Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Three-dimensional U-Net Convolutional Neural Network for Detection and Segmentation of Intracranial Metastases

Jeffrey D. Rudie, MD, PhD • David A. Weiss, MSE • John B. Colby, MD, PhD • Andreas M. Rauschecker, MD, PhD • Benjamin Laguna, MD • Steve Braunstein, MD, PhD • Leo P. Sugrue, MD, PhD • Christopher P. Hess, MD, PhD • Javier E. Villanueva-Meyer, MD

From the Department of Radiology and Biomedical Imaging, University of California, San Francisco, 513 Parnassus Ave, San Francisco, CA 94143. Received August 27, 2020; revision requested October 20; revision received February 5, 2021; accepted February 19. **Address correspondence to** J.D.R. (e-mail: *jeff.rudie@gmail.com*).

Authors declared no funding for this work; see Acknowledgments for material support. Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2021; 3(3):e200204 • https://doi.org/10.1148/ryai.2021200204 • Content codes: AI NR

Purpose: To develop and validate a neural network for automated detection and segmentation of intracranial metastases on brain MRI studies obtained for stereotactic radiosurgery treatment planning.

Materials and Methods: In this retrospective study, 413 patients (average age, 61 years ± 12 [standard deviation]; 238 women) with a total of 5202 intracranial metastases (median volume, 0.05 cm^3; interquartile range, 0.02–0.18 cm^3) undergoing stereotactic radiosurgery at one institution were included (January 2017 to February 2020). A total of 563 MRI examinations were performed among the patients, and studies were split into training (n = 413), validation (n = 50), and test (n = 100) datasets. A three-dimensional (3D) U-Net convolutional network was trained and validated on 413 T1 postcontrast or subtraction scans, and several loss functions were evaluated. After model validation, 100 discrete test patients, who underwent imaging after the training and validation patients, were used for final model evaluation. Performance for detection and segmentation of metastases was evaluated using Dice scores, false discovery rates, and false-negative rates, and a comparison with neuroradiologist interrater reliability was performed.

Results: The median Dice score for segmenting enhancing metastases in the test set was 0.75 (interquartile range, 0.63–0.84). There were strong correlations between manually segmented and predicted metastasis volumes (r = 0.98, $P < .001$) and between the number of manually segmented and predicted metastases (R = 0.95, $P < .001$). Higher Dice scores were strongly correlated with larger metastasis volumes on a logarithmically transformed scale (r = 0.71). Sensitivity across the whole test sample was 70.0% overall and 96.4% for metastases larger than 6 mm. There was an average of 0.46 false-positive results per scan, with the positive predictive value being 91.5%. In comparison, the median Dice score between two neuroradiologists was 0.85 (interquartile range, 0.80–0.89), with sensitivity across the test sample being 87.9% overall and 98.4% for metastases larger than 6 mm.

Conclusion: A 3D U-Net–based convolutional neural network was able to segment brain metastases with high accuracy and perform detection at the level of human interrater reliability for metastases larger than 6 mm.

Brain metastases are the most common central nervous system tumor (1). Recent advancements in systemic treatment of primary tumors have led to an increase in the number of patients with metastatic cancer (2). Increased prevalence of brain metastases combined with more recent effective treatment methods, particularly stereotactic radiosurgery, have in turn led to a dramatic increase in imaging to plan treatment and track the longitudinal course of intracranial metastatic disease (3). Detection and segmentation of intracranial metastases on studies obtained for stereotactic radiosurgery treatment planning is a tedious and time-sensitive task requiring high sensitivity and specificity by neuroradiologists and radiation oncologists. Thus, a reliable automated and quantitative longitudinal assessment tool for brain metastases could have a substantial clinical impact by augmenting the ability of radiologists and radiation oncologists to form rapid and accurate treatment plans.

A variety of approaches have been previously used for automated detection and segmentation of intracranial metastases. Several earlier methods used template-matching approaches (4,5), which demonstrated moderate sensitivity and high false-positive (FP) rates in small sample sizes ($n <$ 30). Within the past 5 years, deep learning–based approaches have been established as being superior to the previous generation of atlas or template-matching approaches. Typically, convolutional neural networks (CNNs) (6) have been used for image-based problems, as they allow for the identification of lower- and intermediate-level image features. Prior studies using CNNs for the detection and/or segmentation of brain metastases (7–12) have shown promise but have also been limited by large numbers of FP results and relatively poor performance in detecting smaller metastases.

Here, we evaluated the performance of a customized, three-dimensional (3D) U-Net–based, fully convolutional CNN (13) for the detection and segmentation of brain metastases in a large sample of patients undergoing stereotactic radiosurgery treatment planning. We evaluated detection and segmentation performance by using different input images and loss functions, including focal loss (14,15), which is thought to improve performance for detection of small lesions. We also carefully evaluated the dependence of detection and segmentation performance on metastasis size and established a performance baseline by evaluating neuroradiologist interrater reliability for detection and segmentation of metastases in the test set.

Abbreviations

CNN = convolutional neural network, FD = false discovery, FN = false negative, FP = false positive, 3D = three dimensional, TP = true positive

Summary

A three-dimensional U-Net convolutional network was developed and evaluated for detection and segmentation of intracranial metastases on brain MRI studies obtained for radiosurgery planning; the network was able to segment metastases with high accuracy and detect metastases larger than 6 mm with a reliability similar to neuroradiologist interrater reliability.

Key Points

- A three-dimensional U-Net developed on 463 brain MRI studies of patients with 4494 brain metastases undergoing radiosurgery achieved a median Dice score of 0.75 in a held-out test set of 100 patients (708 metastases) and had an overall sensitivity of 70% and a positive predictive value of 91.5%.
- There were strong correlations between manually segmented and predicted metastasis volumes ($r = 0.98$, $P < .001$) and the number of metastases ($r = 0.95$, $P < .001$).
- Performance was highly dependent on metastasis size, with sensitivity for metastases larger or smaller than 6 mm being 96.4% and 59.1%, respectively.

Keywords

- MR Imaging, Neuro-Oncology, Neural Networks, CNS, Brain/Brain Stem, Segmentation/Feature Detection/Quantification (Vision and Application Domain)

Materials and Methods

Study Design and Patients

As part of an institutional review board–approved Health Insurance Portability and Accountability Act–compliant study, 563 brain MRI studies from 413 patients (mean age, 61 years ± 12 [standard deviation]; 238 women) (Table 1) undergoing stereotactic radiosurgery planning from the University of California, San Francisco, Medical Center were included with a waiver for written consent in this retrospective study.

The training and validation sample (463 MRI studies) represented 313 distinct patients (mean age, 61 years ± 11; 186 women) identified through a search of institutional radiology archives (mPower; Nuance Communications) of stereotactic radiosurgery studies performed between January 1, 2017, and March 30, 2019. The validation sample consisted of 50 randomly selected MRI studies from the total 463 MRI studies, with the remaining 413 MRI studies reflecting the training dataset. The final test set ($n = 100$) represented 100 discrete patients (mean age, 62 years ± 12; 52 women), who were distinct from the patients in the training and validation group and subsequently underwent imaging at the same institution between April 1, 2019, and February 29, 2020.

Exclusion criteria for all datasets included patients without definitive enhancing intracranial metastases ($n = 26$), patients who presented with only dura-based or leptomeningeal metastases ($n = 9$), and examinations that had missing sequences or yielded corrupted data ($n = 8$). Although we did not exclude patients with resection cavities for the training ($n = 124$) and validation ($n = 463$) sample, we excluded patients with resection cavities for the final test set ($n = 30$ from an initial 130 selected) to more accurately evaluate the detection of individual intracranial metastases rather than evaluating the postoperative enhancement typically seen along resection cavities.

Imaging Data Acquisition

T1-weighted and T1-weighted postcontrast images (spoiled gradient echo sequences) were included for each patient. The majority (374 of 563) of studies were acquired with a 1.5-T Signa HDxt scanner (GE Healthcare). There were 150 studies acquired with a 1.5-T Achieva scanner (Philips Healthcare) and 39 acquired with a 3.0-T Discovery MR750 scanner (GE Healthcare). Although acquisition parameters varied slightly, representative values from a 1.5-T Signa HDxt scanner were as follows: repetition time, 8.8 msec; echo time, 3.3 msec; inversion time, 450 msec, flip angle, 11°; matrix size, 256 × 256 × 106; and voxel size, 0.71 × 0.71 × 1.5 mm. Across all acquisitions, the in-plane axial voxel dimension was less than 1 × 1 mm, and the Z dimension was less than or equal to 1.5 mm in 72% of all the scans and 100% of test scans.

Brain Metastases Annotations

To provide voxelwise reference-standard segmentations for training, all MRI studies were hand-segmented using ITK-SNAP (16) (*https://www.itksnap.org/*) by a neuroradiology fellow (J.D.R. and B.L.) or attending neuroradiologist (L.P.S. and J.E.V., with 5 and 4 years of experience as neuroradiology attending physicians) with the final radiology report used as a reference. T1-weighted, T1-weighted postcontrast, and subtraction images were available to guide the manual segmentations. Only the enhancing portions of the metastases were segmented. Areas of central necrosis and areas of T1 intrinsic hyperintensity were not included in the segmentation masks. In addition, extra-axial calvarial and entirely dural metastases were not included. To determine interrater reliability of segmentations in the test set, all test set scans were independently hand-segmented by two radiologists (one neuroradiology fellow [J.D.R.] and one of two neuroradiology attending physicians [J.E.V. and L.P.S.]) without reference to the final radiology report. To generate the final reference-standard segmentations on the test set scans, the two segmentations were combined and then refined with reference to the final radiology report.

Image Preprocessing

T1-weighted images were registered to the T1-weighted postcontrast images by rigid registration (six degrees of freedom) using the FMRIB Software Library's Linear Image Registration Tool (17). Subtraction images were generated by subtracting the registered T1-weighted images from the T1-weighted postcontrast images. The input images were then resampled to 1-mm^3 isotropic resolution by linear interpolation. Prior to input into the network, intensity normalization was performed to achieve a zero mean and unit standard deviation. The images

Table 1: Patient Demographics and Characteristics

Parameter	All Patients	Training and Validation	Testing	*P* Value
Demographics				
No. of patients	413	313	100	NA
No. of MRI studies	563	463	100	NA
Average age (y)	61 ± 12	61 ± 11	62 ± 12	.92
No. of women	238 (57.6)	186 (59.4)	52 (52)	.21
Primary cancer types				
Lung	173 (41.9)	128 (40.9)	45 (45)	.54
Breast	96 (23.2)	73 (23.3)	23 (23)	.94
Melanoma	52 (12.6)	46 (14.7)	6 (6)	.02
Renal	22 (5.3)	16 (5.1)	6 (6)	.73
Head and neck	11 (2.7)	8 (2.6)	3 (3)	.81
Other genitourinary	10 (2.4)	9 (2.9)	1 (1)	.29
Other gastrointestinal	9 (2.2)	8 (2.6)	1 (1)	.35
Rectal	8 (1.9)	5 (1.6)	3 (3)	.38
Neuroendocrine	5 (1.2)	5 (1.6)	0 (0)	NA
Colon	6 (1.5)	4 (1.3)	2 (2)	.60
Prostate	10 (2.4)	4 (1.3)	6 (6)	.01
Thyroid	4 (1.0)	2 (0.6)	2 (2)	.23
Other or unknown	7 (1.7)	5 (1.6)	2 (2)	.79
Metastasis information				
Total no. of metastases	5202	4494	708	NA
No. of metastases				
Average ± SD	9.32 ± 12.9	9.7 ± 13.5	7.1 ± 9.7	.07
Median (IQR)	5 (2–10)	5 (2–11)	3 (2–8)	
Total metastasis volume (cm^3)				
Average ± SD	5.3 ± 6.9	5.6 ± 7.2	3.6 ± 4.9	.01
Median (IQR)	2.7 (0.6–6.8)	3.0 (0.6–7.5)	1.8 (0.5–4.8)	
Individual metastasis volume (cm^3)				
Average ± SD	0.56 ± 2.0	0.57 ± 2.0	0.50 ± 1.7	.36
Median (IQR)	0.05 (0.02–0.18)	0.04 (0.02–0.18)	0.05 (0.02–0.19)	

Note.—Continuous variables are shown as the average ± SD or the median and interquartile range (in parentheses). Categorical variables are shown as the number and percentage (in parentheses). *P* values are for comparisons between the training and validation dataset compared with the test dataset using *t* tests or χ^2 tests. NA = not applicable. IQR = interquartile range, SD = standard deviation.

were then augmented three times using elastic transformations (18), including rotate, flip, and skew, with small random affine transformations stacked on top of small random freeform deformations. The augmented images, either the T1-weighted postcontrast images or the subtraction images, were then split into 96-mm^3 cubes ("3D patches") and used as input for the CNN.

U-Net CNN

We used a previously detailed 3D U-Net CNN (13), which consisted of four consecutive downsampled blocks, followed by four consecutive upsampled blocks with residual (skip) connections (Fig 1). Batch normalization was used for regularization with a batch size of six and the rectified linear unit for nonlinearity. The Adam optimizer (19) and a learning rate of 10^{-4} was used for all experiments. Patches with and without lesions were equivalently sampled during training. The networks were trained until validation loss stabilized, which was at approximately 10 epochs over the course of approximately 24 hours. During testing, the brain volume was densely sampled using a step size of 32 in each direction, and overlapping segmentation predictions were averaged. The network was implemented with TensorFlow (20) (CUDA version 9.2.148) on a Titan Xp graphics processing unit (NVIDIA; 12-GB memory).

Experiments within the Validation Sample

We trained networks using either the T1-weighted postcontrast images or the subtraction images and also tested four different loss functions for these two different inputs. The four different loss functions included cross-entropy, balanced cross-entropy,

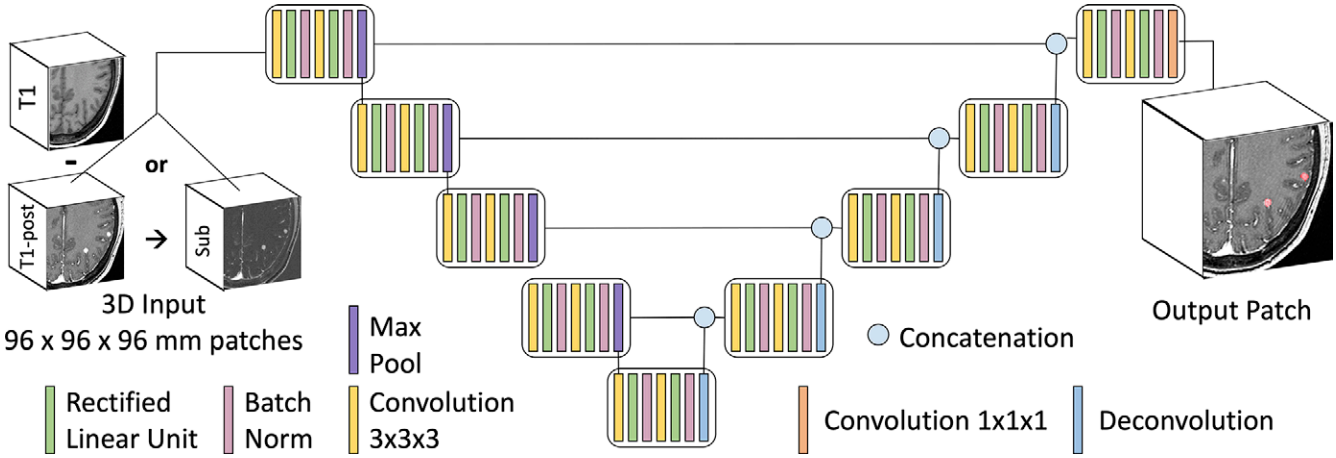

Figure 1: Three-dimensional (3D) convolutional neural network U-Net schematic. The U-Net has four encoding layers and four decoding layers. Input consisted of 3D 96-mm^3 patches of either T1 postcontrast (T1-post) images or subtraction images (T1-post images − T1-weighted images), and the output consisted of predicted segmentation maps of brain metastases.

soft Dice loss (21), and focal loss (13,14) (with alpha = 0.25 and gamma = 2). We also bag ensembled the T1 postcontrast and subtraction models by averaging their predictions for each of the loss functions, across the different loss functions, and across all eight models. Predictions were evaluated across different probability thresholds (final softmax layer) ranging from 0.2 through 0.8 in 0.1 increments, given that the U-Net generates a probabilistic prediction map.

Performance Metrics

We evaluated segmentation performance in the validation and test sets by using voxelwise Dice coefficient (22): $2 \times TP/(2 \times TP + FP + FN)$, where TP is true positive, FP is false positive, and FN is false negative. This calculation was performed for each scan and for individual metastases. We evaluated the performance for detection of metastases in each scan by calculating the metastasis FN rate (FN/TP + FN) and false discovery (FD) rate (FP/TP + FN). Additionally, we evaluated the overall metastasis sensitivity (1 − FN/(TP + FN) and positive predictive value (TP/TP + FP) across all 100 patients in the test sample. A TP metastasis was defined as a connected component of the manual segmentation that had any overlap with the predicted segmentation. Finally, to measure the shape fidelity of the predicted segmentations for detected metastases in the test set, we calculated Hausdorff distances as the 95th percentile of minimum distances from all points in the predicted segmentations to the reference-standard set of voxels.

Statistical Analyses in the Validation Sample

For statistical comparison of different models evaluated in the validation sample, we computed *P* values for the differences between the mean Dice scores, FD rates, and FN rates by using bootstrapping with paired sampling and 10 000 replicates with custom Python scripts. Significance was defined by a *P* value less than .05, and all reported *P* values represent the proportion of tests in which the difference of means was less than 0.

Statistical Analyses in the Test Sample

We applied the highest-performing model (or ensemble) at the peak probabilistic threshold from the validation set to the final test set. To further interrogate the performance of the U-Net on the test set, we performed Pearson correlations between the manually segmented and predicted volume and the number of brain metastases. Given the known relationship between lesion volume and Dice scores (13,23), we evaluated the correlation between the Dice score and the total metastasis volume per case as well as the individual metastasis Dice score and volume on a $\log_{10}$-transformed scale. In addition, we evaluated detection sensitivity across a range of different metastasis sizes.

We evaluated radiologist interrater reliability by computing Dice scores between the two initial manual annotations of the test set and then compared them with Dice scores for the U-Net relative to the final reference-standard segmentations by using bootstrapping. We then evaluated the sensitivity for metastasis detection of the two initial manual annotations relative to the combined final reference-standard annotations and compared detection sensitivity between the U-Net relative to the reference-standard segmentations and different metastasis sizes by using χ^2 tests. We further evaluated sensitivity for detection across different metastasis sizes as well as the relationship between the metastasis volume and Dice score between the two manual segmentations.

Results

Size and Distribution of Brain Metastases

A total of 4494 brain metastases were segmented in the training and validation dataset, and 708 metastases were segmented in the test dataset. The average number of metastases per patient was similar between the two datasets (9.7 ± 13.5 vs 7.1 ± 9.7; *P* = .07). The most common types of primary cancers were lung cancer, breast cancer, and melanoma. Table 1 provides an overview of patient and metastasis characteristics.

The average total volume of metastases was higher in the training and validation datasets than in the test dataset (5.6 cm^3

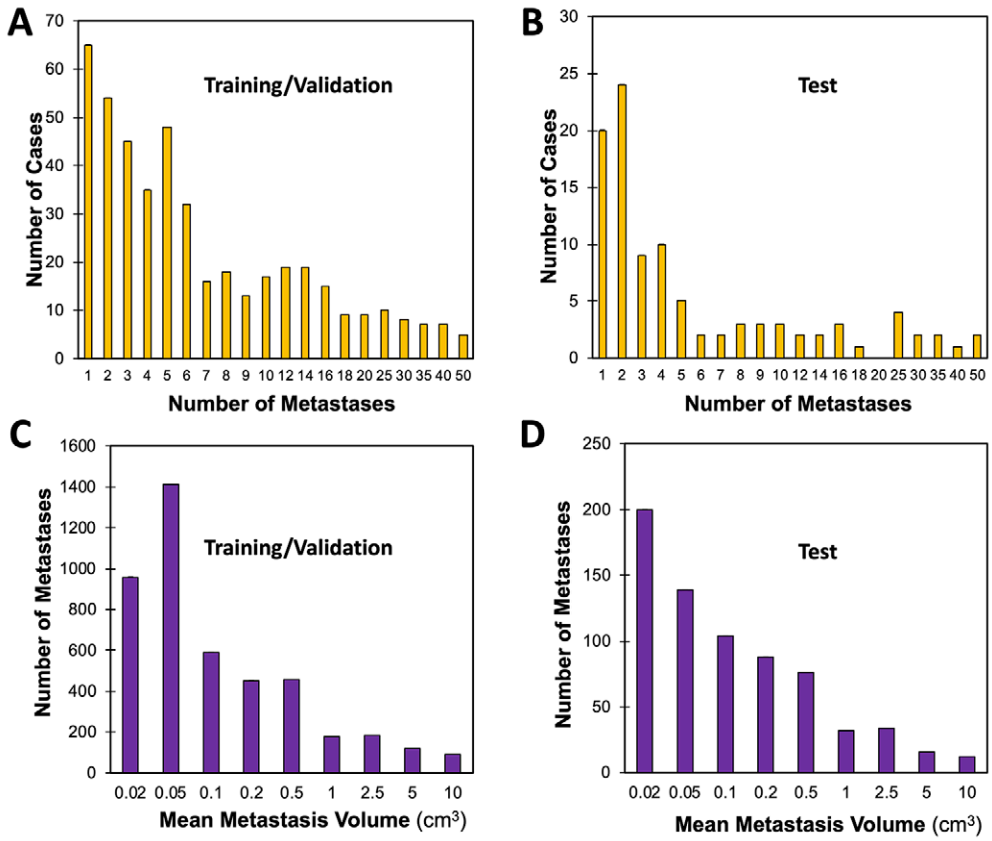

Figure 2: Distribution of number and size of brain metastases in the training and validation sample and test sample. Counts of number of manually segmented brain metastases per case for the *A*, training and validation and *B*, test cases. Average brain metastasis volume for the, *C*, training and validation and, *D*, test cases.

± 7.2 vs 3.6 cm^3 ± 4.9; *P* = .01); however, the average individual metastasis sizes were similar (0.57 cm^3 ± 2.0 vs 0.50 cm^3 ± 1.7; *P* = .36). Metastasis size distributions are shown in Figure 2.

Performance in Validation Dataset

The median Dice scores and average FD and FN rates for T1 postcontrast and subtraction models across the four different loss functions and averaged across the loss functions for the validation sample are shown in Table 2. The median Dice scores across the four different loss functions for the T1 postcontrast and subtraction images ranged from 0.71 to 0.77. The ensemble model predictions across all T1 postcontrast and subtraction models were best at a probabilistic threshold of 0.4 with a median Dice score of 0.78 (interquartile range, 0.65–0.86), which was significantly higher than those of all 14 other models.

The average FD and FN rates ranged from 15% to 121% and 14% to 34%, respectively. The FD rates for the ensemble model (17% ± 40) were significantly lower for six of the other 14 models. The FN rates for the ensemble model (22% ± 25) were significantly lower for 11 of the other 14 models.

Although a lower probabilistic threshold of 0.3 did improve the average FN rate from 22% to 19%, it increased the average FD rate from 17% to 53%. Charts displaying median Dice scores, average FD rates, and average FN rates as a function of the probabilistic threshold for the average of the T1 postcontrast and subtraction models are shown in Figure 3. The ensembled model with a probabilistic threshold of 0.4 was chosen for the test set.

Test Set Performance

Example predicted segmentations of the distinct test set patients using the bagged ensemble of the eight different individual models are shown in Figure 4. The median Dice score in the test set was 0.75 (interquartile range, 0.67–0.83), and the average score was 0.73 ± 0.14. There were three test cases in which Dice scores were 2 standard deviations below the average, all of which were cases in which metastases were missed or undersegmented. The average FD and FN rates across cases were 17.5% ± 41.5 and 17.7% ± 22.5, respectively, corresponding to an average case sensitivity of 82.5% and a positive predictive value of 91.5%. Across the entire test sample, the sensitivity was 70.0% (496 of 708 metastases detected).

Overall sensitivity varied as a function of metastasis size: it was 84.3% (482 of 572) for metastases larger than 0.014 cm^3 (≥3 mm in diameter), 94.5% (308 of 326) for those larger than 0.065 cm^3 (≥5 mm in diameter), and 96.4% (217 of 225) for those larger than 0.113 cm^3 (≥6 mm in diameter). The sensitivity for detection of metastases between 3 and 6 mm was 76.7% (266 of 347). Sensitivities for metastases smaller than 3 mm, 5 mm, or 6 mm were 14.7% (20 of 136), 50.9% (195 of 383), and 59.1% (286 of 484), respectively.

Table 2: Validation Set Performance Metrics of Models with Different Inputs and Loss Functions

Loss Function	Dice	FD Rate (%)	FN Rate (%)
T1 postcontrast			
Cross-entropy	0.75 (0.54–0.85)*	35 ± 67†	28 ± 28‡
Balanced cross-entropy	0.77 (0.60–0.84)*	47 ± 70†	24 ± 25‡
Soft Dice loss	0.73 (0.48–0.85)*	17 ± 32	16 ± 22
Focal loss	0.72 (0.54–0.82)*	39 ± 69†	34 ± 28‡
Ensemble of all loss functions	0.79 (0.58–0.85)*	27 ± 59	27 ± 28‡
Subtraction			
Cross-entropy	0.71 (0.57–0.83)*	35 ± 57†	28 ± 28‡
Balanced cross-entropy	0.73 (0.53–0.83)*	31 ± 59†	31 ± 28‡
Soft Dice loss	0.71 (0.53–0.83)*	15 ± 31	14 ± 19
Focal loss	0.72 (0.56–0.83)*	121 ± 194†	28 ± 30‡
Ensemble of all loss functions	0.77 (0.63–0.86)*	23 ± 58	26 ± 28‡
T1 postcontrast and subtraction model ensemble			
Cross-entropy	0.75 (0.62–0.85)*	13 ± 36	29 ± 30‡
Balanced cross-entropy	0.75 (0.63–0.86)*	16 ± 43	28 ± 28‡
Soft Dice loss	0.77 (0.61–0.86)*	67 ± 107	24 ± 26
Focal loss	0.76 (0.54–0.85)*	21 ± 40	33 ± 30‡
Ensemble of all loss functions	0.78 (0.65–0.86)	17 ± 40	22 ± 25

Note.—Dice scores are the median and interquartile range (25th to 75th percentile in parentheses), and FD and FN rates are the average ± standard deviation. FD = false discovery, FN = false-negative.
* Ensemble of all models resulted in a higher Dice score than all 14 of the 14 other models (P = .01 for T1 postcontrast balanced cross-entropy, P = .003 for T1 postcontrast ensemble and subtraction ensemble, and $P < .001$ for all others).
† Ensemble of all models resulted in a lower FD rate in six of the 14 other models (P = .04 for T1 postcontrast cross-entropy, P = .005 for T1 postcontrast focal loss, P = .01 for subtraction cross-entropy, and $P < .001$ for all others).
‡ Ensemble of all models resulted in a lower FN rate in 11 of the 14 other models (P = .04 for T1 postcontrast balanced cross-entropy, P = .008 for T1 postcontrast and subtraction model ensemble balanced cross-entropy, and $P < .001$ for all others).

The average and median sizes of FN metastases were 0.03 cm^3 and 0.0013 cm^3 (approximately 4.0 mm and 1.4 mm in diameter), respectively. Although most FN metastases were smaller than 3 mm, larger FN metastases included those with subtle and/or minimal enhancement or metastases adjacent to dura and veins and venous sinuses (Fig 5). For the 496 detected metastases in the test set, the average ± standard deviation and median 95th percentile Hausdorff distances were 1.9 mm ± 1.6 and 1.5 mm.

There was a total of 46 FP findings in the test set with a total FD rate of 6.5% (46 of 708) and a positive predictive value of 91.5% (496 of 542). The average and median sizes of FP metastases were 0.05 cm^3 and 0.01 cm^3. FP findings included osseous metastases and benign extra-axial schwannomas, noting that skull stripping was not performed and that calvarial or purely extra-axial metastases or benign enhancing lesions were not included in reference-standard segmentations. FP findings also consisted of prominent vessels, including benign vascular lesions, such as capillary telangiectasias and developmental venous anomalies (Fig 5). Given the presence of calvarial and extra-axial FP findings, we performed skull stripping as a post hoc step, and the overall number of FP findings decreased to 35 (4.9% FD rate and positive predictive value of 93.4%) without affecting sensitivity.

Correlation between Manual and Predicted Volume and Dice Score

There was a strong correlation between manually segmented and predicted metastasis volume in the test set (Pearson r = 0.98, $P < .001$; Fig 6, *A*). There was also a strong correlation between the manually segmented number of metastases and the predicted number of metastases (Pearson r = 0.95, $P < .001$; Fig 6, *B*).

There was a moderately strong positive correlation between manually segmented $\log_{10}$-transformed volumes and Dice scores for total volumes of metastases (Pearson r = 0.63, $P <$ 0.001; Fig 6, *C*) and individual metastases (Pearson r = 0.71, $P < .001$; Fig 6, *D*), such that 51% of the variance (R^2) in individual metastasis Dice scores was explained by $\log_{10}$-transformed metastasis volumes.

Interrater Reliability

The average Dice score between the two manual annotations for the test set was 0.83 ± 0.09 (median, 0.85; interquartile range, 0.80–0.89), which was higher than the average U-Net Dice score (0.75 ± 0.14; $P < .001$ by bootstrapping). The average per-scan sensitivity and total sensitivity of the neuroradiology fellow and attending physician segmentations versus the final reference standard segmentations were similar: 88% ± 52 and 88.4% (626 of 708) and 87% ± 48 and 87.4% (619 of 708), respectively (P = .57, χ^2 = 0.32).

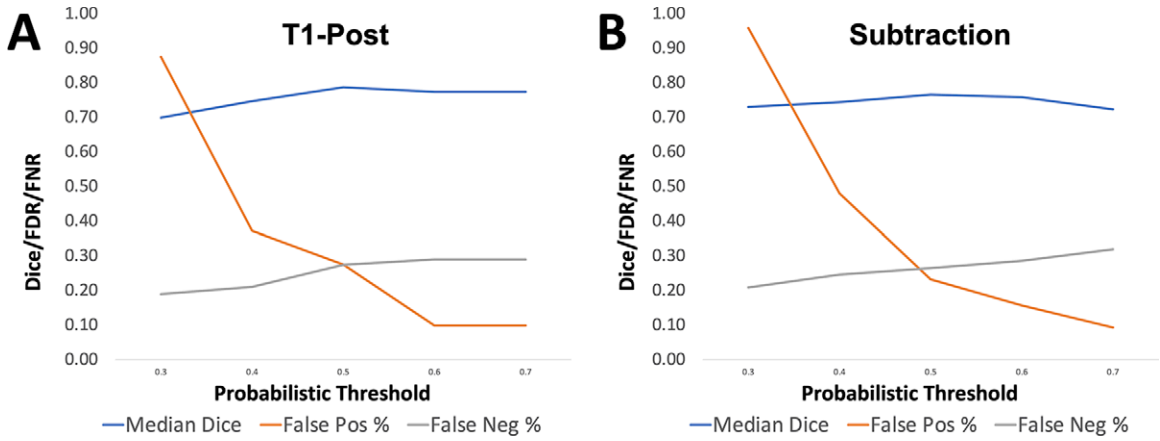

Figure 3: Performance across different probabilistic thresholds. The median Dice score, average false discovery rate (FDR), and average false-negative rate (FNR) are plotted as function of the probabilistic threshold chosen for the models averaged across the, *A*, T1 postcontrast (T1-post) models and, *B*, subtraction models. neg = negative, pos = positive.

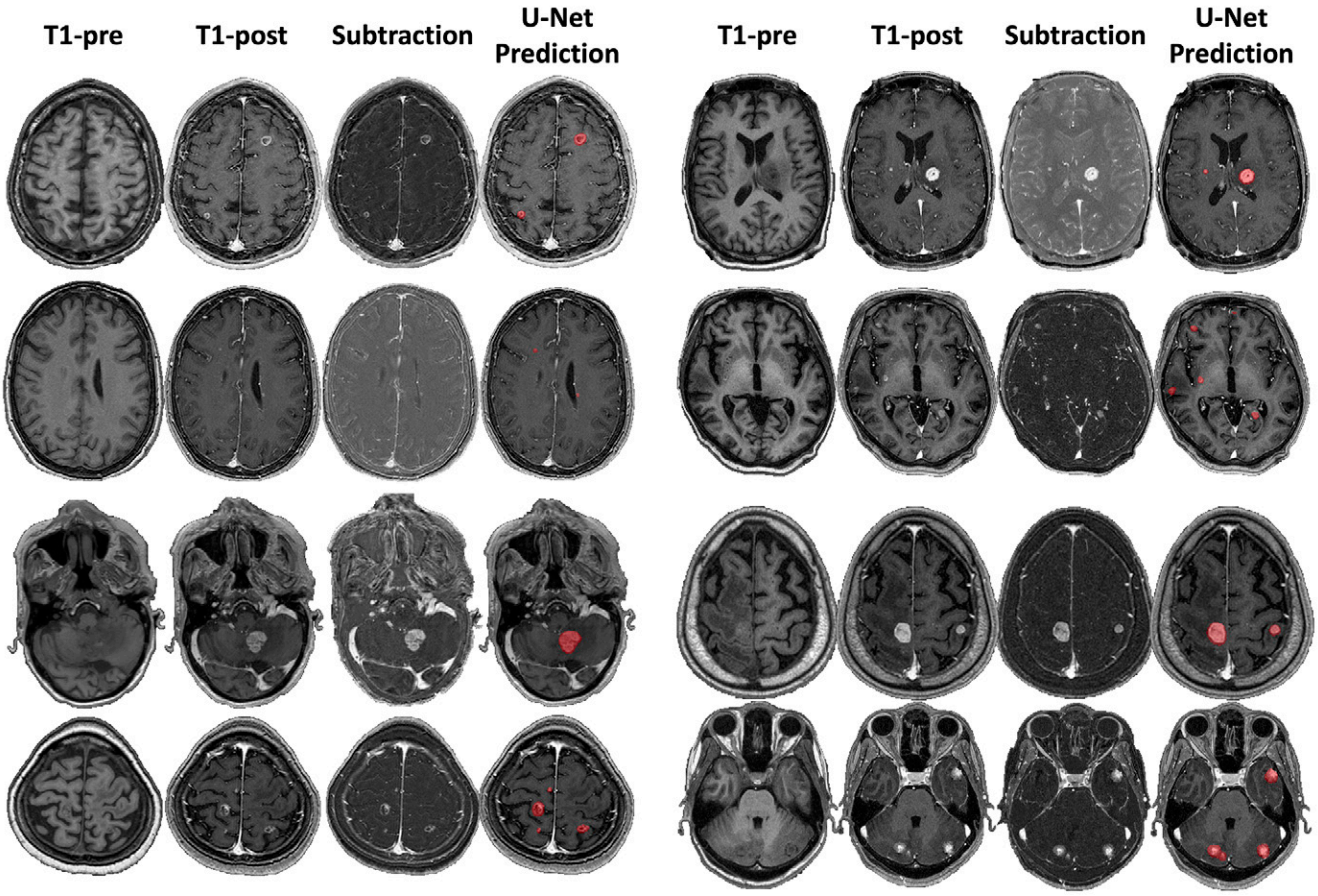

Figure 4: Example U-Net–predicted segmentations. Eight example MRI studies with axial T1 precontrast (T1-pre), T1 postcontrast (T1-post), and subtraction images, with example segmentations overlaid on the T1-post images (U-Net prediction).

Across both the neuroradiology fellow and attending physician segmentations, sensitivities for metastases smaller than 3 mm, metastases between 3 and 6 mm, metastases larger than 6 mm, and all metastases were 63.2% (172 of 272), 90.8% (630 of 694), 98.4% (443 of 450), and 87.9% (1245 of 1416), respectively. Although the sensitivity for detection was higher for human annotators than for the U-Net for metastases smaller than 3 mm (63.2% vs 14.7%, $P < .001$, $\chi^2 = 83.7$) and metastases between 3 and 6 mm (90.8% vs 76.7%, $P < .001$, $\chi^2 = 37.3$), there was no significant difference in the detection of metastases greater than 6 mm (98.4% vs 96.4%, $P = .09$, $\chi^2 = 2.76$).

Similar to the U-Net, the two manual annotations demonstrated a strong positive correlation between the reference standard manually segmented $\log_{10}$-transformed volumes and the Dice scores (Pearson $r = 0.50$, $P < .001$).

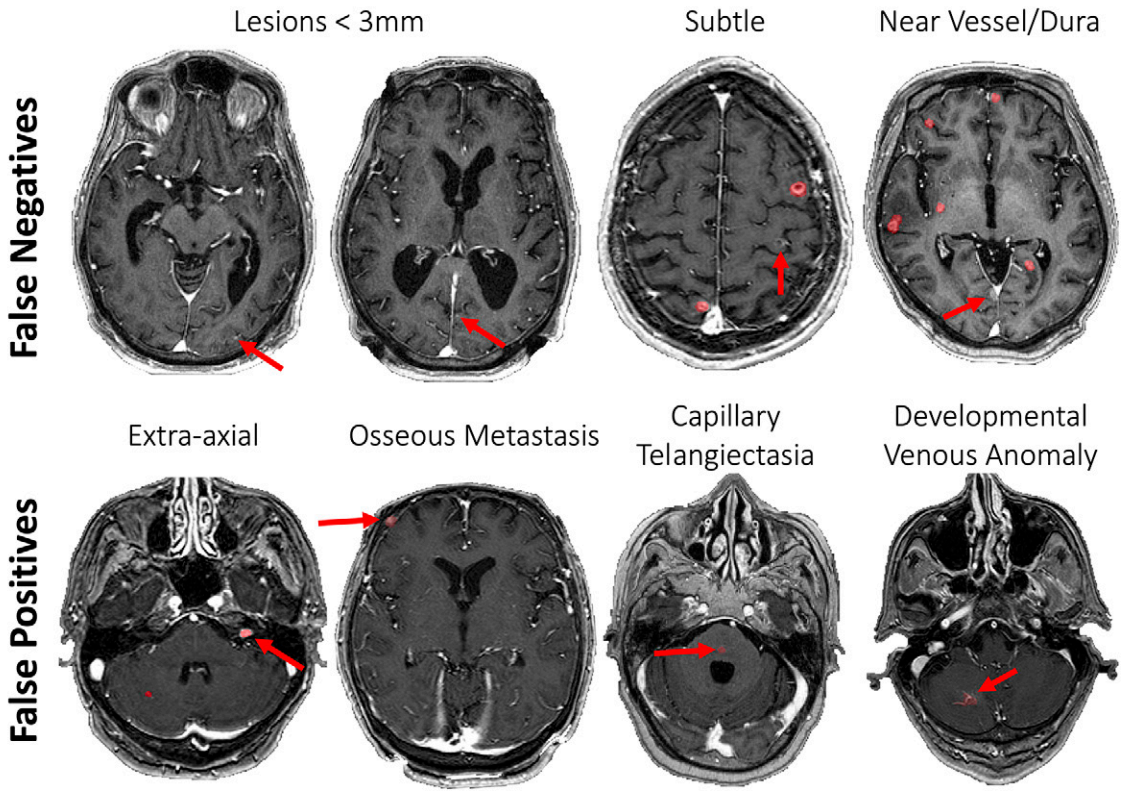

Figure 5: Example false-negative and false-positive findings. False-negative metastases (top row) were typically smaller than 3 mm or had other challenging characteristics, such as a subtle enhancement pattern or close proximity to the dura or prominent vessels. False-positive findings included calvarial metastases, as well as benign lesions, such as schwannomas or vascular lesions, including capillary telangiectasias and developmental venous anomalies.

Manual Segmentation and U-Net Inference Times

Manual segmentation of test cases took approximately 15–20 minutes per scan. The average inference time of the U-Net for test patient studies was 20 seconds for preprocessing (interpolation and patch construction) and 25 seconds for inference with a single model.

Discussion

Artificial intelligence methods are poised to improve diagnostic and treatment options for patients diagnosed with central nervous system neoplasms (24). In particular, the assessment of brain metastases represents an ideal use case for the translation of artificial intelligence methods into clinical practice. Here, we evaluated a 3D U-Net CNN for the detection and segmentation of brain metastases in a large sample of brain MRI studies acquired for stereotactic radiosurgery of brain metastases. The network performed well at detecting and segmenting metastases, with sensitivity equivalent to the interrater reliability of neuroradiologists for metastases larger than 6 mm.

This 3D U-Net architecture was previously shown to perform well across a large variety of lesions on T2 fluid-attenuated inversion recovery MRI (13). There were minimal differences between different loss functions or between using T1 postcontrast images and using subtraction images. Interestingly, the focal loss function, which has been shown to improve robustness against class imbalance, did not improve performance for detection of small metastases, which is consistent with another recent study for the detection of brain metastases using a two-dimensional single shot detector deep learning–based algorithm with bounding boxes (10).

Although directly comparing the performance of this algorithm with algorithm performance in other recent studies is not entirely possible, given the variation in the data used and the metrics reported, this network appeared to achieve performance similar to that of networks in other more recently published studies (9–12). Although our reported Dice scores were similar to those reported in prior studies (most ranging from 0.6 to 0.8), the average FD rate of 17% and the total positive predictive value of 91.5% in our study appear to be better than those from multiple prior studies (7–10) that have reported more than four FP findings per patient, including the study by Zhou et al (10), which reported a positive predictive value of 36%; our results appear to be more similar to the findings of Bousabarah et al (12), who also used a U-Net architecture with an average FP rate of 8%–35%. As expected, lowering the model's probabilistic threshold for detection slightly improved sensitivity, but this was at the cost of many more FP findings.

Consistent with prior studies showing the important impact of lesion size on the Dice score (13,23), we found a strong relationship between Dice scores and metastasis size, such that metastasis size transformed to a $\log_{10}$ scale explained 51% of the variance in the Dice scores. Likewise, sensitivity

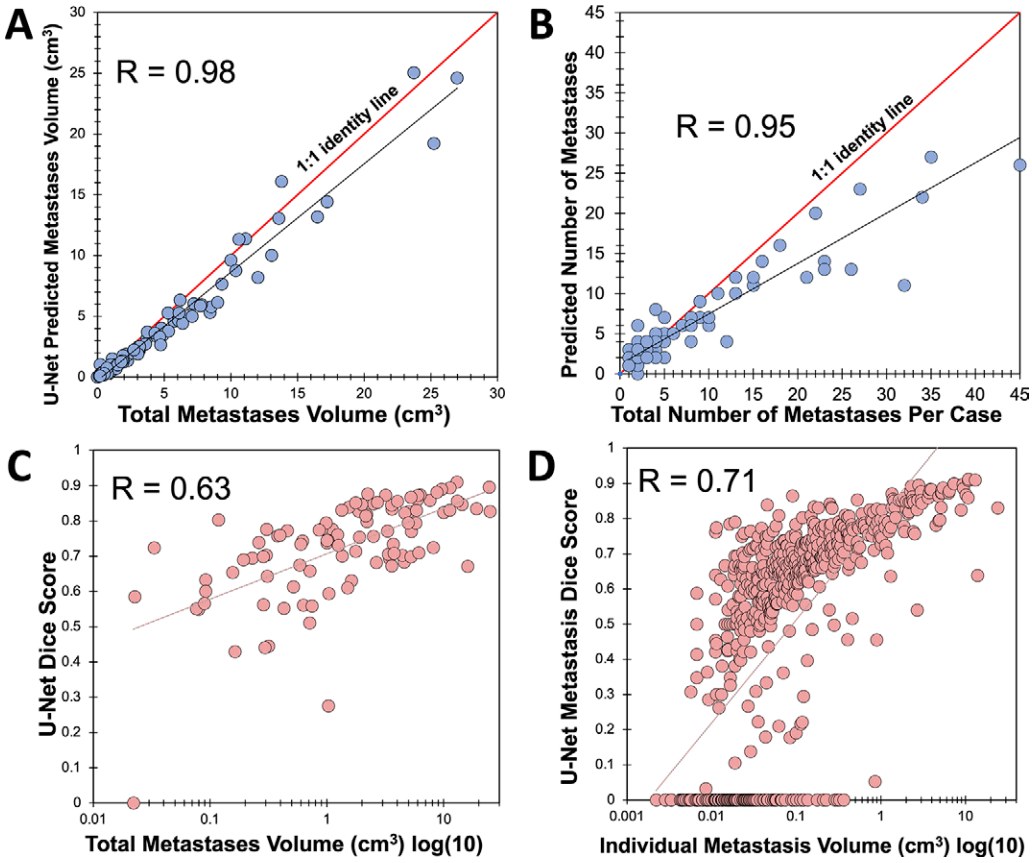

Figure 6: Relationship between manually segmented metastasis volume or number and U-Net–predicted volume or number and Dice scores. *A*, Scatterplot and Pearson correlation between manually segmented tumor volume and U-Net–predicted tumor volume. *B*, Scatterplot and Pearson correlation between manually segmented number of metastases per case and predicted number of metastases per case. Scatterplot and Pearson correlation between $\log_{10}$-transformed volumes and Dice scores for, *C*, individual patients and, *D*, individual brain metastases.

for metastases smaller than 3 mm was only 14.6%, whereas sensitivity for metastases greater than 3 mm was 84.3%. Lower Dice scores and lower sensitivity for smaller lesions were also found when comparing the two sets of manual annotations. The relatively lower overall sensitivity reported in our study (70%) compared with some other recent studies is likely a function of many small metastases in our sample. The average number of metastases and the median metastasis size in our sample were 9.3 and 0.05 cm^3, compared with an average of four metastases and a size of 2.2 cm^3 (44 times larger) in a study by Xue et al (11) and an average of 2.4 metastases and a size of 0.47 cm^3 (9.4 times larger) in a study by Bousabarah et al (12), which respectively reported metastasis sensitivities of 100% and 83%. Nevertheless, we found that performance for metastases smaller than 0.113 cm^3 (approximately 6 mm in diameter) and particularly for metastases smaller than 0.014 cm^3 (approximately 3 mm in diameter) was inferior to the interrater reliability of neuroradiologists. It should also be noted that the interrater reliability is likely even higher in clinical practice, in which multiple readers (a radiologist, trainee radiologist if at a training institution, and radiation oncologist) typically review the images on a picture archiving and communication system and have the clinical history and prior imaging studies available for review.

In addition to poor sensitivity for very small metastases, limitations of the current study include a lack of an external test set to evaluate the generalization of the model across more heterogeneous or multisite data. Overall, these limitations point toward the importance of data-sharing initiatives and public competitions, such as those sponsored by the Radiological Society of North America (25) or the Multimodal Brain Tumor Segmentation Challenge (26). Larger sample sizes with more heterogeneous data should allow for improved algorithm performance and generalizability, particularly for difficult-to-detect small metastases, as well as a clearer comparison of different algorithms.

Ideally, detection of metastases smaller than 6 mm should be improved prior to clinical implementation, and it is unlikely that the current level of performance could substantially improve radiologist metastasis detection. However, even an imperfect system that missed smaller metastases, when integrated into clinical systems, has the potential to improve workflow efficiency for radiologists and radiation oncologists by automatically generating preliminary radiology reports and treatment plans and contours for radiation treatment-planning software for metastases larger than 6 mm. This is supported by the time savings of the U-Net segmentations as compared with manual segmentations, which are required for radiation treatment plans. However, this would need to be assessed in a prospective fashion as part of future

study. Notably, a system that also performs longitudinal assessment of lesions would also be ideal for clinical practice in which most studies are done in comparison with a prior study.

As artificial intelligence tools begin to integrate with clinical workflows for more precise quantitative assessments of disease burdens, it will be necessary to distinguish, quantify, and longitudinally assess a variety of disease processes to assist with more accurate and efficient clinical decision making. The evaluation and treatment of brain metastases represents an excellent use case, and the results of the current study support its potential clinical value.

Acknowledgment: We gratefully acknowledge the support of NVIDIA, who donated the Titan Xp graphics processing units used for this research.

Author contributions: Guarantors of integrity of entire study, J.D.R., J.E.V.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, J.D.R., A.M.R., S.B., C.P.H., J.E.V.; clinical studies, J.D.R., S.B., L.P.S., J.E.V.; experimental studies, J.D.R., D.A.W., J.B.C., A.M.R.; statistical analysis, J.D.R., J.B.C., A.M.R.; and manuscript editing, all authors

Disclosures of Conflicts of Interest: J.D.R. Activities related to the present article: current member of the *Radiology: Artificial Intelligence* trainee editorial board. Activities not related to the present article: institution received a grant from the American Society of Neuroradiology (ASNR Foundation grant in AI). Other relationships: disclosed no relevant relationships. **D.A.W.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: author was a consultant for Galileo CDS (work was done concurrently with position as consultant for Galileo CDS but Galileo GDS was not involved with the work under consideration); author has stock/stock options in Galileo GDS. Other relationships: disclosed no relevant relationships. **J.B.C.** disclosed no relevant relationships. **A.M.R.** Activities related to the present article: former member of the *Radiology: Artificial Intelligence* trainee editorial board. Activities not related to the present article: institution received an RSNA scholar grant; institution received an ASNR trainee grant; institution involved with NIH T-32 training program. Other relationships: disclosed no relevant relationships. **B.L.** disclosed no relevant relationships. **S.B.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **L.P.S.** disclosed no relevant relationships. **C.P.H.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: author is a research consultant for GE Healthcare; DSMB member of Focused Ultrasound Foundation; author received travel accommodations from Siemens Healthineers. Other relationships: disclosed no relevant relationships. **J.E.V.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: institution received a grant from GE Healthcare. Other relationships: disclosed no relevant relationships.

References

1. Ostrom QT, Wright CH, Barnholtz-Sloan JS. Brain metastases: epidemiology. Handb Clin Neurol 2018;149:27–42.
2. Arvold ND, Lee EQ, Mehta MP, et al. Updates in the management of brain metastases. Neuro Oncol 2016;18(8):1043–1065.
3. Mills SJ, Radon MR, Baird RD, et al. Utilization of volumetric magnetic resonance imaging for baseline and surveillance imaging in Neuro-oncology. Br J Radiol 2019;92(1098):20190059.
4. Ambrosini RD, Wang P, O'Dell WG. Computer-aided detection of metastatic brain tumors using automated three-dimensional template matching. J Magn Reson Imaging 2010;31(1):85–93.
5. Pérez-Ramírez Ú, Arana E, Moratal D. Brain metastases detection on MR by means of three-dimensional tumor-appearance template matching. J Magn Reson Imaging 2016;44(3):642–652.
6. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–444.
7. Liu Y, Stojadinovic S, Hrycushko B, et al. A deep convolutional neural network-based automatic delineation strategy for multiple brain metastases stereotactic radiosurgery. PLoS One 2017;12(10):e0185844.
8. Charron O, Lallement A, Jarnet D, Noblet V, Clavier JB, Meyer P. Automatic detection and segmentation of brain metastases on multimodal MR images with a deep convolutional neural network. Comput Biol Med 2018;95:43–54.
9. Grøvik E, Yi D, Iv M, Tong E, Rubin D, Zaharchuk G. Deep learning enables automatic detection and segmentation of brain metastases on multisequence MRI. J Magn Reson Imaging 2020;51(1):175–182.
10. Zhou Z, Sanders JW, Johnson JM, et al. Computer-aided detection of brain metastases in T1-weighted MRI for stereotactic radiosurgery using deep learning single-shot detectors. Radiology 2020;295(2):407–415.
11. Xue J, Wang B, Ming Y, et al. Deep learning-based detection and segmentation-assisted management of brain metastases. Neuro Oncol 2020;22(4):505–514.
12. Bousabarah K, Ruge M, Brand JS, et al. Deep convolutional neural networks for automated segmentation of brain metastases trained on clinical data. Radiat Oncol 2020;15(1):87.
13. Duong MT, Rudie JD, Wang J, et al. Convolutional neural network for automated FLAIR lesion segmentation on clinical brain MR imaging. AJNR Am J Neuroradiol 2019;40(8):1282–1290.
14. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 2020;42(2):318–327.
15. Abraham A, Khan K. A novel focal Tversky loss function with improved attention U-Net for lesion segmentation. In: Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2019; 683–687.
16. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 2006;31(3):1116–1128.
17. Smith SM. Fast robust automated brain extraction. Hum Brain Mapp 2002;17(3):143–155.
18. Simard PY, Steinkraus D, Platt JC. Best practices for convolutional neural networks applied to visual document analysis. In: Proceedings of the Seventh International Conference on Document Analysis and Recognition, 2003. Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2003; 958–963.
19. Kingma DP, Ba L. ADAM: a method for stochastic optimization. Presented at the Third International Conference on Learning Representations, San Diego, Calif, May 7–9, 2015.
20. Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous systems. Computer science: distributed, parallel, and cluster computing. ArXiv 1603.04467 [preprint] https://arxiv.org/abs/1603.04467. Posted March 14, 2016. Accessed December 2019.
21. Milletari F, Navab N, Ahmado S. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the Fourth International Conference on 3D Vision (3DV). Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2016; 565–571.
22. Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26(3):297–302.
23. Rudie JD, Weiss DA, Saluja R, et al. Multi-disease segmentation of gliomas and white matter hyperintensities in the BraTS data using a 3D convolutional neural network. Front Comput Neurosci 2019;13:84.
24. Rudie JD, Rauschecker AM, Bryan RN, Davatzikos C, Mohan S. Emerging applications of artificial intelligence in neuro-oncology. Radiology 2019;290(3):607–618.
25. Flanders AE, Prevedello LM, Shih G, et al. Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. Radiol Artif Intell 2020;2:e209002 [Published correction appears in Radiol Artif Intell 2020;2(4).].
26. Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. ArXiv 1811.02629v2 [preprint] https://arxiv.org/abs/1811.02629v2. Posted November 5, 2019. Accessed December 2019.