10.1148/radiol.2020190283

Appendix E1

Supplemental Methods

Patients and Diseases: Details

We identified study patients by searches of radiology reports (using mPower search engine, Nuance Communications, Burlington, MA) at our academic tertiary care center for 19 prespecified diagnoses. Searches were aimed to be broad, using multiple possible names for each diagnosis. Inclusion criteria included studies acquired between January 2008 and January 2018 at our own institution or uploaded to our PACS for secondary interpretation from an outside institution during the same timeframe. Patients were randomly chosen (up to a maximum of 15) from this search for each diagnosis, resulting in 279 potentially eligible patients. Diagnoses were confirmed by comprehensive chart review to ensure accurate reference standard diagnoses, using pathologic data when available, or by follow-up clinical and radiologic assessments over time, resulting in exclusion of 56 patients (Fig 1). After identifying patients with the diseases of interest, we chose the first diagnostic MRI for each patient as the specific study to include, which often resulted in studies prior to the ultimate diagnosis, therefore necessitating a differential diagnosis. Exclusion criteria included multiple neurologic diagnoses or history of prior cranial surgery causing fluid-attenuated inversion recovery (FLAIR) abnormality, excessive imaging artifact precluding radiologic interpretation, lack of FLAIR sequence in the study protocol, or no imaging findings within the cerebral hemispheres (Fig 1).

The 19 diseases included in the test sample were: low-grade glioma (grade I and grade II), high-grade glioma (grade IV glioblastoma), primary central nervous system (CNS) lymphoma, metastatic disease, vascular disease (acute or subacute ischemia), small vessel ischemic disease (SVID), Susac's syndrome, active multiple sclerosis, inactive multiple sclerosis, tumefactive multiple sclerosis, neuromyelitis optica (NMO), acute disseminated encephalomyelitis (ADEM), X-linked adrenoleukodystrophy, cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL), human immunodeficiency virus (HIV) encephalopathy, migraine, progressive multifocal leukoencephalopathy (PML), toxic leukoencephalopathy, and posterior reversible encephalopathy syndrome (PRES). Up to 12 patients from each diagnostic category were included in the initial study cohort.

MRI manufacturers included GE (Chicago, IL), Philips (Amsterdam, Netherlands), Siemens (Munich, Germany), and Toshiba (Otawara, Japan).

Five patients with each of the 19 diseases were randomly chosen according to the procedure described above from this initial study cohort to form the test sample. If a diagnosis was so rare that less than five patients with the disease exist in our institution's PACS (which was only the case with Susac's syndrome), then all patients with that disease were included in the test sample (none in the training sample). The remainder of the patients were assigned to the training set (Table 3 in Materials and Methods).

All FLAIR lesions on training images were hand-segmented by a radiologist using ITK-SNAP (29), to provide segmentation masks for training the FLAIR U-Net. The same procedure was followed for the GRE and T1 U-Nets, with all abnormalities hand-segmented on that modality without the use of other modalities. Diagnosis was not available to the radiologist at the time of hand segmentation.

Image Pre-Processing and Convolution Neural Networks

For input to the first portion of the AI algorithm, the T1-weighted images were first skull-stripped using Advanced Normalization Tools (ANTs), and the skull-stripping was then transferred to coregistered FLAIR volumes. Tissue segmentation into gray matter, white matter, and CSF was performed by ANTs. Due to intensity changes in T1-weighted MRI caused by many pathologic lesions, lesions in white matter are often mislabeled (mostly as CSF or background) in the default ANTs tissue segmentation (known as Atropos). To correct for this error and generate a white matter mask that covers these lesions, a tailored method was applied. This method consisted of running Atropos twice, first with default setting and then again with high weight (0.75) given to the atlas-based spatial prior probabilities, to generate a segmentation that is less sensitive to intensity changes (referred to as high-weight segmentation). Then the labels of the voxels that are segmented as CSF/background in the original segmentation but are segmented as white matter in the high-weight segmentation were changed to white matter. Visual inspection on training MRIs found that this step improved lesion coverage of the white matter mask and mislabeling of lesions within white matter as other tissue types.

As input to the U-Net convolutional neural networks, images were normalized by the mean and standard deviation signal intensity to zero mean and unit standard deviation. Image volumes were resampled to 1 mm^3 isotropic resolution via linear interpolation. Elastic transformations (30) were applied to the images for augmentation purposes, including small random affine transformations and small random free-form deformations. We then split the augmented image into 96 mm^3 cubes ("3D patches") as the network input. During training, the cubes were randomly sampled across the full brain volumes. To prevent sample imbalance during training, the same number of patches that included lesions as those that excluded lesions were sampled. During testing, the brain volume was densely sampled with the cubes using a step size of 32 in each direction, with a 64 pixel overlap between cubes. Overlapping segmentation predictions were averaged.

We employed a 3D U-Net architecture (31,32) to predict lesion segmentations on the FLAIR images, as previously described (9) and as shown in the schematic of Figure 3. We also used the same network architecture for predicting abnormal T1 signal and abnormal susceptibility-related signal loss, using separate training MRIs consisting only of those modalities. The network architecture consists of 4 consecutive down-sampled blocks followed by 4 consecutive up-sampled blocks, using a rectified linear unit for nonlinearity. For down-sampling, we used a stride-2 convolution; for up-sampling, we used a stride-2 deconvolution. We used kernel size 3 × 3 across the network. In the down-sampling block, we applied a dilation factor of 2 in all convolutional layers. In addition to the cross-links between corresponding up-sampling and down-sampling blocks, there is also a residual connection between subsequent layers with number of features matched by a plain 1 × 1 convolution. After the final up-sampling block, three additional convolutional, rectified linear unit, batched-normalized (conv-ReLu-bn)

layers were added, prior to the final normalized exponential (softmax) function. A batch consisted of six patches (Fig 3a).

We used standard cross-entropy loss in all U-Nets. We used an Adam optimizer with learning rate of 10^{-5}. No hyperparameter optimization was performed, since network training had been previously performed on independent data and the network weights were applied to the test MRIs only a single time. All deep learning processing (training and testing) was implemented on an NVIDIA Titan Xp GPU (NVIDIA corporation, Santa Clara, California) using Python version 3.7 and Tensorflow (33).

Finally, a cluster threshold of 5 mm^3 was applied to the lesion mask, excluding clusters of voxels smaller than this threshold.

Quantitative Feature Extraction Methods

Findings on each MRI were extracted using a combination of image processing techniques, described in detail below. Briefly, for five *signal features*, we identified abnormal signal using independent U-Nets and in-house MATLAB (Mathworks, Natick, MA) scripts detecting signal deviations relative to the mean signal in the cerebral hemispheres. Resulting abnormal signal maps were coregistered with the FLAIR lesion mask. Six *volumetric features* of lesions (eg, lesion volume) were calculated directly from FLAIR lesion masks. Seven *spatial features* were calculated by lesion overlap with atlas-based regions of interest. Quantitative outputs were thresholded to convert them to one of two or three possible "feature states" (eg, "present" or "absent"), which were then used as input to the Bayesian inference mechanism.

Specifically, first, intermodality rigid-body registration and deformable registration to the OASIS standard atlas template (34) was performed using ANTs. T1 W images were up-sampled to reduce effective slice thickness, using a patch-based superresolution technique (35) for this portion of the pipeline to preserve in-plane resolution. Individual quantitative or semiquantitative features were then extracted as follows, followed by thresholding each feature to convert it from the calculated quantitative or semiquantitative measure into a "feature state." The feature state is one of 2 or 3 possibilities for each feature, as listed for each feature below. For instances where the feature could not be calculated (eg, missing sequence or other missing data, as well as indeterminate feature states as defined below), the feature received an "N/A" state and was not used as input to the Bayesian network for that patient. Similarly, predefined thresholds were chosen such that intermediate values could result in an "N/A" state, implying that the available information was not strong enough to result in a confident assessment of either feature state and should not be used for Bayesian calculations. Thresholds were chosen based on maximizing performance on the *training* dataset, defined as the highest accuracy in predicting neuroradiology attending "reference-standard" feature descriptions as well as maximizing diagnostic performance on "top 3" differential diagnosis on the same training set.

Signal Features

T1 signal.—

An independent U-Net determined abnormally higher or lower T1 signal on a voxelwise basis. The predicted abnormal signal mask was overlaid onto the predicted lesion mask (from the FLAIR U-Net) to determine the volume of abnormal (high and low) T1 signal within lesions

(defined as a region of FLAIR abnormality). Only the top half of lesions in terms of size are considered to avoid noise from small lesions.

Feature states: "High," "Low," or "Normal"

Thresholds: High T1 signal takes precedence over low T1 signal. Threshold for high T1 signal is ≥ 10 mm^3. Threshold for low T1 signal is ≥ 20% of total lesion volume.

Susceptibility.—

An independent U-Net determined abnormally low signal on GRE or SWI sequences (whichever was available) on a voxelwise basis. The predicted abnormal signal mask was overlaid with the predicted lesion mask (from the FLAIR U-Net) to determine the volume of abnormal susceptibility-related signal loss within each lesion.

Feature states: "Present" or "Absent"

Threshold: ≥ 150 mm^3 of abnormal susceptibility within lesions is "present"

T2 Signal.—

Average T2 signal within the white matter (based on ANTs tissue segmentation) and outside of lesions (based on FLAIR U-Net prediction) was calculated. The relative ratio of the T2 signal within the lesion mask compared with this "normal white matter T2 signal" was calculated.

Feature states: "High" or "Normal"

Threshold: Ratio of > 1. A low threshold is chosen to be maximally sensitive to high T2 signal.

Diffusion.—

Reduced diffusion was identified by a combination of DWI and ADC volumes, modeling the steps a neuroradiologist takes to identify foci of reduced diffusion. First, a mean DWI value across all cortical gray and white matter DWI values for that patient was calculated. Candidate regions of reduced diffusion were identified as voxels higher than 2.5 standard deviations above this mean value on the DWI volume. This candidate region map was overlaid on another map of ADC values that were 0.5 standard deviations below the mean ADC value across all gray and white matter. Regions of overlap between the abnormal DWI and ADC masks were labeled as voxels with reduced diffusion. The resulting mask was overlaid with the lesion mask from the FLAIR U-Net, and the volume of reduced diffusion was calculated for each lesion in each patient. A threshold was set for volume of reduced diffusion within a lesion, and the patient's MRI was considered to show reduced diffusion if one or more lesions passed this threshold.

Feature states: Reduced diffusion "Present" or "Absent"

Threshold: > 100 mm^3 within at least one FLAIR lesion is "present"

Enhancement.—

The T1pre sequence was coregistered with and then subtracted from the T1post sequence. The resulting subtraction map was overlaid on the FLAIR U-Net lesion prediction mask. The volume of voxels two standard deviations above the mean of the subtraction map (again counting only voxels within the cerebral hemispheres' gray matter or white matter) overlapping with the predicted lesion mask was calculated for each lesion. Similar to the calculation for reduced

diffusion, a threshold of enhancement volume for each lesion was set, and the patient's MRI was considered to show enhancement if one or more lesions passed this threshold.

Feature states: "Present" or "Absent"

Threshold: All lesions with greater than 40% of its voxels showing enhancement are considered candidate enhancing lesions. If the volume of the largest region of enhancement is greater than 1000 mm^3, enhancement is definitely present. If the volume is between 10 and 1000 mm^3, the feature state is indeterminate (so the conditional probability is unused, similar to a missing sequence). Enhancing volume less than 10 mm^3, or no lesions passing the 40% threshold, is considered as enhancement "absent."

Volumetric Features

Number of lesions.—

The predicted lesion mask from the FLAIR CNN was automatically divided into clusters of contiguous voxels, with each cluster corresponding to a separate lesion. The number of lesions (ie, contiguous clusters of voxels) was then extracted for each patient.

Feature states: "Few" or "Many"

Thresholds: < =3 is "few," ≥ 6 is "many," in between is indeterminate

Lesion volume.—

The volume of each individual FLAIR lesion was calculated by multiplying the number of voxels in each predicted lesion by the scan resolution. This resulted in an array of lesion volumes. The threshold to convert to the feature state was based on the maximum individual lesion volume (ie, size of largest lesion).

Feature states: "Large," "Medium," or "Small"

Thresholds: < =1000 mm^3 is "small," 1000–15000 mm^3 is "medium," and ≥ 15000 mm^3 is "large"

Lesion extent.—

The total volume of abnormal FLAIR signal was calculated based on the FLAIR U-Net prediction and was termed lesion extent. While lesion volume (above) was intended to capture the size of individual lesions, lesion extent was intended to capture the total volume of abnormal tissue within the cerebral hemispheres.

Feature states: "Extensive" or "Limited"

Thresholds: < =10000 mm^3 is "limited," ≥ 50000 is "extensive," in between is indeterminate

Enhancement ratio.—

For each lesion, the volume of enhancing voxels (as defined above) was divided by the individual lesion volume to determine an enhancement ratio (0 to 1) that estimates the percentage of lesion that shows enhancement. The median ratio of the top third largest lesions (in terms of lesion volume) was used for thresholding.

Feature states: “Large” or “Small”

Thresholds: < =0.1 is “small,” ≥ 0.12 is “large,” in between is indeterminate

Mass effect.—

Lesions with fewer surrounding CSF voxels than the same lesion mask transformed to the contralateral hemisphere were considered to exert positive “mass effect” (ie, effacement of the CSF). To accomplish this calculation, the predicted FLAIR lesion with the largest volume was dilated with a kernel of 15 mm^3, and then the lesion mask itself was subtracted, to derive a mask of voxels *surrounding* the largest lesion. This “rim” mask was reflected to the contralateral “non-lesion” hemisphere as well. The local sum of overlap between the rim masks and CSF (based on ANTs tissue segmentation and excluding lateral ventricle voxels) was calculated in each hemisphere using a 5 mm^3 convolutional filter. The difference of local sums between lesion and nonlesion hemispheres was calculated and normalized by the size of the rim lesion and by the total number of lesions, resulting in a value between 0 and 1 (0 = less mass effect, 1 = more mass effect).

Feature states: “Present” or “Absent”

Thresholds: < =0.04 is “absent,” ≥ 0.119 is “present,” in between is indeterminate

Ventricular volume.—

The volume of the lateral ventricles was directly calculated based on the ANTs tissue segmentation.

Feature states: “Enlarged” or “Normal”

Thresholds: < =31000 mm^3 is “normal,” ≥ 41000 mm^3 is “enlarged,” in between is indeterminate

Spatial Features

Lobar distribution.—

We counted the number of voxels in the FLAIR U-Net lesion prediction mask overlapping with each lobe (frontal, parietal, temporal, and occipital), as based on atlas-based coregistration with the T1-weighted image. The percentage of lesion volume within each lobe as a function of total lesion volume was calculated, and if this percentage surpassed an individual threshold for any one (and no more than one) lobe, then this was counted as the “predominant” lobe. Otherwise, an “N/A” value was assigned.

Feature states: “Frontal,” “Parietal,” “Temporal,” or “Occipital”

Frontal threshold: 60% (ie, 60% of total lesion volume is contained within frontal lobes)

Parietal threshold: 50%

Temporal threshold: 50%

Occipital threshold: 20%

Anterior temporal lobe involvement.—

The number of predicted lesion voxels within the anterior third of the temporal lobe was calculated in each hemisphere. If this value surpassed the threshold in both hemispheres, the MRI was considered to have prominent anterior temporal lobe involvement.

Feature states: "Present" or "Absent"

Thresholds: "Present" if ≥ 10 mm^3 total anterior temporal lobe lesion volume and absolute value of (L-R)/(L+R)<0.3, where L is the volume of left hemisphere anterior temporal lobe lesions and R is the volume of right hemisphere anterior temporal lobe lesions.

Symmetry.—

The predicted lesion volume in each hemisphere was calculated, and symmetry was calculated as a ratio of (left hemisphere–right hemisphere)/(left hemisphere + right hemisphere), resulting in a value between-1 and 1. The absolute value of this ratio, if high, determined asymmetry.

Feature states: "Symmetric" or "Asymmetric"

Thresholds: ≥ 0.55 is "asymmetric," < 0.55 is "symmetric"

Corpus callosum involvement.—

The predicted lesion volume overlapping with core corpus callosum (based on atlas coregistration) was calculated, and a ratio to total lesion volume was calculated. If either the total volume of lesions within the corpus callosum passed a threshold *or* the percentage of total lesion volume involving the corpus callosum passed a separate threshold, the MRI was considered to prominently involve the corpus callosum.

Feature states: "Yes" or "No"

Thresholds: Volume thresholds of 500 and 1000 mm^3 ("no," indeterminate, "yes"). Percentage thresholds of 1.6% and 3.3% ("no," indeterminate, "yes")

Cortical gray matter involvement.—

The percentage of predicted lesion volume overlapping with cortical gray matter (based on ANTs tissue segmentation) was calculated.

Feature states: "Yes" or "No"

Thresholds: < =10% is "no," ≥ 70% is "yes," in between is indeterminate

Periventricular lesions.—

For each lesion, the distance between lateral ventricles (based on ANTs tissue segmentation) and the closest lesion voxel was calculated, and lesions within a certain distance were termed "periventricular." All periventricular lesions' volumes were summed to determine a total "periventricular lesion volume." If the total periventricular lesion volume normalized by total lesion volume (across the brain) passed a threshold, the MRI was considered to have a prominent amount of periventricular lesions.

Feature states: "Yes" or "No"

Thresholds: Threshold for distance from ventricles is 5 mm. Thresholds for percentage of periventricular lesion volume: < =20% is "no," ≥ 40% is "yes," in between is indeterminate.

Juxtacortical lesions.—

For each lesion, the distance between cortical gray matter (based on ANTs tissue segmentation) and the closest lesion voxel was calculated, and the percentage of lesions falling within a certain distance was calculated. If the percentage of juxtacortical lesions exceeded a threshold, the MRI was considered to demonstrate a prominent amount of juxtacortical lesions.

Feature states: "Yes" or "No"

Thresholds: Threshold for distance from gray matter was 2 mm. Threshold for percentage of lesions is ≥ 50%.

Clinical Features

Age.—

The age of the patient was extracted from the medical record and thresholded. Exact age was provided to radiologists, while thresholded ages (feature states) were provided to the AI system.

Feature states (with thresholds): "Young" (< =40 yrs), "Adult" (41–60 yrs), "Old" (≥ 61 yrs)

Sex.—

A value of man (M) or woman (F) was assigned to each patient based on the sex stated in the patient's medical chart.

Feature states: "Woman" or "Man"

Immunocompromised.—

Whether or not patients were immunocompromised was determined from the medical chart based on a list of predefined qualifying conditions or current medications: HIV (regardless of current CD4 count), azathioprine, natalizumab, dimethyl fumarate, fingolimod, ocrelizumab, current chemotherapies (including intrathecal methotrexate), immunosuppressive therapies following organ transplantation, and recent (< 2 weeks) radiation therapy of any kind. This list of qualifying conditions/medications and the feature state was provided to radiologists reviewing MRIs, but the particular condition/medication for that particular patient was not provided.

Feature states: "Yes" or "No"

Viral prodrome.—

The presence of a viral prodrome was considered present if any clinical note within the electronic medical record mentioned a history of viral illness (respiratory, flu-like, or gastrointestinal) within a 2-week time period prior to the onset of neurologic symptoms. The details of the prodrome were not provided to the radiologist or the AI algorithm. If no viral prodrome was specifically mentioned, it was presumed to be absent.

Feature states: "Yes" or "No"

Chronicity of clinical symptoms.—

Acuity of neurologic symptoms, but not a description of the symptoms, was provided to the radiologist and the AI system for each patient. “Acute” was defined as neurologic symptoms occurring within 7 days or less of the MRI examination. “Chronic” was defined as neurologic symptoms lasting for more than 7 days prior to the MRI examination. “Acute on chronic” symptoms (eg, history of headaches, with recent severe headache 2 days prior to MRI different from prior headaches) were coded as “acute,” unless the chronic symptoms were specifically mentioned as the indication for imaging. If the patient was not experiencing neurologic symptoms prior to the MRI scan (eg, for routine cancer screening examination with incidental finding), then chronicity was coded as “N/A,” so that this feature was not used in the Bayesian inference.

Feature states: “Acute” or “Chronic”

Importantly, these features above were chosen because they are of general interest for radiologists and clinicians when reading MRIs (eg, number of lesions, degree of mass effect, and total lesion volume), *and* because they are thought to be important for differentiating the 19 diseases of interest. Additional potentially interesting quantitative features, such as lesion overlap with Brodmann area cortical parcellations, are also automatically included in the output of the image processing portion of the system, but these were not included in this version of the AI system because of limited relevance to differentiating the particular diagnoses included in the study.

Bayesian Network

The Bayesian network is a table of conditional probabilities that define the relationships between the neurologic diseases and the clinical and imaging features included in the study (see supplemental file “ConditionalProbabilities.xlsx” on https://github.com/rauscheck/radai). For each cell of this table, a probability of finding X feature given Y disease was estimated. In cases where such epidemiologic information was not available in the literature, we resorted to defining the probabilities using a consensus of two expert neuroradiologists’ estimates. These probabilities were defined without reference to any particular patients but rather represent the overall experience of the radiologists in their clinical practice.

Training data (n = 86) were primarily used to adjust the thresholds described above to accurately represent reference standard features. However, the same training data were also used to make minor adjustments to the probability table (to achieve higher overall top 3 accuracy performance on the training data). To avoid overfitting the probabilities to the training data, we gave only a small amount of weight ($\sim$ 3% weight per training study) to the training data and a much larger weight (the remainder) to the expert neuroradiologists’ estimates, which represent the accumulated experience of years of practice. Probabilities derived from the literature were not updated by training data, as these were considered “fact.” Since no “hyperparameter optimization” was performed, there was no separate validation set.

The prior (or “pretest”) probabilities of the probability table were set to reflect the distribution of diseases in the test dataset, which was equal probabilities for all diseases except Susac’s syndrome, which had a prior probability of 2.2% (2/92), compared with 5.26% (5/92) for all others. Analogously, radiologists were told that there was an equal probability of all diagnoses except Susac’s syndrome. As a result, prior probabilities had nearly negligible effects

on the AI system outputs, with differences in posterior probabilities almost completely driven by the specific features and their associated conditional probabilities.

Bayesian Inference

During testing, the thresholded features (ie, "feature states") were calculated as above for each patient. The final probability of each diagnosis was calculated by Naïve Bayesian inference, in effect a simple multiplication of the conditional probabilities according to the feature states for each possible diagnosis (code available at: https://github.com/rauscheck/radai). The AI-derived "top 3 differential diagnosis" represents a rank-ordered list of the three diagnoses with the highest calculated probabilities.

Data Analysis Details

Prevalence Analysis

Two academic neuroradiologists rated the relative clinical prevalence of each diagnosis on imaging, blinded to any performance measures, with respect to how often they make that diagnosis based on imaging. In cases where the two neuroradiologists disagreed, they came to a consensus after further discussion. For analysis, performance was averaged across each group of radiologists for each of the diagnoses within a prevalence category, with a standard deviation measured across diseases in that category.

ROC Analysis

For both radiologists and the AI system, the diagnostic choice data (each reader's first, second, and third diagnosis) were used as the basis of an ordinal scale of confidence (with four levels) in that diagnosis, which then allows for ROC analysis (36). Specifically, for each study, placing the target diagnosis in the first position of the differential diagnosis indicated high confidence in this diagnosis. Similarly, placing the target diagnosis in the second position of the differential diagnosis corresponded to slightly lower confidence, and placing the target diagnosis in the third position of the differential diagnosis corresponded to an even lower confidence. Not placing the target diagnosis in the top 3 differential diagnosis was counted as a false negative. This ordinal scale of confidence for each diagnosis was used to construct nonparametric ROC curves according to classic signal detection theory as applied to medical decision making (37–39), with the AUC used as a measure of criterion-independent performance–that is, this analysis takes into account the top 1, top 2, and top 3 diagnoses provided by all readers. Nonparametric AUCs were calculated with 95% confidence intervals determined by 100 bootstrapping samples on this nonparametric ROC curve, using the MATLAB routine "paramROC" (http://www.mathworks.com/matlabcentral/fileexchange/39127-parametric-roc-curve). In addition, for the purposes of visualization only, we constructed parametric ROC curves by fitting the responses using the gamma distribution.

Confusion Matrices

Confusion matrices were calculated for each reader (radiologists and AI system). For each reference standard diagnosis (x-axis), the proportion of times that the reader answered each possible diagnosis (as the top diagnosis) was calculated. Each column in the matrix sums to 100%.

To determine the similarity between any two confusion matrices, a correlation was calculated between matrix A and matrix B using the MATLAB function corr2, which can be understood as the degree of overlap or similarity between these two matrices. Because the confusion matrices evaluate the specific errors that are made between the possible diagnoses, the correlation coefficient is a measure of error similarity between two individuals. Statistical significance of differences in correlation coefficients was determined by the Fisher r-to-Z transformation.

Table E1. Full results of generalized estimating equation (GEE) for top 3 differential diagnosis accuracy, expressed as odds ratios for each group of radiologists relative to the AI system as baseline.

Top 3	Odds Ratio	Robust Std. Err.	z	*P* value	95% Conf. interval	
Residents	0.124	0.049	−5.32	<0.001	0.057	0.267
General Radiol.	0.113	0.045	−5.51	<0.001	0.052	0.246
Fellows	0.312	0.124	−2.92	0.003	0.143	0.681
Attending Neuroradiol.	0.579	0.245	−1.29	0.20	0.253	1.326
AI System	1	(omitted)				

Table E2. Full results of generalized estimating equation (GEE) for top 2 differential diagnosis accuracy, expressed as odds ratios for each group of radiologists relative to the AI system as baseline.

Top 2	Odds Ratio	Robust Std. Err.	z	*P* > |z|	95% Conf. interval	
Residents	0.302	0.075	−4.80	0.000	0.185	0.492
General Radiol.	0.306	0.081	−4.46	0.000	0.181	0.515
Fellows	0.743	0.195	−1.13	0.256	0.444	1.241
Attending Neuroradiol.	1.370	0.340	1.27	0.204	0.843	2.227
AI System	1	(omitted)				

Table E3. Full results of generalized estimating equation (GEE) for top 1 diagnosis accuracy, expressed as odds ratios for each group of radiologists relative to the AI system.

Top 1	Odds Ratio	Robust Std. Err.	z	*P* > |z|	95% Conf. interval	
Residents	0.399	0.098	−3.75	0.000	0.247	0.644
General Radiol.	0.361	0.091	−4.03	0.000	0.220	0.592
Fellows	1.069	0.247	0.29	0.775	0.679	1.682
Attending Neuroradiol.	1.507	0.348	1.78	0.076	0.958	2.371
AI System	1	(omitted)				

Table E4. Full set of P values associated with comparisons of the AI system to each individual radiologist evaluated on the same set of patients (McNemar's test), with regard to various outcome measures (top 1, top 2, or top 3 accuracy).

	Top 1	**Top 2**	**Top 3**
Resident 1	**<0.001**	**<0.001**	**<0.001**
Resident 2	0.07	**0.001**	**<0.001**
Resident 3	**0.01**	**0.008**	**<0.001**
Resident 4	**<0.001**	**<0.001**	**<0.001**
General 1	**0.02**	**0.001**	**<0.001**
General 2	**<0.001**	**<0.001**	**<0.001**

Fellow 1	0.27	**0.02**	**<0.001**
Fellow 2	0.09	0.59	0.14
Attending 1	0.15	0.33	0.08
Attending 2	0.10	0.26	0.61

Significant values (P < .05), without correction for multiple comparisons, are highlighted in **bold**.

Table E5. Partial AUC values and associated 95% confidence intervals (CIs) at high specificity values, ranging from 0.7 to 1.

	pAUC	95% CI
Residents	0.59	[0.55 0.64]
General Radiol.	0.58	[0.53 0.62]
Fellows	0.79	[0.74 0.83]
Attending Neuroradiologists	0.90	[0.84 0.94]
AI System	0.88	[0.81 0.94]

Results again demonstrate similar performance between the AI system and academic neuroradiology attending physicians, with neuroradiology fellows approaching a similar level of performance. Compared with the full (nonpartial) AUC values, where the AI system performs qualitatively highest, here the academic attending physicians perform qualitatively highest. The AI system, academic neuroradiology attending physicians, and neuroradiology fellows all exceeded the performance of general radiologists and radiology residents. pAUC = partial AUC.