Title
Interinstitutional Portability of a Deep Learning Brain MRI Lesion Segmentation Algorithm

Journal
Radiology Artificial Intelligence, 4(1)

ISSN
2638-6100

Authors
Rauschecker, Andreas M
Gleason, Tyler J
Nedelec, Pierre
et al.

Publication Date
2022

DOI
10.1148/ryai.2021200152

Peer reviewed

Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Interinstitutional Portability of a Deep Learning Brain MRI Lesion Segmentation Algorithm

Andreas M. Rauschecker, MD, PhD • Tyler J. Gleason, MD • Pierre Nedelec, MS, MTM • Michael Tran Duong, BA • David A. Weiss, MSE • Evan Calabrese, MD, PhD • John B. Colby, MD, PhD • Leo P. Sugrue, MD, PhD • Jeffrey D. Rudie, MD, PhD • Christopher P. Hess, MD, PhD

From the Department of Radiology & Biomedical Imaging, University of California, San Francisco, 513 Parnassus Ave, Room S-261, Box 0628, San Francisco, CA 94143-0628 (A.M.R., T.J.G., P.N., D.A.W., E.C., J.B.C., L.P.S., J.D.R., C.P.H.); and Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia, Pa (M.T.D., D.W.). Received June 23, 2020; revision requested September 1; revision received September 28, 2021; accepted October 22. **Address correspondence to** A.M.R. (e-mail: *andreas.rauschecker@ucsf.edu*).

A.M.R. supported by an American Society of Neuroradiology trainee research award, a National Institutes of Health T-32 institutional training grant (T32EB001631-14), and a Carestream Health/RSNA Research Scholar Grant. The project described was supported by RSNA Research & Education (R&E) Foundation, through grant number RSCH2025. The content is solely the responsibility of the authors and does not necessarily represent the official views of the RSNA R&E Foundation.

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2022; 4(1):e200152 • https://doi.org/10.1148/ryai.2021200152 • Content codes: AI MR NR

Purpose: To assess how well a brain MRI lesion segmentation algorithm trained at one institution performed at another institution, and to assess the effect of multi-institutional training datasets for mitigating performance loss.

Materials and Methods: In this retrospective study, a three-dimensional U-Net for brain MRI abnormality segmentation was trained on data from 293 patients from one institution (IN1) (median age, 54 years; 165 women; patients treated between 2008 and 2018) and tested on data from 51 patients from a second institution (IN2) (median age, 46 years; 27 women; patients treated between 2003 and 2019). The model was then trained on additional data from various sources: *(a)* 285 multi-institution brain tumor segmentations, *(b)* 198 IN2 brain tumor segmentations, and *(c)* 34 IN2 lesion segmentations from various brain pathologic conditions. All trained models were tested on IN1 and external IN2 test datasets, assessing segmentation performance using Dice coefficients.

Results: The U-Net accurately segmented brain MRI lesions across various pathologic conditions. Performance was lower when tested at an external institution (median Dice score, 0.70 [IN2] vs 0.76 [IN1]). Addition of 483 training cases of a single pathologic condition, including from IN2, did not raise performance (median Dice score, 0.72; $P = .10$). Addition of IN2 training data with heterogeneous pathologic features, representing only 10% (34 of 329) of total training data, increased performance to baseline (Dice score, 0.77; $P < .001$). This final model produced total lesion volumes with a high correlation to the reference standard (Spearman $r = 0.98$).

Conclusion: For brain MRI lesion segmentation, adding a modest amount of relevant training data from an external institution to a previously trained model supported successful application of the model to this external institution.

Supplemental material is available for this article.

One of the promises of artificial intelligence (AI) for neuroimaging is its potential to automate the detection of abnormal findings at brain MRI, thereby reducing measurement variability and perceptual errors. However, generalization and portability of AI technologies across different health care practice settings, such as across hospital systems, patient populations, imaging technology manufacturers, and diseases, is critical to the success of such technologies (1,2).

Well-tested AI algorithms may perform poorly on external test datasets because of a mismatch between the distribution of training data and external test data statistics. For example, performance on the external test dataset can suffer because of differences in prevalence of target classification outcomes (eg, disease states) between institutions (3). Data may be acquired at a different resolution, using different MRI parameters (eg, different echo or repetition times), or with different imaging hardware with unique artifacts (4). In addition, patient factors, such as differences in ethnic background, age, body habitus, genetics, or comorbidities, could affect the average imaging appearance of the same diseases across different populations at different institutions (5).

We recently reported an algorithm that detects and segments fluid-attenuated inversion recovery (FLAIR) abnormalities on MR images, regardless of underlying pathologic condition (6). This algorithm uses a three-dimensional (3D) U-Net architecture to detect abnormal FLAIR signal, which is a very sensitive marker of lesions in a variety of pathologic conditions, although it is not specific for any particular cause. The goal of developing this algorithm was to use it to detect and segment FLAIR signal abnormalities across multiple different clinical contexts, including when the disease process is unknown, rather than being limited to any specific pathologic condition. The algorithm was found to perform well across the 19 diseases tested. The purpose of this algorithm is for lesion segmentation and volume quantification, information that may be directly clinically useful, for example, in tracking lesion sizes over time. FLAIR lesion segmentations have also been shown to be useful for diagnostic support if combined with other imaging processing techniques that offer a detailed description of the lesions (7,8).

The performance of the FLAIR U-Net (6), while tested across many pathologic conditions, was tested at only a

Abbreviations

AI = artificial intelligence, BraTS = Multimodal Brain Tumor Segmentation challenge, CNN = convolutional neural network, FLAIR = fluid-attenuated inversion recovery, IN1 = institution 1, IN2 = institution 2, IQR = interquartile range, M_1 = model trained on IN1 patient data, M_{1+B} = model trained on IN1 patient data + BraTS, M_{1+B+2T} = model trained on IN1 patient data + BraTS + data from IN2 patients with tumors, M_{1+2} = model trained on IN1 + IN2 data, $M_{FT\ (1+2)}$ = model trained on IN1 with sequential fine-tuning with IN2, M_2 = model trained on IN2 data, 3D = three dimensional, 2D = two dimensional

Summary

For a brain lesion segmentation model trained on a single institution's data, performance was lower when applied at a second institution; however, the addition of a small amount (10%) of training data from the second institution allowed the model to achieve its full potential performance level at the second institution.

Key Points

- A U-Net for detecting abnormal fluid-attenuated inversion recovery (FLAIR) signal trained at one institution (IN1) had consistently lower performance when applied to a second institution (IN2) (median Dice score, 0.76 and 0.70, respectively; $P = .001$).
- Addition of data with abnormal FLAIR signal from a single pathologic condition (primary brain tumors) from IN2 did not improve performance of the model at IN2 (Dice score, 0.72; $P = .10$).
- Addition of IN2 data that included a variety of pathologic conditions increased model performance on the IN2 test dataset (Dice, 0.77; $P < .001$) despite representing only a small portion (10%) of the total training data.

Keywords

Neural Networks, Brain/Brain Stem, Segmentation

single institution, which inherently calls into question its utility beyond the institution at which it was trained. Although the algorithm was broadly trained on images created using a wide variety of imaging parameters from multiple different MRI machine vendors and models, these factors do not guarantee high performance at another institution. Were external test performances of AI systems to be significantly lower than the reported performances of AI algorithms using internal test datasets, there would be substantial patient safety implications when deploying AI algorithms such as this one across hospital systems. Therefore, we sought to test the external validity of this particular neural network and evaluate strategies for improving interinstitutional portability. The goal of this work is to offer one specific practical strategy for portability of a deep neural network from one institution to another, which is to fine-tune the network using a manageable amount of training data from the target institution.

Materials and Methods

Patients, Data, and Models

This Health Insurance Portability and Accountability Act–compliant retrospective study was approved by the institutional review board of the University of California San Francisco, with a waiver for consent. A total of 51 patients (median age, 46 years [range, 3–83 years]; 27 women) treated at institution 2 (IN2) from 2003 through 2019 served as the primary study sample, such that for each of the 17 diseases listed in the "Diseases" section, three patients with that disease were included. The patients were identified by searching the radiology archives of our tertiary care university hospital for the diagnoses included in the study (see below). The convolutional neural network (CNN) was previously trained on MR images from 293 patients (median age, 54 years [range, 15–95 years]; 165 women) treated at institution 1 (IN1) from 2008 through 2018, as described previously (6). One experiment in the study used training data from an additional 198 patients from IN2 with primary brain tumors (World Health Organization grade II–IV gliomas). Another experiment in the study used 285 patients' open-source data from the 2018 Multimodal Brain Tumor Segmentation challenge (BraTS) (9), which also included manual segmentations of white matter hyperintensities, as described in Rudie et al (10).

Diseases

A total of 17 diseases were included in the test sample, with three patients for each disease. Given that our study was focused on overall performance of the algorithm across a wide range of diseases rather than on performance for any particular disease, a sample size of 51 patients was sufficient for detecting small differences in performance between models. The 17 diseases included in the test sample included: low-grade glioma, high-grade glioma (glioblastoma), primary central nervous system lymphoma, metastatic disease, acute or subacute ischemia, small-vessel ischemic disease, multiple sclerosis, tumefactive multiple sclerosis, neuromyelitis optica, acute disseminated encephalomyelitis, adrenoleukodystrophy, cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy, HIV encephalopathy, progressive multifocal leukoencephalopathy, toxic leukoencephalopathy, posterior reversible encephalopathy syndrome, and migraine. These diseases encompass a large range of pathologic conditions causing FLAIR abnormalities on brain MR images, with variable size, shape, and extent of those abnormalities in individual patients. The diseases were chosen to match the diseases on which the original model (6) was trained, to enable a fair comparison of performance across institutions. Reference standard diagnoses were established by chart review, using pathologic data if available (eg, for brain tumors) or otherwise using a combination of clinical and radiologic follow-up to ensure accurate diagnoses. Exclusion criteria for the study were identical to those of the prior study (6) and included lack of reference standard diagnosis, inadequate imaging (no FLAIR sequence or excessive imaging artifact making diagnostic interpretation impossible), multiple diagnoses or prior surgery causing FLAIR abnormality, and presence of all imaging findings outside of the cerebral hemispheres.

Training and Testing Assignments

For all initial experiments, the 51 study patients from IN2 formed the test set for the FLAIR U-Net. The training set for

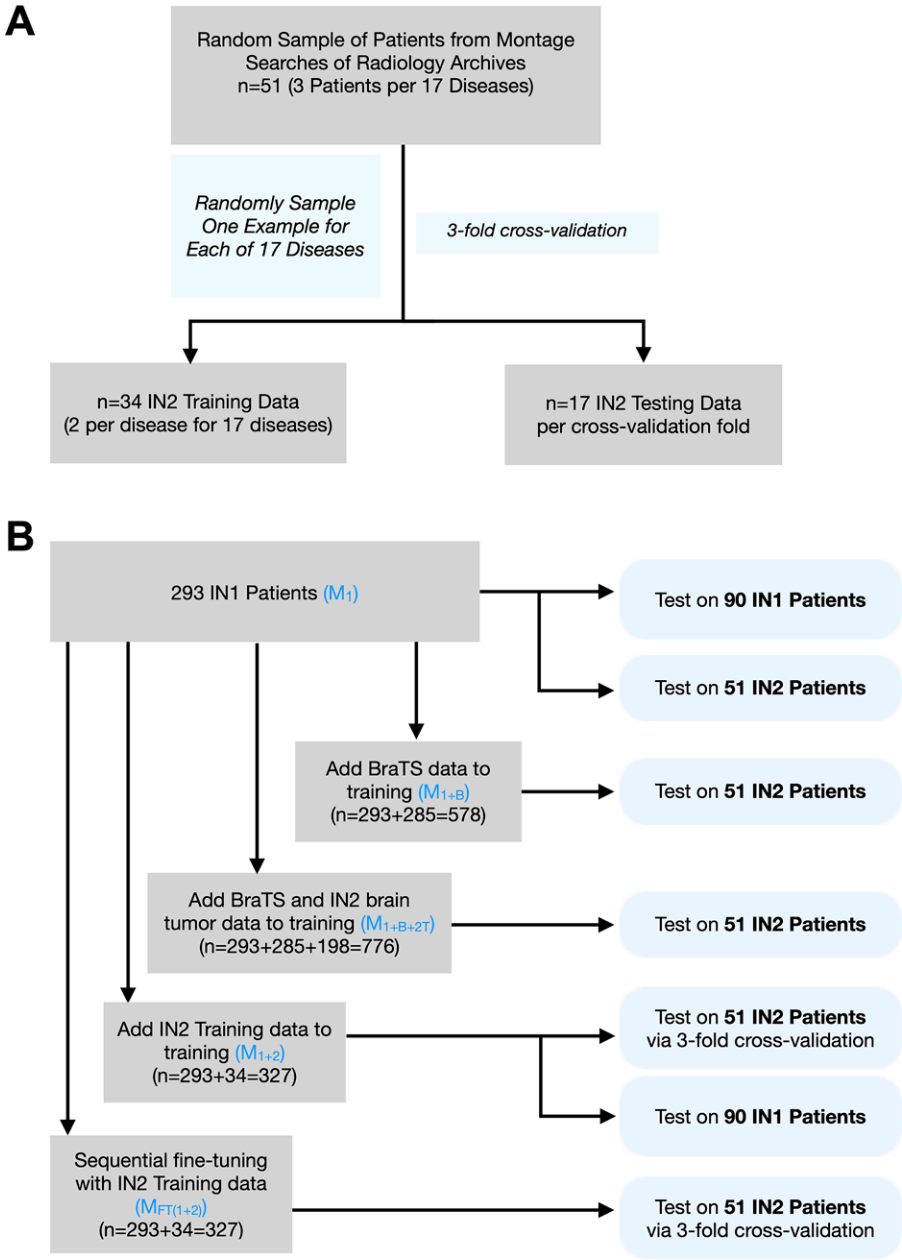

Figure 1: Flow diagram demonstrates use of data from various institutions for training and testing. **(A)** Data from the 51 patients from institution 2 (IN2) formed the primary test set on which the previously trained model was tested. This dataset was then divided into three subsets of training data (n = 34) and independent test data (n = 17), forming a threefold cross-validation test sample. **(B)** The original model (M_1) was trained on 293 patients' data from one institution (IN1) and tested on both the same institution's independent test data (replication of Duong et al [6]) and on IN2 data. Additional experiments were performed using training data composed of other data sources combined with the original dataset (BraTS [M_{1+B}], IN2 brain tumor data [M_{1+B+2T}], and IN2 training data [M_{1+2}]). Fine-tuning with IN2 data ($M_{FT\,(1+2)}$) was also explored in another set of experiments. BraTS = Multimodal Brain Tumor Segmentation challenge, M_1 = model trained on IN1 patient data, M_{1+B} = model trained on IN1 patient data + BraTS, M_{1+B+2T} = model trained on IN1 patient data + BraTS + data from IN2 patients with tumors, M_{1+2} = model trained on IN1 + IN2 data, model $M_{FT\,(1+2)}$ = trained on IN1 with sequential fine-tuning with IN2, M_2 = model trained on IN2 patient data.

the U-Net varied between these experiments and consisted of: *(a)* IN1 data only (M_1); *(b)* IN1 data and 285 BraTS brain tumor segmentations (M_{1+B}); *(c)* IN1 data, 285 BraTS brain tumor segmentations, and data from 198 IN2 patients with brain tumors that were previously manually segmented on FLAIR imaging for a different study (M_{1+B+2T}); *(d)* IN1 data and data from 34 IN2 patients (two per disease for 17 diseases [M_{1+2}]); and *(e)* the same data as M_{1+2}, except that sequential fine-tuning was performed with IN2 ($M_{FT\,(1+2)}$). A stratified threefold cross-validation procedure was used wherein the folds were stratified by each of the 17 diagnoses, such that each fold consisted of two samples of training data from each disease (34 patients' total training data in each fold) and one sample of test data from each disease (17 patients' total test data in each fold, unique for each of the three folds) (Fig 1).

We also performed experiments to tease out the effects of IN1 and IN2 training data on IN1 and IN2 test data performance. In one experiment, we tested the effect of varying the amount of IN1 training data in M_1 with either all 293 patients or with 204, 102, or 51 patients. These models were then tested on IN1 and IN2 test data. In another experiment, we varied the training data on M_{1+2} in four ways: *(a)* 10.4% IN2 (293 IN1 + 34 IN2 patients), *(b)* 14.3% IN2 (204 IN1 + 34 IN2 patients), *(c)* 25% IN2 (102 IN1 + 34 IN2 patients), and *(d)* 40% IN2 (51 IN1 + 34 IN2 patients) to assess the effect of the proportion of IN2 training data on the model performance on IN2 test data. These four models were compared with M_1 and M_2 on the same test datasets.

MRI Parameters and Manual Reference Segmentations

Only FLAIR imaging was used for this study, which was the only sequence used for manual and automated segmentations. Images were acquired with a variety of scanners from GE, Philips, Siemens, and Toshiba; scanner models are outlined in the Table. The MRI parameters that defined FLAIR imaging in the IN2 dataset varied substantially from those that defined the FLAIR imaging in the IN1 training data (Table). Ground truth lesion segmentations were provided by a senior radiology resident (T.J.G.) and were verified and/or modified by an attending neuroradiologist with 3 years of postresidency experience (A.M.R.) using ITK-SNAP (version 3.8; *www.itksnap.org*) (11). Diagnoses were not available to the radiologist at the time of manual segmentation.

Image Preprocessing

Brain extraction was performed directly on FLAIR images using our in-house CNN dedicated to skull stripping. Brain extraction performance was qualitatively assessed (A.M.R.) and found to be excellent. Because additional manual corrections of resulting brain masks did not significantly improve overall Dice scores, only the results using automated brain extraction are presented here. Similarly, N4 bias field correction (12) did not consistently improve performance; thus, all results are reported without this preprocessing step. Images were normalized by the mean and standard deviation signal intensity to zero mean and unit standard deviations. All image volumes were resampled to 1-mm^3 isotropic image resolution via linear interpolation. For training, elastic transformations were applied to the image volumes, including small random rotations, translations, scaling, and free-form deformations such that each image was augmented three times. Similar to Duong et al (6), we split the full-resolution augmented imaging volumes into 96 × 96 × 96-mm cubes ("3D patches") to fit within graphic memory processor constraints. To prevent sample imbalance during training, we sampled the same number of patches that included lesion voxels as that excluded lesion voxels. During testing, the brain volume was densely sampled with the same size cubes, using a step size of 32 mm in each direction. The overlapped segmentation predictions were averaged.

Heterogeneous Scanning Parameters Used for FLAIR Sequences from Patients from IN1 and IN2

Variable	IN1	IN2
Total no. of patients	293	51
Imaging parameter		
TE (msec)	136 (86–396)	127 (94–149)
TR (msec)	9000 (5000–12 000)	6000 (5000–11 000)
Field strength		
1.5 T	228 (77.8)	24 (47.1)
3 T	65 (22.2)	27 (52.9)
Dimension		
2D	285 (97.3)	11 (21.6)
3D	8 (2.7)	40 (78.4)
Manufacturer and model		
GE		
Discovery MR750	4 (1.4)	23 (45.1)
Genesis Signa	20 (6.8)	1 (2.0)
Optima MR450	15 (5.1)	0 (0.0)
Signa Excite	20 (6.8)	1 (2.0)
Signa HDx	0 (0.0)	7 (13.7)
Signa HDxt	14 (4.8)	13 (25.5)
Philips		
Achieva	0 (0.0)	2 (3.9)
Intera	2 (0.7)	1 (2.0)
Siemens		
Aera	14 (4.8)	0 (0.0)
Avanto	38 (13.0)	1 (2.0)
Espree	83 (28.3)	1 (2.0)
Magnetom Essenza	9 (3.1)	0 (0.0)
Skyra	8 (2.7)	0 (0.0)
Symphony	4 (1.4)	0 (0.0)
Symphony Tim	5 (1.7)	1 (2.0)
Trio Tim	37 (12.6)	0 (0.0)
Verio	16 (5.5)	0 (0.0)
Toshiba		
Titan	4 (1.4)	0 (0.0)

Note.—Continuous variables are shown as median, with range in parentheses, and categorical variables are shown as numbers, with percentages in parentheses. Percentages do not equal 100 due to rounding. FLAIR = fluid-attenuated inversion recovery, IN1 = institution 1, IN2 = institution 2, TE = echo time, TR = repetition time, 3D = three dimensional, 2D = two dimensional.

CNN Model Architecture

The 3D U-Net architecture used here was described previously (6) and was replicated without modification. Hyperparameters used for this U-Net included a kernel size of 3 × 3 × 3, cross-entropy loss function, and an Adam optimizer with learning rate of 4 × 10^{-4}, implemented using TensorFlow 2 *(https://www.tensorflow.org)* (13) within the Python programming language. The network was trained for 30 epochs in each experiment described, with a batch size of 24 3D patches (96 × 96 × 96 mm each). The implementation was on a DGX-2 AI server (version 4.5.0; NVIDIA).

Statistical Analyses

The performance of the U-Net was tested against the manual segmentation reference standard using Dice coefficients (14) on the test data described in Figure 1. Threefold cross-validation was used to keep the test data independent from the training data in any experiment in which a portion of IN2 data was used for additional training. The cross-validation approach allowed for calculation of a Dice score for each individual patient, with each patient falling within the test set once. Baseline performance was defined as the Dice scores resulting from the original one-institution model as applied to the same institution, a replication of Duong et al (6). Results of all other trained models were compared with this baseline performance. Mann-Whitney *U* and Kruskal-Wallis H tests were used to compare median Dice scores for unpaired data (eg, IN1 vs

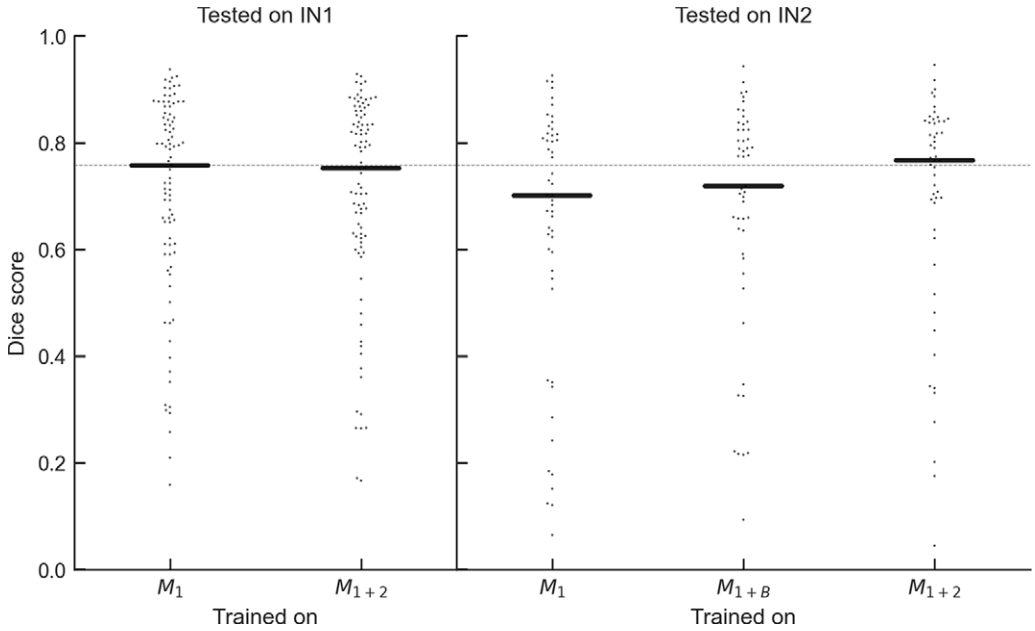

Figure 2: Segmentation accuracy (Dice scores) for five different versions of the convolutional neural network, with individual Dice scores for each patient indicated by a data point and median Dice scores indicated by horizontal black bars. The architecture and hyperparameters of the model remained identical, but training and testing cases varied. The horizontal dashed lined demonstrates baseline performance of the model trained at one institution and applied to the same institution, with a median Dice score (0.76) for comparison to the four other models, which used various mixtures of interinstitutional training data. IN1 = institution 1, IN2 = institution 2, M_1 = model trained on IN1 patient data, M_2 = model trained on IN2 data, M_{1+B} = model trained on IN1 patient data + Multimodal Brain Tumor Segmentation challenge data, M_{1+2} = trained on IN1 + IN2 data.

IN2 test data). For any paired comparisons (eg, two different models tested on IN2 test data, with one pair of Dice scores per patient), Wilcoxon signed rank tests were used. Statistical significance was defined as *P* less than .05. Total lesion volume was calculated from the manual and predicted segmentation masks, and Spearman correlations were calculated.

Results

Differences in Data Acquisition Parameters between IN1 and IN2

Differences in the MRI acquisition parameters and hardware between IN1 and IN2 are presented in the Table. Most notably, the majority of images from IN1 (73%) were acquired with Siemens MRI machines, while the majority of IN2 images (88%) were acquired with GE MRI machines. Furthermore, 78% of IN1 FLAIR images were acquired at 1.5-T field strength, while only 47% of IN2 FLAIR images were acquired at 1.5 T. At IN1, 97% of FLAIR images were two-dimensional (2D) acquisitions, while only 22% of IN2 FLAIR images were 2D acquisitions (with 78% being 3D acquisitions).

CNN-based FLAIR Lesion Segmentation Accuracy on External Institution Data

Lesion segmentation performance was marginally lower for the model trained on IN1 data (M_1) when tested on IN2 data than when tested on IN1 data (median Dice score, 0.70 [interquartile range {IQR}, 0.55–0.81] and 0.76 [IQR, 0.60–0.85], respectively; $P = .08$) (Fig 2). To evaluate the consistency of this effect given training stochasticity, we trained the same architecture on the same IN1 data five separate times, and we tested each of these training instances of M_1 on IN1 and IN2 data. The performance on IN1 test data (median Dice score range, 0.733–0.765) was consistently higher than on IN2 test data (median Dice score range, 0.693–0.733) across independently trained models (paired *t* test, $P = .001$), with the decrease in median Dice score ranging from 0.027 to 0.059 (mean, 0.042) when applied to the external (IN2) test dataset compared with the internal (IN1) test dataset.

We observed the same effect when using IN2 data only ($n = 34$) to train the model. M_2 demonstrated a median Dice score of 0.74 (IQR, 0.60–0.82) on IN2 test data and a median Dice score of 0.64 (IQR, 0.51–0.77) on IN1 test data ($P < .001$). Furthermore, training a model with all 51 IN2 patients resulted in a median Dice score of 0.60 (IQR, 0.50–0.74) when tested on IN1 test data, which was lower than the score from a model trained on the same number of IN1 training patients ($n = 51$) when tested on IN1 test data (median Dice score, 0.74 [IQR, 0.51–0.84]; $P < .001$).

Addition of Brain Tumor Training Data (M_{1+B} and M_{1+B+2T} Models)

Owing to the commonality of FLAIR hyperintensities across many pathologic conditions, we attempted to improve portability across institutions by incorporating additional training data with FLAIR hyperintense lesions from other institutions (from the BraTS dataset [M_{1+B}]). Including this multi-institutional dataset in training did not yield a significant change in median Dice score from the initial model's performance on

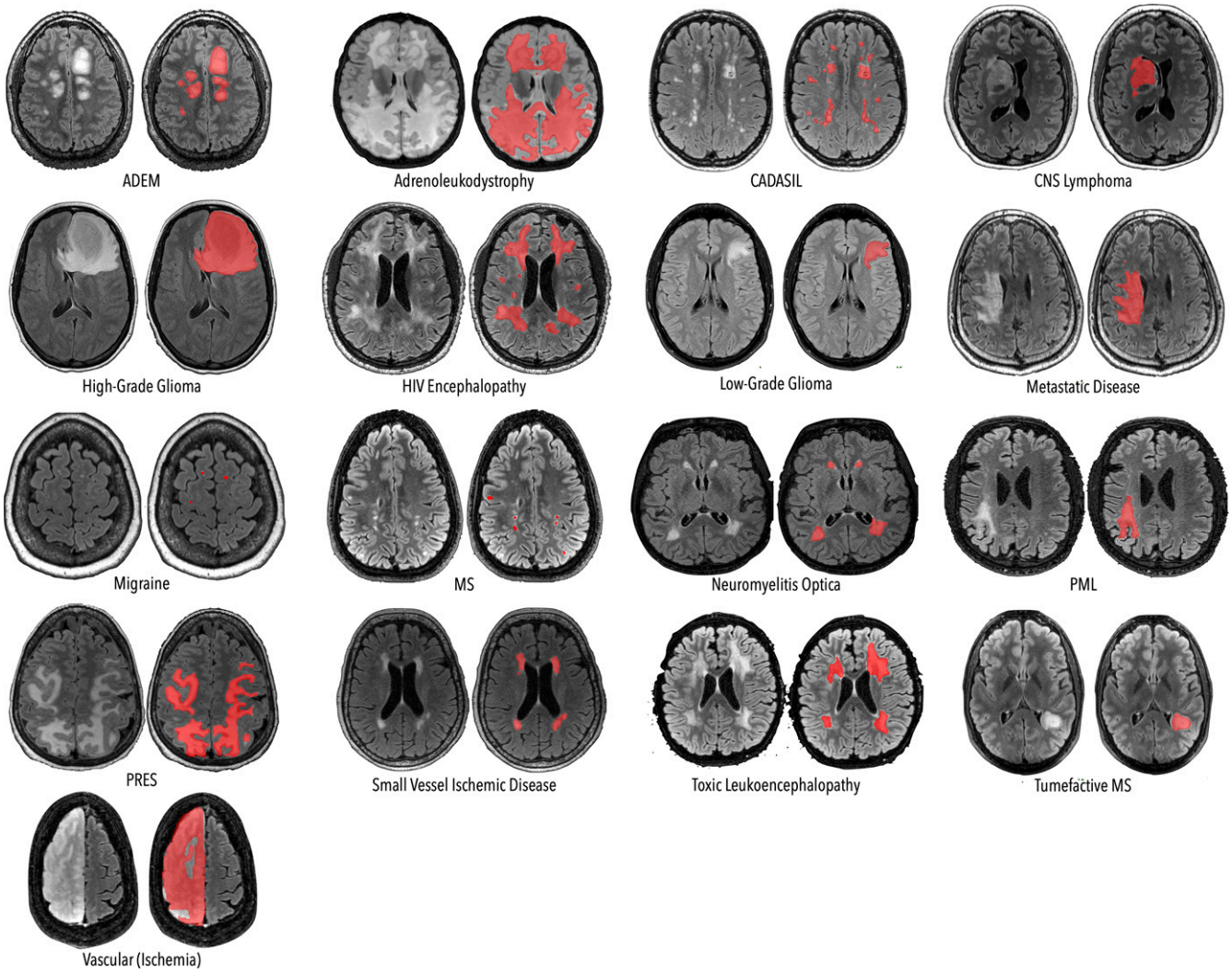

Figure 3: Representative single-section fluid-attenuated inversion recovery (FLAIR) images from 17 test samples using the final model (M_{1+2}), which was trained using data from both institutions. Note the large variation in number, size, and extent of lesions. On the left of each pair is the original image, and on the right of each pair is the automated segmentation overlaid on the original FLAIR image. ADEM = acute disseminated encephalomyelitis, CNS = central nervous system, CADASIL = cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy, MS = multiple sclerosis, M_{1+2} = model trained on institution 1 + institution 2 data, PML = progressive multifocal leukoencephalopathy, PRES = posterior reversible encephalopathy syndrome.

the IN2 test data (0.72 [M_{1+B}] vs 0.70 [M_1]; P = .1). Further addition of IN2 training data with only primary brain tumors as the pathologic feature (M_{1+B+2T}) (total training, n = 776) resulted in a marginal but not significant increase in median Dice score compared with the original model tested on IN2 data (0.73 [M_{1+B+2T}] vs 0.70 [M_1]; P = .06).

Addition of IN2 Training Data (M_{1+2} Model)

Compared with M_1 applied to the IN2 test dataset, the M_{1+2} model (addition of 34 IN2 training samples with various pathologic conditions, with threefold cross-validation across all IN2 data; 10% of total training sample from IN2 and 90% from IN1) yielded improved segmentation accuracy on the IN2 test dataset (median Dice score, 0.77 [M_{1+2}] vs 0.70 [M_1]; $P < .001$) (Fig 2). The addition of IN2 training data (M_{1+2}) did not degrade performance on IN1 test data (median Dice score, 0.75 [M_{1+2}] vs 0.76 [M_1]; P = .47). The performance of M_{1+2} tested on IN2 data was similar to the performance of M_1 tested on IN1 data (median Dice score, 0.77 [M_{1+2} tested on IN2] vs 0.76 [M_1 tested on IN1]; P = .45) (Fig 2).

Segmentation performance was good across the range of 17 diseases tested (Fig 3), although inherently some interdisease variation in Dice scores existed. The degree of interdisease variation in Dice scores was qualitatively similar to IN1 test data results from M_1 (Fig E1 [supplement]) and was similar to previously observed human interobserver variability (6).

Staged Addition of IN1 Training Data

To investigate the effects of the relationship between number of training cases and segmentation performance, we created models with subsets of IN1 data including either all 293 patients from IN1 (original M_1) or with either 204, 102, or 51 patients. In general, the addition of IN1 training data was helpful for performance on both IN1 and IN2 test data (Fig 4). However, the addition of a small amount of IN2 training

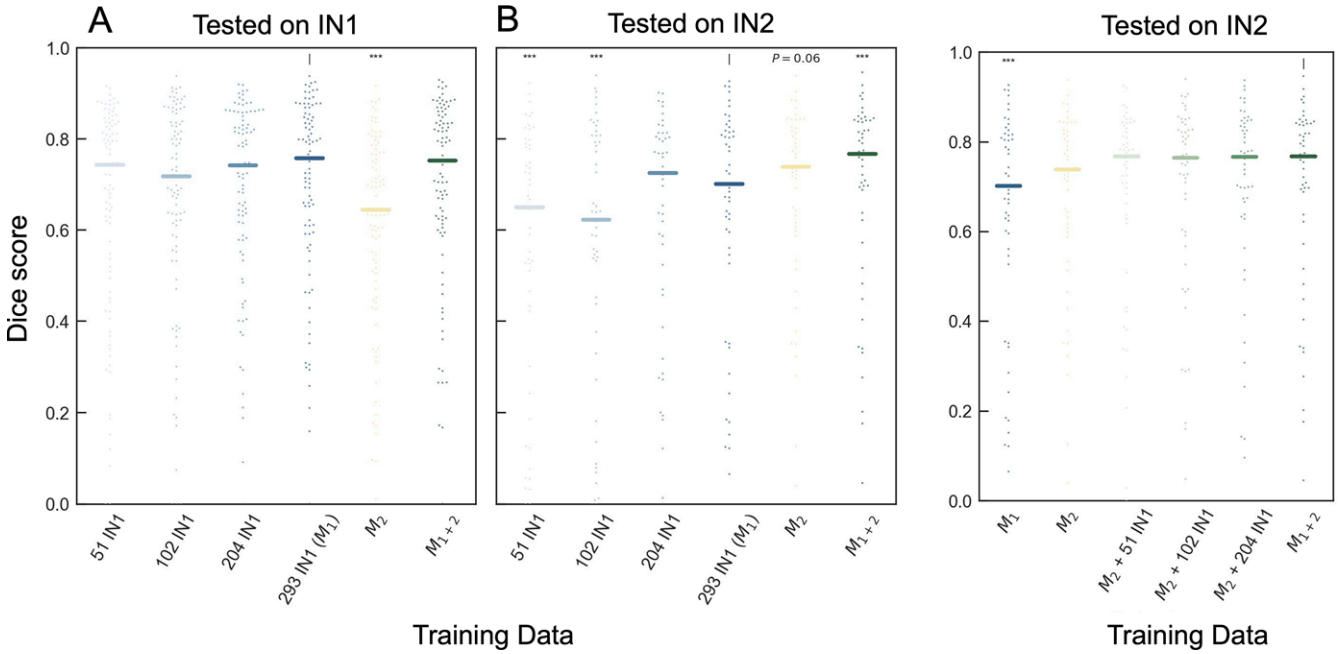

Figure 4: Effect of adding institution 1 (IN1) training data on IN1 and institution 2 (IN2) test data performance. **(A)** Dice scores on IN1 (left panel) and IN2 (right panel) test data for models trained with various datasets, including different amounts of IN1 training data. Median Dice scores are indicated by horizontal bars. A small amount of IN1 training data (blue shading) was sufficient for high performance when tested on IN1 test data, while a small amount of IN2 training data (M_2 or M_{1+2}) improved performance on IN2 test data. **(B)** IN2 test data performance for various models with incremental addition of IN1 training data (green shading) to a baseline M_2 model (yellow). Addition of IN1 training data had no effect on IN2-tested performance, while a small amount of IN2 data improved performance over the M_1 baseline model (*** = $P < .001$, $P = .06$ where indicated). Statistical comparisons are to the bars in each panel indicated by "|" using Wilcoxon signed-rank test. M_1 = trained on IN1 patient data, M_2 = trained on IN2 data, M_{1+2} = trained on IN1 + IN2 data.

data had a much greater effect than a large amount of IN1 training data on IN2 test data performance. Similarly, removing portions of IN1 training data when IN2 training data were used had no measurable effect on performance of the model tested on IN2 data, unless all IN1 training data were removed (Fig 4B).

Transfer Learning versus Training on Combination of Data

We applied fine-tuning methods to test whether transfer learning starting with M_1 and using IN2 data ($M_{FT\,(1+2)}$) could result in performance improvements on IN2 test data similar to the performance improvements of training a model with a combination of the same IN1 and IN2 training data from scratch (M_{1+2}). We performed multiple fine-tuning experiments that involved separately fine-tuning the following layers: first layer, middle layers, last layer, first and last layers, all layers. The best-performing model was the model in which all layers were fine-tuned, demonstrating a median Dice score of 0.78 (IQR, 0.67–0.85) on IN2 test data. Owing to consistent improvements in Dice score across the majority of patients in the test datasets, $M_{FT\,(1+2)}$ had higher performance than M_{1+2} (median Dice score, 0.78 and 0.77, respectively; $P < .001$) (Fig E2 [supplement]).

Fine-tuning the original IN1-trained model resulted in substantial time savings, with fine-tuning taking 291 seconds per epoch, compared with 2374 seconds per epoch when training from scratch. Across 30 epochs, using our hardware, the time savings was more than 17 hours for training an individual model.

Lesion Volume Quantification and Its Effect on Segmentation Accuracy

Using the segmentation masks, total lesion volume could be directly calculated for each patient. Using manually delineated segmentation masks as the reference standard, the correlation between true total lesion volume and U-Net–predicted total lesion volume was very high, with a Spearman correlation of 0.98 ($P < .001$) and a Pearson correlation coefficient of 0.90 ($P < .001$) for M_{1+2} (Fig 5A). Performance was noticeably lower for patients with extremely high lesion volumes, which were far outside (>3 standard deviations) the range of lesion volumes in the training data.

Deviations of predicted lesion volume from true lesion volume were similar between the M_{1+2} model tested on IN2 and the M_1 model tested on IN1 (Fig 5B). Interestingly, M_{1+2} had higher performance than M_1 on small (<10 cm^3) lesions (median Dice score, 0.81 [M_{1+2} on IN2 test data] vs 0.54 [M_1 on IN1 test data]; $P < .05$) (Fig 5C).

Imaging Parameter Effects on Segmentation Accuracy

Scan parameters other than underlying disease cause or lesion volume may theoretically also have substantial effects on segmentation performance. For example, differing MRI hardware may cause images to appear different and therefore cause variable lesion segmentation performance. However, we found no difference in performance across different scanner models when M_{1+2} was tested on IN1 (one-way Kruskal-Wallis test of Dice across 10 scanner types, H = 6.83; $P = .66$) or manufacturers (one-way Kruskal-Wallis test across three manufacturers,

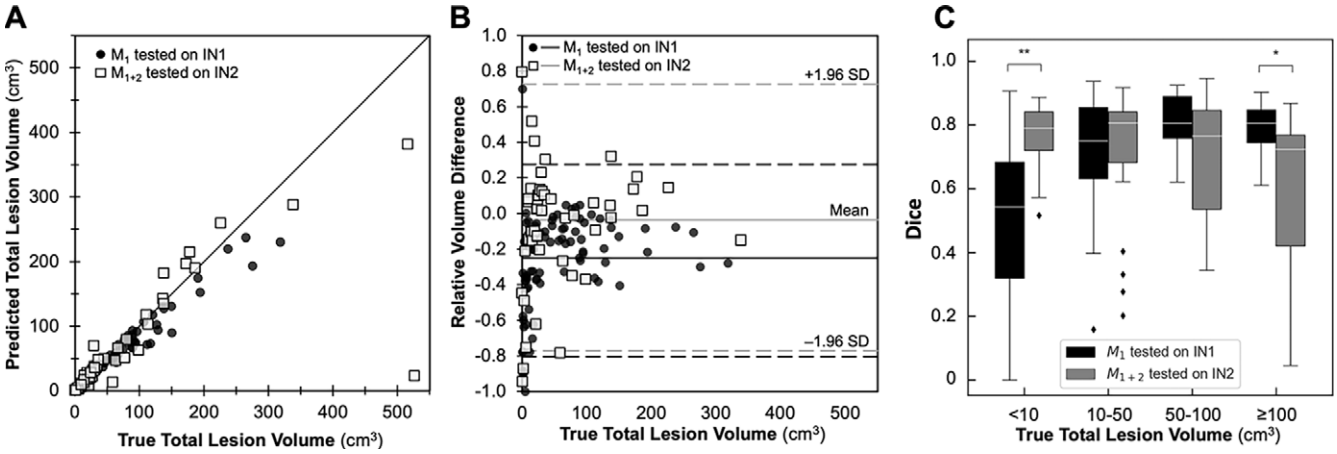

Figure 5: Lesion volume estimates and volume effect on Dice score. **(A)** The correlation between true total lesion volume and predicted total lesion volume for M_{1+2} model (white squares) was high (Spearman $r = 0.98$). Black circles represent data from M_1 tested on institution 1 (IN1) data as a reference. **(B)** Bland-Altman plot demonstrating the difference between predicted and true total lesion volume on the test set as a function of true total lesion volume in the range 0 to 500 cm^3. Lines represent the mean (solid line) and ± 1.96 standard deviations (SD; dashed line) for M_1 with IN1 test (black) and M_{1+2} with institution 2 (IN2) test (gray) models. Lesions larger than three SD above the mean training data lesion volume ($n = 2$) are excluded from this plot because they represent clear outliers. **(C)** Box plot of Dice score as a function of true total lesion volume for IN1-trained, IN2–fine-tuned model ($M_{FT\,(1+2)}$). (*) indicates $P < .05$, (**) indicates $P < .01$. M_1 = trained on IN1 patient data, M_{1+2} = trained on IN1 + IN2 data.

H = 0.68; $P = .71$) (Fig 6A), although the lack of significance may be partially related to low numbers of patients in each category (total, $n = 51$ [Table]).

There was no obvious relationship between the number of training cases on a particular scanner model and the segmentation performance on that model during testing (Fig 6B). There was no effect of field strength on segmentation performance ($P = .45$) (Fig 6C). There was also no effect of imaging acquisition dimensionality (2D vs 3D acquisitions) on segmentation performance ($P = .69$) (Fig 6C).

Discussion

The practice of radiology, including clinical neuroimaging, can be conceptually broken down into perceptual and cognitive components. Most errors in radiology are perceptual (15), and much of neuroradiologists' time is spent searching for and measuring lesions. With imaging volumes steadily increasing, thereby allowing less time for each imaging study, such ancillary tasks could be relegated to automated algorithms. These automated algorithms have less bias, less variability, and are much faster than human search patterns. For those reasons, we had previously created and trained a 3D U-Net CNN for detection of abnormal FLAIR signal (6), a sensitive marker of disease types that often guides the description of abnormal MRI findings. This AI algorithm was previously found to generalize well across disease types and across different scanners at a single institution.

However, a major limitation to the use of AI technologies in the clinical setting has been a poor ability to generalize outside of single health care institutions. Just as external validity is critical for testing the generalizability of randomized clinical trial results (16), AI technologies must be proven to function beyond single institutions prior to clinical deployment. In our data, we found a decrease in performance when attempting to port the FLAIR U-Net outside of the institution in which it was originally trained (M_1 tested on IN1 vs IN2 [median Dice score, 0.76 and 0.70, respectively] and M_2 tested on IN2 vs IN1 [median Dice score, 0.74 and 0.64, respectively]), despite substantial heterogeneity in the initial training data, including variable MRI scanner manufacturers, imaging parameters (echo and repetition times), spatial resolution, and other factors.

The segmentation performance of the externally tested model recovered to internally tested performance with the addition of a very small amount (10%) of training data that closely approximates the external dataset in terms of lesion and imaging characteristics. This result suggests that only limited additional data are needed to recover performance and return it to levels that approach previously established human performance for this task (6). Clearly, the amount of local data that are needed to recover intrainstitutional performance will depend on initial generalizability of the model and on the statistical similarities between the data from the two institutions. Currently, no clear guidelines exist on the amount of data that are sufficient to guarantee portability to a new institution. However, it appears that it is not necessarily the quantity of data, but rather its relevance, that increases performance. Our experiments demonstrate that only training data from the same institution with similar disease pathologic features as the test set increased segmentation performance. The addition of only brain tumor segmentations, which have idiosyncratic FLAIR patterns, did not increase the algorithm's portability to the new institution.

One major barrier to amassing the diverse dataset required for increasing generalizability is the concern for data privacy (17,18). The use of federated learning models, wherein network weights and parameters, rather than patient data, are shared between institutions, may help to appease these concerns (19,20). The results of our study support such efforts, as they demonstrate the potentially large gains from limited additional local training data. We demonstrate that the local data are used most efficiently when they are used to fine-tune the model, with an approximate

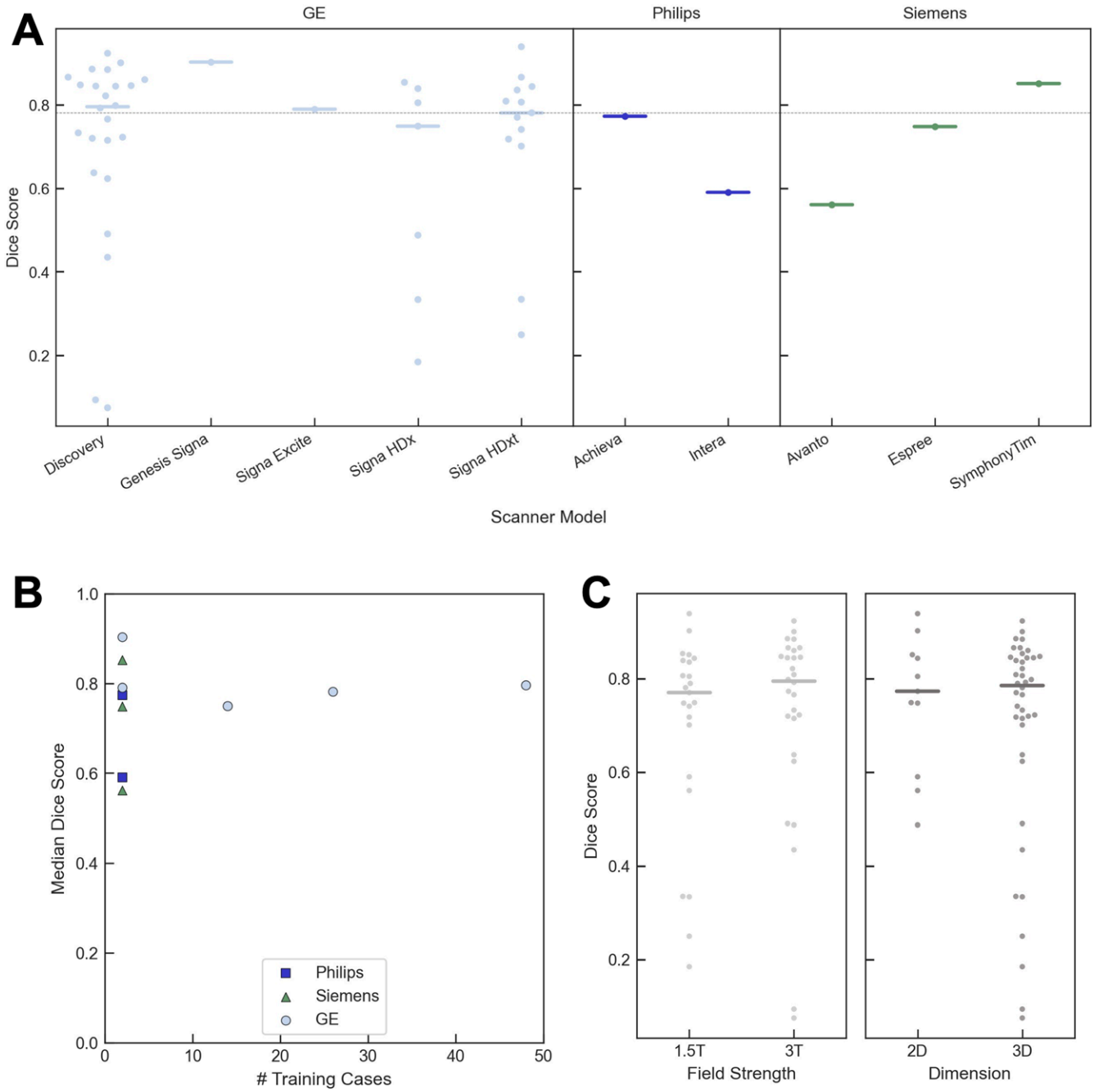

Figure 6: Lesion segmentation performance of the bi-institutionally trained model according to various imaging parameters. **(A)** Dice score based on institution 2 (IN2) test data as a function of MRI machine model, grouped by MRI machine manufacturer. Dashed gray line represents median Dice score for the institution 1 (IN1)–trained, IN2–fine-tuned model ($M_{FT\,(1+2)}$) tested on IN2 (0.78). There were no significant differences in median Dice score across 10 machine models ($P = .66$) or across three manufacturers ($P = .71$). **(B)** Dice scores in test data according to the number of training cases from the same machine model. **(C)** Dice scores as a function of field strength (left) and acquisition dimension (right). There was no effect of field strength ($P = .45$) or acquisition dimension ($P = .69$). Horizontal bars represent median Dice scores. All data shown are for $M_{FT\,(1+2)}$ tested on IN2 data using threefold cross-validation to ensure independent training and testing data. # = number of, 3D = three dimensional, 2D = two dimensional.

10-fold time savings and slightly increased segmentation performance compared with combining the data from institutions and training from scratch. The alternative, to completely retrain a model at each new institution with its own training data, may result in similar performance but is extremely time-consuming and expensive owing to the effort required to curate high-quality labeled training data.

This study had several limitations. It tested the portability of a single AI algorithm to a single new institution on a modest dataset, and it is unclear whether the same results would hold true for other algorithms and other institutions, although we suspect that these are general principles for AI in radiology. The specific amount of local training data required before deploying to the new institution, however, will be both model- and

institution-dependent and will need to be evaluated for any given model, including disease-specific models trained at one institution. For example, while initial external validation performance of the FLAIR U-Net was lower than internal validation performance, the overall decrease in performance was relatively modest, likely owing to the image acquisition parameter diversity of the original training dataset. The FLAIR U-Net itself also has limitations, including only being tested on the 17 specific disease entities included in this study, and not performing as well on small lesions (6) and extremely large lesions outside the range of training data.

Ongoing work aims to identify whether similar principles of fine-tuning allow efficient adaptation of segmentation algorithms from one MRI sequence to another or from one disease (or set of diseases) to another, such as from primary brain tumors to metastases. Additionally, experiments are underway to examine whether it may be possible to apply generative adversarial networks using input from one institution to generate synthetic data outputs that more closely resemble the distribution of acquisition parameters of another institution, thereby further mitigating the need for labeling external datasets.

In conclusion, we tested the interinstitutional portability of an AI algorithm for detection of FLAIR lesions at brain MRI. The AI algorithm performed well on its intended task of lesion segmentation across a variety of neurologic diseases. A model trained and tested at one institution had lower performance on data from an outside institution. However, with the addition of a small amount of highly relevant training data from the outside institution, full performance was recovered. These results suggest a means to AI algorithm portability from one institution to another without requiring extensive new training data.

Acknowledgment: We gratefully acknowledge the support of NVIDIA, which donated a Titan Xp GPU used for this research.

Author contributions: Guarantors of integrity of entire study, **A.M.R.**, **M.T.D.**; 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, **A.M.R.**, **T.J.G.**, **P.N.**, **M.T.D.**, **C.P.H.**; clinical studies, **T.G.**; experimental studies, **A.M.R.**, **P.N.**, **D.A.W.**, **E.C.**, **J.D.R.**; statistical analysis, **A.M.R.**, **P.N.**, **M.T.D.**, **J.B.C.**, **J.D.R.**; and manuscript editing, all authors

Disclosures of conflicts of interest: **A.M.R.** Institution received American Society of Neuroradiology (ASNR) trainee grant and Carestream Health/Radiological Society of North America (RSNA) research scholar grant; former member of *Radiology: Artificial Intelligence* trainee editorial board. **T.J.G.** No relevant relationships. **P.N.** Formerly employed by Perceus and PrinterPress (urologic devices and orthopedic devices); grants/grants pending from Perceus for a urology device; patent from Perceus for a urology device; stock/stock options in PrinterPress (orthopedic devices). **M.T.D.** No relevant relationships. **D.A.W.** Consultancy fee from Galileo CDS for work not related to this publication; stock/stock options in Galileo CDS received for work not related to this publication. **E.C.** No relevant relationships. **J.B.C.** No relevant relationships. **L.P.S.** No relevant relationships. **J.D.R.** Institution received ASNR neuroradiology research grant in AI; former member of *Radiology: Artificial Intelligence* trainee editorial board. **C.P.H.** Medical imaging consultant for GE Healthcare; research travel expenses from Siemens Healthineers; member of data monitoring and safety boards for Insightec and UniQure; former member of *Radiology* editorial board.

References

1. Beam AL, Manrai AK, Ghassemi M. Challenges to the Reproducibility of Machine Learning Models in Health Care. JAMA 2020;323(4):305–306.
2. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18(8):500–510.
3. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 2018;15(11):e1002683.
4. Onofrey JA, Casetti-Dinescu DI, Lauritzen AD, et al. Generalizable Multi-Site Training and Testing Of Deep Neural Networks Using Image Normalization. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Conference Location, Conference date. Piscataway, NJ: IEEE, 2019; 348–351 https://doi.org/10.1109/ISBI.2019.8759295.
5. Bhuva AN, Bai W, Lau C, et al. A Multicenter, Scan-Rescan, Human and Machine Learning CMR Study to Test Generalizability and Precision in Imaging Biomarker Analysis. Circ Cardiovasc Imaging 2019;12(10):e009214.
6. 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.
7. Rauschecker AM, Rudie JD, Xie L, et al. Artificial Intelligence System Approaching Neuroradiologist-level Differential Diagnosis Accuracy at Brain MRI. Radiology 2020;295(3):626–637.
8. Rudie JD, Rauschecker AM, Xie L, et al. Subspecialty-Level Deep Gray Matter Differential Diagnoses with Deep Learning and Bayesian Networks on Clinical Brain MRI: A Pilot Study. Radiol Artif Intell 2020;2(5):e190146.
9. Bakas S, Reyes M, Jakab A, et al. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. arXiv 1811.02629 [preprint]. https://arxiv.org/abs/1811.02629. Posted November 5, 2018. Accessed May 7, 2019.
10. 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.
11. 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.
12. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 2010;29(6):1310–1320.
13. Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv 1603.04467 [preprint]. https://arxiv.org/abs/1603.04467. Posted March 14, 2016. Accessed May 7, 2019.
14. Dice LR. Measures of the Amount of Ecologic Association Between Species. Ecology 1945;26(3):297–302.
15. Bruno MA, Walker EA, Abujudeh HH. Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction. RadioGraphics 2015;35(6):1668–1676.
16. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet 2005;365(9453):82–93.
17. Langlotz CP, Allen B, Erickson BJ, et al. A Roadmap for Foundational Research on Artificial Intelligence in Medical Imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop. Radiology 2019;291(3):781–791.
18. Larson DB, Magnus DC, Lungren MP, Shah NH, Langlotz CP. Ethics of Using and Sharing Clinical Imaging Data for Artificial Intelligence: A Proposed Framework. Radiology 2020;295(3):675–682.
19. Chang K, Balachandar N, Lam C, et al. Distributed deep learning networks among institutions for medical imaging. J Am Med Inform Assoc 2018;25(8):945–954.
20. Sheller MJ, Reina GA, Edwards B, Martin J, Bakas S. Multi-institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation. In: Crimi A, Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T, eds. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2018. Lecture Notes in Computer Science, vol 11383. Cham, Switzerland: Springer, 2019; 92–104.