimal decrease in Sensitivity.

The most common cause of FP cases with CPM-Net and the second most common with 3D-CNN-TR were venous structures. We removed the CVS region from the venous mask because a significant number of TP arises in or adjacent to the CVS. Understandably, FP at the highly vascular choroid plexus were removed by the venous masks. More surprisingly, FP at the pineal gland were also removed, possibly due to their proximity to internal cerebral veins. The pipelines employing our venous masks, comprising all venous structures except the CVS, removed 94% (67/71) of CPM-Net, and 92% (49/53) of 3D-CNN-TR venous FP in the private dataset.

In locations where the venous and arterial masks are in close proximity, we investigated two different methods to reduce FP without inadvertently removing TP aneurysm detections. Method 2 was to remove outputs that overlapped with the venous mask. Method 3 was to remove outputs that had greater overlap with the venous mask than the arterial mask. Pipelines integrating method 3 effectively removed venous FP while removing no TP for the private dataset and under 3% of the TP for the RSNA dataset.

FP related to intracranial arterial structures were the most challenging. With 3D-CNN-TR, this was the most common category of FP. Although some of the FP represented focal arterial pathologies such as occlusion, stenosis, or calcification, many represented normal branch points or infundibula. We note that detection of these abnormalities (particularly vessel occlusions or significant stenoses) is often of clinical significance and may provide an ancillary benefit to somewhat offset the negative impact of the FP detection. This is not true of normal branch points, typically characterized by caliber changes (e.g., basilar tip or vertebrobasilar confluence), or mildly ectatic areas, that were also detected by both models. Although a few were removed by pipelines employing methods 2 and 4, which removed output that had any overlap with the vein mask, overall these were not effectively removed, likely due to the close proximity of the veins and arteries in these locations.

Limitations of our study include the fact that we only carry out a detailed analysis of TP and FP locations for the smaller private dataset. Although the pipeline is fully automated and requires no human annotation during inference, as in all AI research, testing requires expert human-annotated ground-truth data that is costly to prepare. A larger scale analysis that tests the effect of our hybrid pipeline on a wider variety of aneurysm detection models is planned to allow us to extend the analysis by investigating generalizability to different AI model types. Likewise, we note the need to test more methods for segmenting arteries and veins. Since several of our post-processing methods rely on such segmentations, we may further reduce the amount of removed TP by using higher quality segmentations.

Other important future directions include developing hybrid models or post-processing methods that can effectively parse non-pathologic but not perfectly cylindrical arterial structures such as branchpoints and infundibula. Additional post-processing methods can likely be developed to remove other FP cases such as the posterior clinoid process and cases of choroid plexus and pineal gland not removed by the venous template. The tissue background and venous suppression techniques we report can be thought of as methods of post-processing CTA data to mimic MRA and DSA, which benefit from bone, soft tissue, and venous structure suppression intrinsic to the image acquisition. Hence, our pipeline and masks are unlikely to be directly beneficial to AI models for aneurysm detection on MRA or DSA, in which background suppression (by static spin saturation pulses in MRA and subtraction in DSA) and venous structure suppression (by marching inferior saturation bands in MRA, and dynamic acquisition in DSA) makes background tissue and venous masks less critical. Nevertheless, very short T1-relaxation time materials such as intracranial hemorrhage, fat, and bone can create back-ground 'T1-shine-through' artifacts on MRA, and patient motion can result in bone and soft tissue subtraction artifacts in DSA. Similarly, in MRA venous structures may sometimes appear hyper-intense due to tortuous and/or caudal-to-cephalad oriented flow and/or non-laminar and fast flow. In such cases, background and venous tissue suppression masks may increase AI model performance. Investigating this will require future work starting with detailed analysis of FP in MRA and DSA aneurysm detection model outputs, and the design and validation of a mask-based heuristic pipeline appropriate for each modality.

In conclusion, we demonstrate design, production, and testing of fully automated, hybrid heuristic-DL software pipelines integrating domain-specific, anatomy-based, heuristic post-processing modules that operate at inference without human interaction to allow end-to-end detection of aneurysms from unlabeled CTA data. The pipelines markedly reduced FPR with a marginal impact on Sensitivity, as measured on the outputs of two high-performing contemporary CTA aneurysm detection DL models on 2 separate external test datasets. This illustrates the potential of such integrated hybrid heuristic-DL pipelines in general. Such approaches provide the additional value of being highly interpretable by domain experts, and can be easily adapted to incorporate newer improved AI models without substantial alteration to the pipeline. This strategy promises to be a significant step toward increasing radiologist acceptance, efficiency benefit, and trust in AI assistant models, addressing critical barriers to clinical translation of radiology domain AI. Further research is indicated to investigate whether such strategies can improve generalizability and aid detection, assessment, and remediation of AI pipeline performance over time.

Data availability

The training datasets used during the current study are available in the 'Large IA Segmentation' dataset repository, https://zenodo.org/records/6801398 [11]. The private evaluation dataset generated during the current study is available from the corresponding author on reasonable request. The RSNA dataset used for evaluation is available on Kaggle, [https://www.kaggle.com/competitions/rsna-intracranial-aneurysm-detection](https:/www.kagg

le.com/competitions/rsna-intracranial-aneurysm-detection) [17], we will make our filtering and pre-processing scripts for this dataset available online.

Received: 29 August 2025; Accepted: 16 December 2025
Published online: 22 December 2025

References

1. Park, S. W. et al. Short- and long-term mortality of subarachnoid hemorrhage according to hospital volume and severity using a nationwide multicenter registry study. *Front. Neurol.* **13**, 952794 (2022).
2. Ceballos-Arroyo, A. M. et al. Vessel-aware aneurysm detection using multi-scale deformable 3D attention. *Med. Image Comput. Comput. Assist. Interv.* **15005**, 754–765 (2024).
3. Wang, J. et al. Detection of intracranial aneurysms using multiphase CT angiography with a deep learning model. *Acad. Radiol.* **30** (11), 2477–2486 (2023).
4. Din, M. et al. Detection of cerebral aneurysms using artificial intelligence: a systematic review and meta-analysis. *J. Neurointerv Surg.* **15** (3), 262–271 (2023).
5. Kuwabara, M. et al. Effectiveness of tuning an artificial intelligence algorithm for cerebral aneurysm diagnosis: a study of 10,000 consecutive cases. *Sci. Rep.* **13** (1), 16202 (2023).
6. Sichtermann, T. et al. Deep Learning-Based detection of intracranial aneurysms in 3D TOF-MRA. *AJNR Am. J. Neuroradiol.* **40** (1), 25–32 (2019).
7. Terasaki, Y. et al. Multidimensional deep learning reduces False-Positives in the automated detection of cerebral aneurysms on Time-Of-Flight magnetic resonance angiography: A Multi-Center study. *Front. Neurol.* **12**, 742126 (2021).
8. Banerjee, A. et al. Automated seizure onset zone locator from resting-state functional MRI in drug-resistant epilepsy. *Front. Neuroimaging.* **1**, 1007668 (2023).
9. PMCID: PMC10406253, Kamboj, P., Banerjee, A., Boerwinkle, V. L. & Gupta, S. K. S. The expert's knowledge combined with AI outperforms AI alone in seizure onset zone localization using resting state fMRI. *Front. Neurol.* **14**, 1324461 (2024).
10. Song, T. et al. CPM-Net: A 3D Center-Points Matching Network for Pulmonary Nodule Detection in CT Scans. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, et al., editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2020 [Internet]. Cham: Springer International Publishing; [cited 2025 Jul 1]. pp. 550–9. (2020). Available from: https://doi.org/10.1007/978-3-030-59725-2_53
11. Bo, Z. H. et al. Toward human intervention-free clinical diagnosis of intracranial aneurysm via deep neural network. *Patterns (N Y).* **2** (2), 100197 (2021).
12. Fedorov, A. et al. 3D slicer as an image computing platform for the quantitative imaging network. *Magn. Reson. Imaging.* **30** (9), 1323–1341 (2012).
13. Yadav, S., Kim, J., Young, G. & Qin, L. Dynamic-Computed Tomography Angiography for Cerebral Vessel Templates and Segmentation [Internet]. arXiv; [cited 2025 Jul 1]. (2025). Available from: http://arxiv.org/abs/2502.09893
14. Wasserthal, J. et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiol. Artif. Intell.* **5** (5), e230024 (2023).
15. Faust, L. et al. Considerations for quality control monitoring of machine learning models in clinical practice. *JMIR Med. Inf.* **12**, e50437 (2024).
16. Jeff Rudie, E. et al. Maria Correia de Verdier, Luciano Prevedello, Tyler Richards, Rachit Saluja, Greg Zaharchuk, Jason Sho, Maryam Vazirabad. RSNA 2025 Intracranial Aneurysm Detection. (2025). https://kaggle.competitions/rsna-2025-intracranial-aneurysm-detection, Kaggle. RSNA Intracranial Aneurysm Detection. https://kaggle.com/competitions/rsna-intracranial-aneurysm-detection, 2025. Kaggle.
17. Yadav, S. M., Qin, L., Wan, Q., Kim, J. & Young, G. S. *Simplifying Vessel and Skull Base Segmentation in CT Angiography Imaging* (American Association of Physicists in Medicine (AAPM), 2024).
18. Baumgartner, M., Jäger, P. F., Isensee, F. & Maier-Hein, K. H. NnDetection: A Self-configuring method for medical object detection. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021* Vol. 12905 (eds de Bruijne, M. et al.) (Springer, 2021). https://doi.org/10.1007/978-3-030-87240-3_51.

Acknowledgements

Kim J, Ceballos-Arroyo A, Lin C-H, Jiang H, Yadav S, Qin L, Young GS acknowedge support from NIH R01LM013891. More than 50% of the research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under award number R01LM013891 (total program support $850,000 direct / $1,521,500). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Ceballos-Arroyo, A gratefully acknowledgeds funding support provided by the Fulbright Foundation and Colombia's Ministry of Sciences under the Fulbright Minciencias 2021 program. Ceballos-Arroyo A. and Jiang H. acknowledge the National Artificial Intelligence Research Resource (NAIRR) Pilot (NAIRR240236) and Microsoft Azure for contributing to this research result.

Author contributions

J.K. and A.C., as co-first authors, contributed equally to the final version of the manuscript. J.K. wrote the first draft of the manuscript, reviewed all model output, and conducted all data analysis. A.C. completed substantial additional work for the major revision, obtaining and preparing the additional testing dataset, testing the pipelines on this dataset, analyzing the results, preparing additional figures and revising the manuscript. C.L. and A.C. were in charge of the programming pipeline, coding and running the codes to generate the model output, and post-processing. C.L. and P.L. reviewed parts of the data output. H.J. supervised and advised the overall programming pipeline. S.Y. generated the artery-vein vessel segmentation pipeline. Q.W. annotated the private aneurysm dataset. L.Q. and G.Y., as co-senior authors, supervised and advised on the overall concept of the paper and workflow of the project. Both heavily edited the manuscript and provided the direction of the paper. All authors reviewed the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Ethics approval
This study was approved by our Institutional Review Board. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. A waiver of informed consent was obtained from the Mass General Brigham (MGB) Human Research Committee (HRC) Institutional Review Board (IRB) (MGB HRC IRB Protocol #2022P000792).

Additional information
Correspondence and requests for materials should be addressed to G.S.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.